Essays on Modern Market Structure

Marta Khomyn

PhD supervisor: Professor Tālis Putniņš

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

University of Technology Sydney

May 10, 2020

Certificate of Original Authorship

I, Marta Khomyn, declare that this thesis, titled "Essays on Modern Market Structure", is submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy, in the Finance Discipline Group at UTS Business School at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: May 10, 2020

To my parents.

Acknowledgements

I thank my academic supervisor, Professor Tālis Putniņš, for guidance and support throughout the work on this thesis. I am especially grateful to Tālis for instilling in me high standards for research quality and academic integrity. I also appreciate Tālis's patience during my growth journey as a PhD student and junior academic. It has been the foremost learning experience in my academic career to work with Tālis.

I thank my industry supervisor at Chi-X Australia, Dr. Shane Miller, for sharing his expertise in financial markets. Shane's involvement allowed me the freedom to pursue academic work in a high-pressure industry environment. I also thank Mike Aitken, Vic Jokovic, Murraugh O'Brien, Ross Pullen, Michael Somes, Howie Zhang, and all my colleagues at Chi-X Australia for their insights.

Various chapters in this thesis benefited from the helpful feedback by Drs. James Brugler, Paul Lajbcygier, Maureen O'Hara, Dave Michayluk, Vinay Patel, Tom Smith, Avanidhar Subrahmanyam, Zhuo Zhong, Kumar Venkataraman, and Christian Westheide. I also thank my fellow PhD students for their support, especially Dr. Nihad Aliyev, Dr. Marc Bohmann, Junquing Kang, Dr. Huong Nguyen, and Man Nguyen.

I am grateful for funding from the Capital Markets Cooperative Research Centre, International Research Scholarship, UTS Finance Discipline Group, APR Intern funding from Chi-X Australia, and UTS Business School Editing of Thesis Fund. This thesis benefited from the editorial assistance of Marita Smith.

I have been lucky to find a home away from home thanks to the kindness of Dr. Ojars Greste. I thank Ojars for many life lessons. My PhD journey in Australia would not have been the same without the friendship of Baiba Berzins, Cédric Le Gentil, Gerald Gloria, Volha Hrytskevich, Salome Hussein, Maija Kovalevska, Amy Lowe, James Millar, Jay Nam, Ronny Onggo, Ronald Sandoval, Roksolana Savytska, Susan So, Nicole Sun, and many more brave hearts of the UTS Outdoors Adventure Club. I thank Cédric Le Gentil for providing feedback on the very first draft of this thesis, and for inspiring excellence in academic research and in life.

Finally, I thank my parents. Their love has been an ever-present force in my life.

Abstract

Financial markets are different today from what they were two decades ago. This thesis examines recent issues in modern market structure: algorithmic liquidity provision, competition among exchange-traded funds (ETFs), and the shift of trading to the close of the trading day. The findings enhance our understanding of market structure changes resulting from technology, product innovation and market fragmentation.

Chapter 2 of this thesis examines how liquidity provision in fragmented markets affects order-to-trade ratios (OTTRs), a metric used by regulators to detect excessive quoting activity and market misconduct. The theoretical OTTR is determined by the trade-off between the market maker's information monitoring costs and picking-off risk (trading at stale prices). The theory explains why high OTTRs can result from legitimate market making in fragmented markets and are not necessarily a sign of misconduct. The empirical analysis supports the theoretical predictions. The empirical results suggest that recent growth in OTTRs is driven largely by fragmentation of trading across multiple venues and decreasing monitoring costs due to technological improvements. Calibration reveals that OTTRs on a typical day are within levels that are consistent with market-making activity, but occasionally spike beyond such levels. The results imply that regulatory measures designed to curb OTTRs (e.g., messaging taxes) are likely to harm liquidity provision in fragmented markets and create a non-level playing field for trading venues.

Chapter 3 asks how ETFs compete with one another and how their secondary market liquidity shapes this competition. It is puzzling that high-fee ETFs not only survive, but often accumulate greater assets under management (AUM), compared to low-fee ETFs tracking the same index. This chapter develops the equilibrium model of ETF competition, which resolves this puzzle. The main insight from the model is that secondary market liquidity of an ETF plays a key role in determining ETF fees and leads to liquidity clienteles. Greater liquidity attracts high-turnover investors, which sustain the high liquidity in a self-perpetuating cycle. The liquidity advantage allows the high-fee ETF to charge higher fees. The low-fee ETF serves low-turnover clientele, who are more sensitive to fees rather than liquidity. Liquidity clienteles explain the key features of ETF competition, including the first-mover advantage, the "winner-take-all" dynamics in trading volumes and the ability for incumbent ETFs to maintain higher fees. Empirical tests confirm the important role of liquidity clienteles and show that fee differentials for otherwise similar ETFs provide a novel measure of the value of liquidity to investors. Welfare analysis suggests that liquidity can be a source of monopolistic rents for ETF issuers.

Chapter 4 makes a methodological contribution by developing new measures of price discovery for sequential markets. The methodology accounts for the presence of noise in market prices, and hence allows us to study a new array of issues in modern market structure. Price discovery (the incorporation of new information into a security's price) is typically measured when a security trades simultaneously in multiple markets. The method proposed in this thesis extends the classic price discovery model of Hasbrouck (1995) to settings in which a security trades in consecutive phases (e.g., different market mechanisms or time zones) rather than in multiple markets. This approach allows information (efficient price innovations) to be separated from noise (microstructure frictions and liquidity) in each consecutive phase of trading. The Monte Carlo simulations confirm that the empirical estimation recovers correct Information Shares (IS), Noise Shares (NS), and Information-to-Noise ratios (IN). The method is computationally convenient, as it relies only on the output from vector autoregressive models (VARs). The proposed framework accounts for microstructure frictions in prices, and therefore produces more precise estimates of price informativeness compared to existing approaches.

Chapter 5 asks why so much trading has shifted towards the close of the trading day, and whether this tendency has made closing prices more informative. The empirical analysis shows that index investing, including ETFs, is by far the most important driver of trading on close. The price discovery results suggest that closing price informativeness has not improved with greater trading on close. The estimates rely on the novel price discovery methodology developed in the Chapter 4. The results reinforce policymakers' concerns that the increase in trading on close makes closing prices more vulnerable to dislocations.

Overall, this dissertation contributes to the academic and industry debate on the optimal market structure. The analysis of market-making OTTRs suggests that regulators should strike a balance between discouraging excessive quoting activity and encouraging competition between exchanges. The findings from ETF liquidity analysis imply that liquidity can be seen as a public good, with resulting "winner-take-all" externalities. The investigation of trading on close suggests that both market participants and regulators should recognize the potential disconnect between concentrated trading and price discovery. Although trading increasingly concentrates on close, price discovery still happens in continuous limit order books.

Working Papers and Presentations

Chapters 2–5 of this thesis have been concurrently developed as working papers, and presented at various academic conferences. The second working paper is joint work with Prof. Tālis Putniņš at University of Technology Sydney and Prof. Marius Zoican at University of Toronto Mississauga and Rotman School of Management. The list of working papers and conference presentations is below.

1. Khomyn, M. and Putniņš, T., 2017, *"Algos gone wild: Are order-to-trade ratios excessive?"*, Working paper, UTS Business School.
    - 2018 Financial Management Association (FMA) Annual Meeting. San Diego, USA.
    - 2018 Financial Management Association (FMA) Asia-Pacific Conference. Hong Kong, China.
    - 2017 Auckland Finance Meeting. Queenstown, New Zealand.
    - 2017 SIRCA Young Researchers Workshop. Melbourne, Australia.
2. Khomyn, M., Putniņš, T., & Zoican, M., 2019, *"The value of ETF liquidity"*, Working paper, UTS Business School.
    - 2019 Financial Intermediation Research Society (FIRS) Conference. Savannah, USA.
    - 2019 American Finance Association (AFA) Annual Meeting. Poster Session. Atlanta, USA.
    - 2018 Australasian Finance and Banking Conference. Sydney, Australia.
    - 2018 Financial Research Network (FIRN) Annual Conference. Brisbane, Australia.
    - 2018 Behavioural Finance and Capital Markets Conference. Melbourne, Australia.
3. Khomyn, M. and Putniņš, T., 2019, *"Measuring information and noise in sequential trading"*, Working paper, UTS Business School.
4. Khomyn, M. and Putniņš, T., 2019, *"The rise in trading on close: Drivers and effects on price formation"*, Working paper, UTS Business School.

- 2019 Australasian Finance and Banking Conference. Sydney, Australia.
- 2019 Financial Research Network (FIRN) Annual Conference. Byron Bay, Australia.
- 2019 Australian PhD Conference. Canberra, Australia.
- 2019 Behavioural Finance and Capital Markets Conference. Melbourne, Australia.
- 2019 International Accounting and Finance Doctoral Consortium. Milan, Italy.
- 2018 / 2019 Research Seminar Series, Baltic International Centre for Economic Policy Studies. Riga, Latvia.
- 2018 / 2019 Research Seminar Series, Kyiv School of Economics. Kyiv, Ukraine.
- 2018 / 2019 Research Seminar Series, Monash University. Melbourne, Australia.

Contents

Statement of Originality **i**

Acknowledgements **iii**

List of Figures **xiii**

List of Tables **xiv**

1 Introduction **1**
1.1 A non-technical introduction to market microstructure . . . 1
1.2 Why market structure matters . . . 4
1.3 What should we know about order-to-trade ratios? . . . 4
1.4 What should we know about exchange-traded fund liquidity? . . . 5
1.5 What should we know about trading on close? . . . 6
1.6 A brief guide to thesis chapters . . . 7

2 Are order-to-trade ratios excessive? **9**
2.1 Introduction . . . 9
2.2 Literature review . . . 13
2.2.1 How this chapter contributes to the literature . . . 13
2.2.2 What do order-to-trade ratios measure? . . . 14
2.2.3 Are high order-to-trade ratios harmful? . . . 15
2.2.4 What are the effects of messaging taxes? . . . 16
2.2.5 How does HFT affect market quality? . . . 17
2.2.6 Methods in existing HFT literature . . . 18
2.2.7 HFT and market fragmentation . . . 19
2.2.8 How does fragmentation affect market quality? . . . 20
2.3 A simple model of what drives the OTTR . . . 21
2.3.1 Baseline model structure . . . 21
2.3.2 Model with fragmented markets . . . 24
2.3.3 Propositions . . . 25

2.4 Empirical analysis . . . 29
2.4.1 Data and descriptive statistics . . . 29
2.4.2 Time series trends in OTTRs and concurrent market structure changes . . . 33
2.4.3 Cross-sectional and time-series determinants of OTTRs . . . 35
2.4.4 How OTTRs vary across markets . . . 39
2.4.5 The non-linear effects of fragmentation and market share . . . 41
2.5 A benchmark for monitoring OTTRs . . . 44
2.6 Conclusions and policy implications . . . 45
Appendix 2.1. Proofs . . . 48

3 The value of ETF liquidity 52
3.1 Introduction . . . 52
3.2 Literature review . . . 57
3.2.1 What are the effects of ETFs? . . . 57
3.2.2 How do investment funds compete? . . . 58
3.2.3 How much is liquidity worth? . . . 59
3.2.4 Why do liquidity externalities arise? . . . 61
3.3 Institutional details . . . 62
3.3.1 Creation-redemption process . . . 62
3.3.2 Liquidity . . . 63
3.3.3 Regulatory Structure . . . 64
3.3.4 Tax . . . 65
3.4 A model of ETF competition . . . 66
3.4.1 Model primitives . . . 66
3.4.2 Equilibrium . . . 69
3.4.2.1 Competitive market-making in the ETF market. . . 69
3.4.2.2 Investors' ETF selection . . . 70
3.4.2.3 Follower ETF fee-setting at $t = -1$ . . . 72
3.4.2.4 Leader fee-setting and entry deterrence . . . 73
3.4.3 Comparative statics and predictions . . . 77
3.5 Welfare implications . . . 79
3.5.1 Welfare benchmark . . . 80
3.5.2 Equilibrium welfare . . . 82
3.6 Empirical analysis . . . 84
3.6.1 Data and descriptive statistics . . . 84
3.6.2 OLS regression results . . . 85
3.6.3 Probit regression results . . . 89
3.6.4 Robustness checks . . . 90
3.7 Conclusions . . . 93
Appendix 3.1. List of ETFs . . . 94
Appendix 3.2. Proofs . . . 101

4 Measuring information and noise in sequential trading 103

4.1 Introduction . . . 103
4.2 Literature review . . . 106
4.2.1 Price discovery in one security and many markets . . . 107
4.2.2 Price informativeness in one security and one market . . . 108
4.2.3 Intraday / overnight return patterns . . . 110
4.2.4 Price discovery in sequential trading: challenges and methodological choices . . . 110
4.3 Model for separating information and noise in sequential markets . 112
4.3.1 Structural model of price formation . . . 112
4.3.2 Empirical estimation of the model . . . 117
4.4 Monte Carlo simulation evidence . . . 119
4.4.1 Model estimation with n=3 phases . . . 120
4.4.2 Model estimation with n=2 phases . . . 123
4.5 Illustrative application to intraday patterns in price discovery . . . 126
4.6 Conclusions and future research . . . 132
Appendix 4.1. Additional results from Monte Carlo simulations . . . 133

5 The rise in trading on close: Drivers and effects on price formation 141
5.1 Introduction . . . 141
5.2 Literature review and hypotheses . . . 145
5.2.1 Literature on closing auctions . . . 146
5.2.2 Informativeness of auction prices and continuous trading prices 147
5.2.3 Trading on close and passive investing . . . 148
5.2.4 Trading on close and dark / block trading . . . 149
5.2.5 Trading on close and HFT . . . 150
5.2.6 Trading on close and price discovery . . . 151
5.3 Institutional setting . . . 152
5.3.1 Trading on close in the US and Australian markets . . . 152
5.3.2 The microstructure of Australian markets . . . 153
5.4 Data . . . 155
5.5 Drivers of trading on close . . . 159
5.5.1 Descriptive analysis of trading on close . . . 159
5.5.2 2SLS analysis of trading on close . . . 161
5.5.2.1 Passive investing . . . 161
5.5.2.2 Block and dark trading . . . 164
5.5.2.3 High-frequency trading . . . 169
5.6 Price discovery analysis of trading on close . . . 173
5.6.1 Validation checks for the price discovery results . . . 177
5.7 Conclusions . . . 179
Appendix 5.1. Closing mechanism timeline in Australian and US markets 179
Appendix 5.2. Additional regression results for trading on close . . . 180
Appendix 5.3. Validation checks for the price discovery results . . . 185

6 Conclusions and Future Research 190

6.1 What drives the changes in how people trade? 191
6.1.1 How are market structure changes manifested in order-to-trade ratios? 191
6.1.2 How is liquidity priced in new financial products? 193
6.1.3 How informative are prices set by different trading mechanisms? 195
6.2 Work ahead 196
6.2.1 Towards market surveillance 2.0 197
6.2.2 Taking the pulse of ETF liquidity 197
6.2.3 Information and noise around the clock 198
6.2.4 Closing auctions, MiFID II, and price informativeness 198
6.2.5 . . . the game without end 198

Bibliography **200**

List of Figures

2.1 Time series of OTTRs and explanatory variables 34
2.2 Order-to-trade ratios and data processing speed 39
2.3 Standardized regression coefficients for the drivers of OTTRs 40
2.4 The relation between OTTRs, fragmentation, and market shares . . 43
2.5 Distribution of theoretical vs empirical OTTRs 46

3.1 Model timing 69
3.2 ETF entry, assets under management, and investor heterogeneity . 77
3.3 Equilibrium comparative statics 80
3.4 Welfare loss in the oligpolistic equilibrium 83
3.5 Management expense ratios vs liquidity of same-index ETFs 86

4.1 Observed prices and returns in sequential phases 112
4.2 Correspondence between observed returns with three vs two sequential phases 124
4.3 Intraday patterns in the US market, 2018 130
4.4 Intraday patterns in the US market, 2002 – 2018 131

5.1 Trading on close and passive ownership (AU) 162
5.2 Trading on close and dark trading 166
5.3 Trading on close and block trading 168
5.4 Trading on close and HFT 171
5.5 Information shares and noise shares for the Australian market . . . 175
5.6 Information shares and noise shares for the US market 177
5.7 Closing mechanisms in AU and US markets 180
5.8 Information shares and noise shares, with shifting pre-close time . . 186
5.9 Realized variances in three phases of the trading day 187
5.10 Proportion of zero-return observations in three phases of the trading day 188
5.11 Overnight return reversals 189

List of Tables

2.1 Summary of propositions and empirical hypotheses 26
2.2 Variable definitions 31
2.3 Descriptive statistics 32
2.4 Determinants of OTTRs using stock-day observations 36
2.5 Determinants of OTTRs using exchange-day observations 41
2.6 Empirical OTTRs vs a theoretical benchmark 45

3.1 Model notation 68
3.2 Descriptive statistics 85
3.3 Cross-sectional regressions 88
3.4 Probit regressions 91
3.5 Robustness checks 92

4.1 Estimates of information shares from the model with $n = 3$ phases . 120
4.2 Estimates of noise shares from the model with $n = 3$ phases 121
4.3 Estimates of information-to-noise ratios from the model with $n = 3$ phases 122
4.4 Estimates of information shares from the model with $n = 2$ phases . 125
4.5 Estimates of noise shares from the model with $n = 2$ phases 126
4.6 Estimates of information-to-noise ratios from the model with $n = 2$ phases 127
4.7 Estimates of information shares from the model with $n = 3$ phases . 135
4.8 Estimates of noise shares from the model with $n = 3$ phases 136
4.9 Estimates of information-to-noise ratios from the model with $n = 3$ phases 137
4.10 Estimates of information shares from the model with $n = 2$ phases . 138
4.11 Estimates of noise shares from the model with $n = 2$ phases 139
4.12 Estimates of information-to-noise ratios from the model with $n = 2$ phases 140

5.1 Variable definitions 156
5.2 List of passive funds with significant exposure to AU stocks 158
5.3 Descriptive statistics 159
5.4 Stage 1 – 2SLS regression results for passive investing (AU) 164
5.5 Stage 2 – 2SLS regression results for passive investing (AU) 165
5.6 Stage 1 – 2SLS regression results for dark and block trading (AU) . 169

5.7 Stage 2 – 2SLS regression results for dark and block trading (AU) . 170
5.8 Stage 1 – 2SLS regression results for HFT (AU) 172
5.9 Stage 2 – 2SLS regression results for HFT (AU) 173
5.10 Information shares and noise shares results 174
5.11 Time series regression results for trading on close (AU) 181
5.12 Time series regression results for trading on close (US) 182
5.13 Cross-sectional regression results for trading on close (AU) 183
5.14 Cross-sectional regression results for trading on close (US) 184
5.15 Stage 2 – 2SLS regression results for passive investing (AU) 185

Chapter 1

Introduction

The stock market is now run by computers, algorithms, and passive managers.
The Economist (Print edition, October 5, 2019)

1.1 A non-technical introduction to market microstructure

On an average trading day, $356 billion changes hands in US equity markets. This is about the size of annual economic output (GDP) of the Philippines. Unlike goods and services produced in the Philippines, however, money on trading screens seems of little consequence to common citizens. But that is misleading. Trading screens become of common interest, when a crisis strikes. For example, in the 2007–2009 recession, the financial crisis wiped out $19.2 trillion of household wealth, and 8.8 million Americans lost their jobs. Finance gets its publicity in bad times, when markets leave a trading room for a living room. Finance, however, can work in either direction: exacerbate recessions in poor economic conditions, or equally well fuel economic growth when conditions are favorable.

The good news is rarely newsworthy; but the US economy added over 2 million jobs between 2010 and 2018. What was the role of finance in that achievement? Financial markets facilitated the creation of goods and services — a modest role, yet critical to economic recovery. The role of finance is to make it easier to exchange goods and services for money: to allocate resources, monitor investments, mobilize

savings, simplify transactions and distribute risks to those best positioned to bear them (Levine, 2005).

Economists study how businesses use scarce resources efficiently to satisfy people's unlimited needs. Finance examines the merits of diverse business ideas and allocates funds to the most promising ones. In simple terms, financial markets provide a mechanism for investors to vote with their money. The prices of financial assets reflect the average vote of a multitude of investors — a more accurate estimate of true value than if the prices were dictated by a panel of experts.

If an economy is like a giant factory producing everything from tomatoes and tea bags to smartphones and haircuts, then finance is like a network of roads and wires connecting different parts of the factory. Via well-functioning roads, tomatoes can be exchanged for smartphones, and teabags can be delivered to hair salons, without creating prolonged shortages in any part of the factory. Wires connect all parts of the factory to enable information flow about the rates of exchange between tomatoes and teabags, and thus the factory produces the optimal number of each.

The design of roads and wires matters, because even a small fault in the communication network can cause disturbances to traffic and impair information about the rates of exchange. If finance operates all roads and wires in the factory of economics, then market microstructure handles engineering choices. How do we design the roads to withstand increased traffic? How do we ensure that wires deliver information efficiently, without excessive noise? How do we make the network resilient to poor weather and storms? The role of market microstructure, in short, is to enable the seamless flow of information and liquidity in financial markets.

This thesis investigates the microstructure of modern financial markets, which are, according to The Economist, "run by computers, algorithms, and passive managers." Following the earlier analogy, financial markets' roads and wires have undergone a dramatic change. In this new reality, the challenge is to optimize the system so that it works reliably and efficiently. The four chapters in this thesis address different aspects of this challenge.

Chapter 2 sheds new light on liquidity provision that relies on unparalleled data speeds. Nowadays, markets represent a network of servers communicating electronic messages ultra-fast. Therefore, modern market making is quite different from traders interacting through human-input electronic platforms. In these conditions, old measures of market misconduct (e.g., order-to-trade ratios) should be

re-examined. Chapter 2 shows how to assess whether order-to-trade ratios are excessive or not. The insights from this chapter can help financial market regulators to distinguish between legitimate market making and illicit quoting activity.

Chapter 3 zooms in on the value of liquidity. It shows that new financial products, such as exchange-traded funds, operate in "winner-take-all" markets, akin to technology markets dominated by Facebook and Google. Facebook is more attractive to each user, the more other users are already on the network. Similarly, the S&P 500 ETF (SPY) is traded more often, the more high-turnover investors are already using it. This allows SPY to charge monopolistic fees. More broadly, Chapter 3 shows that competition between exchange-traded funds is subject to self-perpetuating liquidity cycles called network externalities.

Chapter 4 recognizes that modern markets are not without friction. Although the selection of available trading mechanisms is wider than ever, market participants' excessive reliance on any one mechanism can generate noise in security prices. For example, the recent trend has been to trade in closing call auctions more than ever before. This shift in volumes can add noise to prices. On the other hand, it can also add more informed voices to the trading chorus. To understand which effect dominates, Chapter 4 develops a novel price discovery methodology. The methodology separates information from noise in sequential trading phases.

Chapter 5 examines why market participants increasingly trade in closing auctions. The price generated in the closing call auction has always been the most important price of the day, as it is used to value many financial products, compute daily returns, and assess asset managers' performance. However, the recent shift to trading on close has been unprecedented. In the past ten years, volumes on close increased four to five times, now accounting for as much as 40% of daily trading in some markets. Chapter 5 finds that this is due to passive managers, who seek to buy and sell stocks exactly at closing prices. Because passive managers just replicate the index rather than look for under- or overvalued stocks, their trading does not contribute much information. This is confirmed by price discovery analysis: closing prices are rather noisy, and information is mostly incorporated through intraday trading rather than through trading on close.

1.2 Why market structure matters

Why is it important to study the microstructure of financial markets? Following the metaphor from the previous section, why should we care about the internal workings of roads and wires of finance? The simple answer is that it is too costly to ignore. AQR, one of the world's largest asset managers, estimates that execution costs are 12 bps[1] for each dollar traded.[2] If investors trade $356 billion daily, they pay $427 million in trading costs. In other words, the *daily* amount spent on trade execution in US equity markets is comparable to the *lifelong* savings of 2,000 retirees.

This thesis contributes to optimizing the roads and wires of financial markets, so that they serve the end investor well. Market structure research seeks to assess whether the trading costs of $427 million a day are justified, whether they can be lowered, and which regulatory measures are suitable to do so. The findings in this thesis can help financial regulators tailor the market design so that it maximizes the value to society rather than benefits some market participants at the expense of others. The findings can also help end users of financial services, such as owners of pension accounts, to save on trading fees. Ultimately, this thesis speaks to the overarching question of market microstructure: how to achieve a fair and efficient market that serves the social good of facilitating exchanges in the real economy.

1.3 What should we know about order-to-trade ratios?

"The US financial markets had always been either corrupt or about to be corrupted," according to the book "Flash Boys" by Michael Lewis. This quote is related to high-frequency trading (HFT) and the way financial market integrity has been arguably compromised by a subset of extremely fast traders. Indeed, when headlines come out, claiming that due to HFTs, "96.8% of all orders are cancelled before they trade", the public sentiment is that markets must be rigged.[3]

[1] 12 bps (basis points) equals 0.0012%.

[2] Assuming they trade 1% or less of total daily volume. If I assumed 5% of total daily volume, the cost would be 25 bps (Frazzini, Israel, & Moskowitz, 2018).

[3] See Quartz (2013) article here: https://qz.com/133695/96-8-of-trades-placed-in-the-us-stock-market-are-cancelled/.

The problem with this sensationalist approach is that it uses "HFT" as an all-purpose term for a set of diverse trading strategies with very different effects on market quality.

What Chapter 2 studies is the reality of modern market making, which is dominated, unsurprisingly, by HFTs. Market making, unlike many other HFT strategies, provides a useful service to market participants. Because of fast data feeds, it's HFTs that are best positioned to provide liquidity in fragmented markets without being "picked off" (trading at stale prices). Chapter 2 of this thesis models order-to-trade ratios (OTTRs), the number of order submissions, cancellations, and amendments divided by the number of trades. Regulators often use OTTRs as a proxy for HFT activity, as well as for detecting illicit behavior such as spoofing. Because of this dual role of OTTRs, regulators might misinterpret high OTTR levels as evidence of illicit activity. This study can help draw the distinction between OTTR levels warranted by market making and those that spike beyond such levels.

The model of OTTRs developed in Chapter 2 and calibrated to the data provides a regulatory tool for assessing OTTR levels observed in financial markets. It addresses the following question: "How do observed OTTR levels compare to the OTTRs that we'd expect from market-making activity in certain volatility conditions, given certain stock characteristics?" This question is important for market surveillance, as it helps disentangle legitimate trading from illicit conduct such as layering and spoofing.

1.4 What should we know about exchange-traded fund liquidity?

"On balance, the financial system subtracts value from society", said John Bogle, the pioneer of passive investing and founder of Vanguard. He argued that investors should turn to passive funds to save on fees. Investors seem to agree: in September 2019, Morningstar reported that passive assets under management surpassed active. An especially fast-growing sub-class of passive funds are index-tracking exchange-traded funds (ETFs), which trade intraday like common shares. Bogle famously declared the high volumes in ETFs "investor's enemy", claiming that their rapid trading doesn't do any social good. The question is, how much,

exactly, are investors paying for accessing ETF liquidity? This question is the subject of Chapter 3 of this thesis.

Chapter 3 shows that competition between ETFs involves a "winner-take-all" dynamics. For example, 50% of all ETF trading is concentrated in the top 15 ETFs (there are over 2,000 US equity ETFs overall). In general, high volumes should not matter to ETF issuers, who make money by charging a fixed percentage fee of assets under management, regardless of how much an ETF trades. However, as Chapter 3 shows, when two ETFs track the same index, the more liquid one typically charges higher fees. This dynamic arises due to liquidity clienteles: high-turnover investors (e.g., institutions) are attracted to the most liquid ETFs, thereby making them more liquid still, and allowing the ETF issuers to charge higher fees in equilibrium. Low-turnover investors (e.g., buy-and-hold retail investors) are more sensitive to the fee and therefore hold low-fee ETFs, which in turn are less liquid due to lower investor turnover. Liquidity clienteles also explain the key features of ETF competition, including the first-mover advantage and the ability for incumbent ETFs to maintain higher fees, despite competition from low-fee funds.

1.5 What should we know about trading on close?

The shift to passive investing is also apparent in the distribution of trading volumes during the day. As passive funds (which seek to minimize their tracking errors against the closing price) receive more inflows, they account for an increasing proportion of trading, most of which is done at or around the market closing time. In Europe's major equity markets, as much as one-quarter to one-half of the entire day's trading volume is now executed at the market close. This is twice the level a mere three years ago, with the US and Australia following similar trends of increasing trading on close. Given that bursts of intense trading activity are associated with increased volatility and large price movements, the concentration of trading in a small window around the close can potentially harm the quality of closing prices. The more trading that occurs in this short window of time, the higher the risk of large order imbalances that can create temporary price swings and distort closing prices.

The quality of closing prices matters, as they serve as benchmarks for portfolio valuation (e.g., in the US, $4.3 trillion passive assets under management are valued using closing prices), derivative pricing, capital budgeting and more. In academic research, most asset pricing and corporate finance studies use closing prices to compute daily returns. For all these applications, it is imperative to understand the relative amount of information vs noise in closing prices. Chapter 4 of this thesis develops a novel empirical methodology to do so. Using Monte Carlo simulations and validation tests, this research confims that the methodology reliably identifies the amount of information and noise in prices at different times throughout the day.

Chapter 5 of this thesis investigates the drivers and effects of increased trading on close. What is driving the strong demand to trade on close? Is the tendency harming closing prices by making them more prone to temporary price pressures? Or is the concentration of trading making closing prices more informative? How is the continuous trading session impacted by trading activity shifting away? Has the rest of the day become "meaningless" for price discovery, as suggested by some market participants? These questions are important to regulators, who are now considering shortening the trading day, as well as for end users of closing prices — academics, investors, and financial markets practitioners.

The results suggest that trading on close is driven mostly by passive investing. Stocks in major indices followed by passive funds experience significantly larger trading on close, compared to similar stocks outside such indices. This has important effects on price formation: closing prices are not contributing much information, while the intraday session remains an important venue for price discovery.

1.6 A brief guide to thesis chapters

This thesis is structured around four questions on modern market structure:

i Are order-to-trade ratios excessive? (Chapter 2)

ii How much is ETF liquidity worth? (Chapter 3)

iii How to measure price discovery in sequential trading phases? (Chapter 4)

iv What are the drivers and effects of increased trading on close? (Chapter 5)

Each thesis chapter has a section dedicated to related literature. The literature sections review existing research and highlight the contribution of this thesis. Chapter 6 draws conclusions and discusses future directions for related research.

Chapter 2

Are order-to-trade ratios excessive?

Who is Wall Street? Think Goldman Sachs, JP Morgan and Morgan Stanley drive today's equity markets? Fuhgeddaboudit. Today's largest trading firms are Citadel Securities, GTS, HRT, IMC, Susquehanna/G1X, and Virtu.

Larry Tabb, founder of TABB Group.

2.1 Introduction

In today's markets, liquidity is provided almost entirely by electronic market makers, operating across fragmented markets, and equipped with fast data feeds and smart market monitoring technologies. One manifestation of electronic market making is high order-to-trade ratios (OTTRs: number of enter/amend/cancel messages divided by the number of trades). OTTRs have increased more than ten-fold since the year 2000 (Committee on Capital Markets Regulation, 2016). In 2013, the US Securities and Exchange Commission (SEC) reported that 96.8% of all orders were cancelled before they traded, with 90% being cancelled within one second (US SEC, 2013). The response of policymakers has been to curb OTTRs by imposing messaging taxes, which have been proposed in some countries (such as the US) and already implemented in others (e.g., Australia, Italy, and Germany). This chapter investigates the drivers of OTTRs, whether their growth warrants concern, and the impacts of regulatory proposals such as messaging taxes.

High OTTRs have been in the public spotlight, with concerns that they are a symptom of predatory or manipulative behavior of high-frequency traders (Biais,

Foucault, & Moinas, 2011). These concerns are accompanied by a recent surge in prosecutions of spoofing, a trading strategy that involves misleading other market participants with orders that are cancelled before they are able to execute. For example, the US Commodity and Futures Trading Commission (CFTC) in 2018 took action against a record number of spoofing cases: five times more than the average in the previous eight years.[1] It is true that market manipulation strategies such as spoofing, layering, or quote stuffing often generate spikes in quoting activity and high cancellation rates (e.g., Egginton, Van Ness, & Van Ness, 2016). Therefore high OTTRs are used by surveillance systems to identify spoofing in markets, and they have been used in courts as evidence of spoofing. However, high OTTRs can also arise from trading strategies that are neither illegal nor harmful, making it crucial to understand how to distinguish between legitimate and illegitimate levels of OTTRs. In fact, as this chapter shows, market making can result in high OTTRs, in particular when posting quotes across multiple trading venues and adjusting the quotes rapidly in response to new information to minimize picking-off risk. The combination of advances in technology and fragmentation of trading requires more quote revisions for market makers to remain competitive.

Regulators' concerns about high OTTRs are two-fold: firstly, high OTTRs can be a symptom of illicit trading activity, and and secondly, high OTTRs increase the regulatory costs of monitoring the large volumes of market data. As a result, a number of regulators have imposed message taxes that charge high-OTTR traders a fee (Friedrich & Payne, 2015). If such regulation curbs harmful HFT behavior, the tax could improve liquidity and other measures of market quality. However, if the regulation negatively affects market makers by constraining legitimate liquidity provision, market liquidity could deteriorate. Therefore the optimal design of message traffic taxes requires an understanding of the levels and variation in OTTRs that result from legitimate market making.

To understand the drivers of the high OTTRs in today's markets and to provide a benchmark against which to evaluate OTTRs, I develop and calibrate a simple model of liquidity provision. In the model, a market maker in a fragmented market monitors several sources of information ("signals") and updates quotes to avoid being "picked off" (trading at stale prices). His monitoring intensity is endogenous: the market maker decides how many and which signals to monitor by weighing

[1] See the The Wall Street Journal (2018) here: https://www.wsj.com/articles/u-s-market-manipulation-cases-reach-record-1540983720.

up the benefit (reduced "picking-off risk") against the cost (the computing and data feed costs). Consequently, the OTTR emerges endogenously as a function of monitoring cost, market conditions (e.g., volume, volatility), stock characteristics (e.g., how closely correlated the stock is with other securities), and the extent of fragmentation of trading across multiple trading venues.

This endogenous OTTR, driven by a multitude of factors, makes it possible to investigate why OTTRs have grown so much in recent years and what explains their time-series variation. The empirical tests of the model suggest that the long-term growth in OTTRs is largely driven by increasing market fragmentation and decreasing monitoring costs, while short-term dynamics can be explained by market volatility. For example, while the average US stock at the start of the sample (year 1998) was traded on around two stock exchanges, by 2018, that number had increased to ten. Microprocessor speeds increased 14-fold during the sample period and storage costs fell by a factor of 2,000, collectively reflecting a substantial reduction in the costs of monitoring a large number of digital signals. Yet around these long-term trends, I find that OTTRs spike around the same time as the VIX index, consistent with those being periods of large changes in security values and therefore high picking-off risk.

The model also sheds light on why OTTRs are considerably higher in some securities compared to others. The empirical tests support the model predictions that OTTRs are higher in more volatile stocks, higher price-to-tick stocks, lower volume stocks, and in ETFs compared to stocks. The intuition for these effects is as follows. In more volatile markets, information arrives more frequently, causing market makers to update (amend, or cancel and replace) quotes more often to avoid being picked off by informed traders. This increases OTTRs. Similarly, market makers update quotes more often in high price-to-tick stocks (stocks with a small relative tick size), because the greater granularity in prices implies that even very small changes in valuations will result in quote updates. This result corroborates the findings of Yao & Ye (2018), albeit through a different channel. ETFs naturally have higher OTTRs compared to stocks due to a number of highly relevant signals available for monitoring, such as the prices of the underlying stocks or the index.

The third dimension of variation in OTTRs explained by the model is across markets. I find that OTTRs are naturally much higher on markets with lower market

shares. The intuition for this effect is as follows. When a security trades on multiple markets, a liquidity provider often duplicates their quotes in each (or at least some) of the different markets. Each time the market maker updates their quoted prices or volumes, they send order amendment or cancellation messages to all of the markets in which they are quoting, and consequently the amount of messaging activity (the OTTR numerator) across the different markets is approximately equal or at least at a similar level. Yet if the number of trades in a market with low market share is considerably lower, the smaller OTTR denominator results in a higher OTTR for that market.

Given the characterization of the large number of factors that drive variation in OTTRs, in the last part of the chapter I apply a calibrated version of the model to assess how often OTTRs spike to levels that could be deemed in excess of regular market making. Applying the model to the most recent period (2018) of the sample reveals that in most cases, empirically observed OTTRs are in line with or below those that would be expected from liquidity provision in a fragmented market, even under conservative assumptions. The distribution of empirical OTTRs is right-skewed, with 19% of observations above the theoretical level. Benchmarking actual OTTRs in this manner against the expected legitimate OTTRs that account for the drivers of OTTRs is an approach regulators could use to detect abnormal quoting activity and penalize illicit behavior in cases that actually require intervention.

The findings suggest that the recent levels of OTTRs in most cases do not necessarily warrant concern, as legitimate market making results in OTTRs that are similar or above those observed in the market data. Therefore, regulatory measures aimed at curbing quoting activity (e.g., messaging taxes) can have adverse effects on market making in securities that already have disadvantageous conditions for market makers. Furthermore, messaging taxes create a non-level playing field between trading venues, because venues with a lower share of total volume naturally have higher OTTRs. Finally, securities with precise, frequent signals (e.g., ETFs that trade as a basket of underlying stocks) are expected to have higher OTTRs compared to individual stocks, so taxing market makers equally across securities on a per message basis is likely to disproportionally harm ETF liquidity provision.

This chapter proceeds as follows. Section 2.2 reviews relevant literature. Section 2.3 develops a model of market-making OTTRs, and Section 2.4 empirically tests

the model predictions. Section 2.5 provides a benchmark for evaluating whether OTTRs are excessive. Section 2.6 concludes the chapter.

2.2 Literature review

The current trading landscape is shaped by two interrelated forces: competition (including regulatory measures that promote competition, such as Reg NMS and MiFID) and technology (e.g., faster data processing and data transmission speeds). Literature on competition (i.e., fragmentation of trading) studies trade-through rules, best execution requirements, smart order routing, maker/taker fees, consolidated data feeds, consolidated data feeds, latency, cross-market arbitrage etc. Literature on the effects of technology explores high-frequency trading (HFT), co-location and direct market access, flash crashes, endogenous market makers, new order types, tick sizes, and dark trading. Order-to-trade ratios reflect the interplay between HFT, fragmentation, and regulatory changes, and hence are the subject of study of several strands of literature.

2.2.1 How this chapter contributes to the literature

Technological improvements have enabled faster speeds of trading, while electronically consolidating liquidity in a fragmented market. The nature of market making changed with the rise of endogenous market makers (O'Hara, 2015). Liquidity provision is now undertaken mostly by HFT firms, which have a speed advantage in rapidly adjusting their quotes to reflect the most recent information across correlated securities, as well as across multiple trading venues. Building on prior literature of HFT market making, this study highlights that the way liquidity is provided in fragmented markets has a direct effect on order-to-trade ratios (OTTRs). Going beyond existing studies, this analysis offers a method to scrutinize observed OTTR levels, taking into account the stock characteristics and prevailing market conditions. The findings also highlight a largely overlooked link between messaging taxes and competition between trading venues.

The simple theory model in this chapter offers a stylized way to examine OTTRs in view of market making activity.The chapter does not seek to make a major contribution to theoretical market microstructure literature, but rather helps draw

a link between OTTRs and market making, which has been mostly overlooked by earlier studies. The theory builds on models of market making with costly monitoring. Similar to Foucault, Röel, & Sandås (2003) and Liu (2009), this theory model considers the trade-off between market makers' monitoring cost and picking-off risk, but it also recognizes fragmentation of trading across different venues. This research differs from Liu (2009) in that it focuses on the drivers of OTTRs, rather than the relation between spreads and limit order revision / cancellation activity. This study also relies on extensive empirical tests, incorporating the complexity of trading in fragmented markets, with decreased cost of monitoring driven by technological advancements (Lyle & Naughton, 2018), and maker-taker fee structures (Foucault, Kadan, & Kandel, 2003).

The empirical findings are in line with Yao & Ye (2018) and Wang & Ye (2017) in that order-to-trade ratios are not driven solely by HFT activity. However, the empirical analysis also investigates additional factors driving OTTRs, not just price vs speed competition in liquidity provision. Similar to this research is Rosu, Sojli, & Tham (2020), who also explain OTTRs using various stock characteristics. Unlike Rosu et al. (2020), I do not study OTTRs in relation to liquidity, price discovery or expected returns, but rather focus on evaluating market-wide levels of OTTRs, as well as stock-level OTTRs. A related paper by Dahlstrom, Hagstromer, & Norden (2018) explains cancellations with the order book variables that determine the limit order profitability. While they focus on when and why orders are cancelled, I look more broadly at what drives high levels of quoting activity and whether high order-to-trade ratios can be justified by market making.

2.2.2 What do order-to-trade ratios measure?

Academics, stock exchanges and regulators often use order-to-trade ratios as a proxy for high-frequency trading. For example, the US Securities and Exchange Commission, the US Congressional Research Service, UK Government Office of Science, and the European Securities and Market Authority rely on OTTRs in their HFT policies. Brogaard, Hendershott, & Riordan (2014) report that some stock exchanges (e.g., NASDAQ) use OTTRs to classify HFT traders. Importantly, regulatory initiatives aimed at curbing HFT activity are usually tied to OTTRs. Chung & Lee's (2015) survey mentions messaging taxes levied in proportion to OTTR in Italy, France and Norway. Germany launched an HFT regulation in 2013,

aiming to decrease OTTRs by HFT firms. Australian and Canadian regulators' cost recovery programs are also based on charging messaging taxes.

However, a number of recent academic studies (Rosu, Sojli, & Tham, 2020; Yao & Ye, 2018; Wang & Ye, 2017) have cast doubt on the merits of OTTR as a proxy for HFT activity. Rosu et al. (2020) show that equilibrium OTTRs reflect a number of factors beyond HFT activity, including an asset's risk-bearing capacity, dealer's inventory, cost of monitoring and monitoring precision. Yao & Ye (2018) provide empirical evidence that order-to-trade ratios are negatively related to HFT liquidity provision in a cross-section of stocks. Wang & Ye (2017) offer a theory model that explains this effect: due to their speed advantage, HFTs post relatively higher fraction of liquidity in high tick-to-price stocks, and once their queue priority is secured, they are less likely to cancel orders. In stocks with lower tick-to-price, there are fewer HFTs, but all market makers compete more on price than time priority, and hence cancel more orders.

If order-to-trade ratios are not necessarily a good proxy for HFT activity, then why are they used by regulators to tax HFTs? Perhaps simply due to the absence of better proxies. However, in that case, it is instrumental for regulators to understand the determinants of OTTR, as messaging fees might be misdirected and harmful to market-making activity.

2.2.3 Are high order-to-trade ratios harmful?

There is no convincing evidence that greater OTTRs are detrimental to market quality. Among empirical studies, Conrad, Wahal, & Xiang (2015) find the opposite: that higher quotation activity (defined as number of price and quantity changes at best quotes) is associated with prices being closer to random walk, and costs of trading being lower. They also show that securities with higher quote updating display lower price impacts after drawdowns in liquidity. Dahlstrom, Hagstromer, & Norden (2018) show that high order cancellation rates are more consistent with market making than with liquidity taking strategies.

Why the are high OTTRs subject to regulatory scrutiny? Because abnormally high OTTR levels might be indicative of illicit activity (by either HFTs or non-HFTs) or harmful HFT strategies. For example, high OTTRs are associated with greater variance ratios in quotes (Hasbrouck, 2018) and higher noise-to-information ratios

in order flow (Yueshen, 2017). Certain market manipulation strategies also generate high OTTRs, including spoofing (Biais, 2011) and quote stuffing (Egginton, van Ness, & van Ness, 2016).

2.2.4 What are the effects of messaging taxes?

The literature addressing the effects of regulatory restrictions on excessive order submissions and cancellations generally finds negative or neutral effects of messaging taxes on liquidity and market quality. For example, van Kervel (2015) provides evidence from a sample of ten FTSE 100 stocks that imposing a cancellation fee discourages competition among trading venues and harms liquidity.

A number of studies investigate the effects of messaging taxes introduced in European countries in 2012. Caivano et al. (2012), Friedrich & Payne (2015), and Capelle-Blancard (2017) study the effect of taxing traders with excessive OTTRs (above 100:1) on Borsa Italiana (Italy's largest stock exchange). The former two studies find the tax to be detrimental to market quality, while the latter study finds no effect (in the time span of three years). Similarly, Colliard & Hoffmann (2017) find no effect on market quality from the French messaging tax. Jorgensen et al. (2018) find that the Norwegian messaging tax (imposed on traders with OTTR above 70) had no harmful effect on the stocks in the treatment group, as relative spreads decreased slightly, while depth and turnover did not change. In Germany, Gomber & Haferkorn (2015) find that the price dispersion across trading venues has increased after implementation of the German HFT Act, which charges HFTs based on their OTTRs. The Canadian regulator (IIROC) imposed a messaging tax as part of its cost recovery program, and charges traders proportionally to their share of submitted messages. Malinova et al. (2013) show that these measures increased quoted and effective spreads in the Canadian market. Similarly, Lepone & Sacco (2013) find that IIROC's cost recovery program coincided with deterioration in liquidity on Chi-X Canada.

While the studies mentioned above provide empirical evidence on negative to neutral effects of messaging taxes, they do not address two relevant concerns. Firstly, they do not offer formal theoretical models for why taxing messaging activity is harmful to liquidity. Secondly, they do not investigate the heterogeneity of these effects in the cross-section of stocks.

I address the first concern by developing a simple theory model that considers liquidity provision in fragmented markets and can be readily calibrated to the data. I also test the model predictions empirically and infer the extent to which order-to-trade ratios might be excessive (and hence need to be curbed through taxes) or the opposite – insufficient (and hence need not be taxed).

I address the second concern by investigating which stock and market characteristics can help explain the variation in OTTRs. This matters for the effect of messaging taxes, as they could negatively affect some securities (e.g., those that naturally tend to have high OTTRs), but not others (e.g., those with naturally low OTTRs). This also matters for providing a level playing field for competing trading venues, as messaging taxes could disadvantage smaller trading venues, which have higher OTTRs even in the absence of greater HFT activity.

2.2.5 How does HFT affect market quality?

High-frequency trading refers to a variety of proprietary trading strategies that aim to make profit from their speed advantage. HFTs are characterized by relatively short holding periods and zero overnight positions. They typically specialize in one of the following trading strategies: endogenous liquidity provision, cross-market arbitrage, statistical arbitrage, predatory trading, or news-based strategies (e.g., trading around earnings announcements, macroeconomic news etc.). Most market observers agree that HFT activity started growing in the early 2000s, peaked around year 2009, and has levelled off since then (Hendershott, Jones, & Menkveld, 2011). The recent period (2015–2018) has witnessed a wave of consolidation in the HFT industry, which is indicative of lower profits. However, HFT still accounts for a substantial share of volumes: 25% in Australia, 27% in the UK, 42% in US large capitalization stocks (Brogaard et al., 2019), and 46% in Canada (Boehmer, Li, & Saar, 2018).

Multiple studies investigate the impact of HFTs on various aspects of market quality (liquidity, volatility, price discovery, institutional execution costs, liquidity co-movement etc.). The results are mixed (Jones, 2013; Menkveld 2016). Theoretically, HFT can increase or decrease adverse selection risks depending on whether fast algorithms predominantly provide liquidity with limit orders or take liquidity with market orders. For example, if an HFT arbitrage algorithm uses market orders, it imposes adverse selection risk on slower traders. Foucault et al. (2016)

find that spreads increase by 4% for each 1% increase in the likelihood of HFT "toxic arbitrage". On the other hand, endogenous HFT market makers are faster than non-HFT, and therefore face lower adverse selection risks (i.e., their limit orders are less likely to be "picked off" by other traders), which means increased liquidity provision. Additionally, HFT market makers are likely to have better inventory control algorithms and therefore lower inventory holding costs. Being technology-intensive, they also face lower fixed order processing costs. Hence, in theory, HFT market makers should decrease the three key components of spreads: fixed order processing costs, inventory holding costs (Ho & Stoll, 1981), and adverse selection costs (Kyle, 1985; Glosten & Milgrom, 1985). Empirical evidence in Hendershott, Jones, & Menkveld (2011), Hasbrouck & Saar (2013), Menkveld (2013), Brogaard, Hendershott, & Riordan (2014), Boehmer, Fong, & Wu (2013), Broggard, Hagstromer, Norden, & Riordan (2015), and van Kervel (2015) suggests that the net effect of HFT is narrower spreads and better price discovery, on average.

Narrower average spreads, however, are not all-encompassing evidence of improved liquidity. Several studies find that HFT activity leads to more fragile liquidity in market downturns (Anand & Venkatarman, 2016), higher commonality in liquidity (Malceniece, Malcenieks, & Putnins, 2019), higher execution costs for large trade packages (Menkveld & van Kervel, 2019), and increased short-term volatility (Hasbrouck & Saar, 2013). On the other hand. Putnins & Barbara (2020) show that HFTs are no more likely than non-HFTs to engage in toxic trading that increases institutional transaction costs.

2.2.6 Methods in existing HFT literature

How is HFT liquidity provision modeled in market microstructure literature? Three common themes are present in the theory models on HFT: (i) adverse selection risk (i.e., picking-off risk), (ii) monitoring intensity, and (iii) competition among market makers. Models of fast and slow traders treat the fast as a source of additional adverse selection in fragmented markets. In van Kervel (2015), adverse selection costs arise due to fast traders being able to observe the order flow before the slow traders. Models with monitoring intensity allow market makers to choose monitoring intensity that allows them to avoid being picked off by other traders (Foucault, Roell, & Sandas, 2003), or to be the first to market among competing

market makers (Foucault, Kadan, & Kandel, 2013). Market makers' monitoring leads to order submissions and cancellations in response to new information arrivals (Liu, 2009). In the spirit of earlier studies, the model in this thesis considers the trade-off between costly monitoring intensity and picking-off risk.

How do finance researchers identify HFTs? Empirical papers commonly use one of the following identification strategies: (i) exogenous entry of HFTs (Brogaard & Gariott, 2015; Malceniece, Malcenieks, & Putnins, 2019), (ii) explicit HFT identifiers (van Kervel & Menkveld, 2019; Goldstein, Kwan, & Philip, 2018), (iii) OTTR as a proxy for high-frequency trading (Malinova, Park, & Riordan, 2016; Hoffman, 2014; Conrad, Wahal, & Xiang, 2015; Brogaard, Hendershott, & Riordan, 2015; Subrahmanyam & Zheng, 2016). Hendershott, Jones, & Menkveld (2011) argue that order-to-trade ratios are indicative of the degree of automation. Depending on the data available, one can use the following characteristics to classify traders as HFT: high position turnover, short holding time, high frequency of order amendments and cancellations, high Sharpe ratio of intraday profits, zero overnight inventory, and high intraday roundtrip volume. The OTTR analysis in this thesis is not concerned with identifying HFTs, but benefits from earlier findings on common HFT-related stock and market characteristics.

2.2.7 HFT and market fragmentation

Financial market regulations in the US (Reg NMS), Canada (the ATS regime) and Europe (MIFID) allowed competition between trading venues in the mid-2000s. Trading venues compete on market structure: latency, transparency, fee structures and order types. In the US, trading fragmented across ∼13 lit markets and ∼44 alternative trading systems, but remained virtually consolidated via the order protection rule and smart order routing. Similarly, in Europe, the primary exchanges lost significant market share to MTFs (multilateral trading facilities, which can be both lit and dark). New trading venues (e.g., Chi-X, Cboe, Turquoise, Posit, Instinet) made it possible for traders to co-locate with the exchange servers, enabling faster trading and data feeds.

Menkveld (2014) argues that fragmentation and HFT are closely related and therefore should be modeled jointly. The following concurrent effects may arise between HFT and fragmentation: (i) both HFT and fragmentation are driven by technological advancements (which enables both HFT activity and seamless trading across

fragmented markets), (ii) HFT firms demand technologically advanced trading platforms, thus causing market fragmentation (as competing trading venues enter the trading business originally dominated by incumbent exchanges), and (iii) market fragmentation requires more complex systems to manage order routing and execution, which gives rise to HFTs (as they are best positioned to capitalize on arbitrage opportunities). The approach in this thesis is to model HFT market making and fragmentation jointly. The descriptive analysis also corroborates the argument that technological advancements go hand in hand with greater market fragmentation and higher degree of HFT activity.

2.2.8 How does fragmentation affect market quality?

Theory does not provide a definitive answer as to whether consolidation or fragmentation is better for market quality. Consolidation is favored by models that consider network externalities (i.e., "liquidity begetting liquidity"), such as Mendelson (1987), Chowdhry & Nanda (1991), Madhavan (1995), and Hendershott & Mendelson (2000). The underlying mechanism is that a single market has greater chances of matching buyers and sellers, hence greater liquidity. Improved liquidity means that traders are more likely to engage in information production and arbitrage, which in turn leads to more informative prices (Kyle, 1985; Chordia et al., 2008). The opposite of this line of reasoning suggests that fragmentation is harmful to liquidity and price discovery, because it increases search costs for traders and decreases competition among market makers (Yin, 2005).

Fragmentation is favoured by models that account for lower trading costs resulting from competition between trading venues (Battalio, 1997; Foucault & Menkveld, 2008; Colliard & Foucault, 2012). Also, Harris (1993) presents the clientele effect argument for fragmentation, wherein trading venues tailor their market structure to suit the needs of different types of traders.

Empirical evidence on the effects of fragmentation is mixed. O'Hara & Ye (2011) find that in the cross-section of US stocks, more fragmented stocks have lower transaction costs, faster execution speeds, higher short-term volatility and more efficient prices. Degryse, De Jong, & van Kernel (2015) provide evidence of improved aggregate liquidity in a panel of Dutch stocks across multiple lit trading venues. Korber, Linton and Vogt (2013) study the UK, and find that overall market volatility is lower, and trading volume higher in the fragmented market.

Aitken, Chen, & Foley (2017) investigate the introduction of competition in the Australian market and conclude that fragmentation is beneficial to market quality, as it improves liquidity and reduces exchange trading fees. Boussetta et. al (2017) provide theoretical and empirical evidence that fragmentation improves global liquidity across markets.

Negative effects are documented in Hendershott & Jones (2005) who find that fragmentation in trading for three ETFs harms liquidity and price discovery. Haslag & Ringgenberg (2017) instrument fragmentation with staggered introduction of Reg NMS in the US, and show that the two opposing effects affect stocks differently: network externalities (negative effect) dominate for small stocks, and competition on fees (positive effect) dominates for large stocks.

A related strand of literature studies the welfare implications of fragmentation. The welfare-increasing mechanisms of fragmentation include lower trading fees (Colliard & Foucault, 2012) and product differentiation that benefits heterogeneus investors (Pagnotta & Philippon, 2018), while welfare-decreasing effects arise from cross-venue arbitrage (Baldauf & Mollner, 2016).

2.3 A simple model of what drives the OTTR

2.3.1 Baseline model structure

Consider a simple model in which a market maker posts bid and ask quotes for a given asset in a given market. The market maker has the option to monitor one or more signals from a set of signals, $\Omega_s = \{s_1, s_2, \ldots, s_N\}$. The signals could be prices of related securities, prices of the same security in another trading venue, the state of the order book, and so on. Each signal is a flow of time-series data that changes at stochastic times (termed "information arrivals") given by Poisson processes with intensity λ_n for the n^{th} signal. The quality of signal n, q_n, is the probability that when there is a change in that signal (an "information arrival"), the market maker will want to update his posted quotes, resulting in a "cancel and enter" or "amend" message from the market maker. Events that trigger such quote revisions are termed "*relevant* information arrivals".

Each event of *relevant* information arrival triggers quote updates by the market maker. The market maker can react by: (i) posting updated bid and ask quotes

that keep the spread unchanged, or (ii) posting updated bid and ask quotes that result in a narrower or wider spread. In either scenario, the market maker reacting to the relevant information arrival results in quoting activity.

The market maker incurs marginal cost c to process an additional information arrival. Hence, the expected monitoring cost for signal n per unit time is proportional to the intensity of information arrivals: $\lambda_n c$. For example, a market maker might need to subscribe to an additional data feed (e.g., a live streaming data feed from an exchange), or add an additional monitoring function to their trading algorithm. This cost can be interpreted as processing capacity that is required to interpret information arrivals and determine whether/how to respond.

Market orders arrive at stochastic times given by a Poisson process with arrival rate λ_m, and trade against the market maker's posted quotes. The market maker's benefit from monitoring comes from avoiding having stale quotes picked off. When a market order arrives after a relevant information update, but the market maker has not updated their quotes in response to the information (this occurs when relevant information arrives for a signal that is not monitored by the market maker), then the market maker's (stale) quotes are picked off and he incurs a picking-off cost, k. It follows that the more signals the market maker monitors, the less often his quotes will be picked off. It also follows that for a given monitoring intensity, the picking-off cost per unit time increases with the frequency of relevant information arrivals, i.e., with the asset's fundamental volatility.

The market maker chooses which signals (if any) to monitor by weighing up the costs of monitoring, $\lambda_n c$, against the benefits of monitoring, namely reducing picking-off risk. The benefits depend on the arrival intensity of market orders and the arrival intensity of relevant information. Hence, the choice of monitoring intensity is endogenous in the model.

To help understand the optimal choice of which signals to monitor, I define a signal's usefulness, u_n, as the arrival intensity of relevant information from the signal (signal changes that cause the market maker to want to revise his quotes): $u_n = \lambda_n q_n$. The expected benefit (per unit time) from monitoring a given signal n is the saved losses that would have occurred from having quotes picked off. This benefit is the expected number of times the market maker's quotes would be hit by a market order when he would have wanted to revise them had he seen the signal, multiplied by the cost of getting hit by a market order without

having updated quotes, k. In one unit of time, the expected number of market order arrivals is λ_m and the probability that a given market order is preceded by useful information from signal n is $\frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n}$. Therefore, the expected benefit per unit time of monitoring signal n is $\lambda_m \left(\frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n}\right) k = \lambda_m \left(\frac{u_n}{\lambda_m+\lambda_n q_n}\right) k$. As this expression reveals, the benefits of monitoring signal n are increasing with the signal's usefulness (u_n), increasing in the picking-off cost (k), and increasing in the market order arrival rate (λ_m).

Recall there is also a cost of monitoring a signal and therefore a market maker will optimally monitor all signals for which the benefit exceeds the cost. The expected cost per unit time of monitoring signal n is $\lambda_n c$, giving a net benefit of $\lambda_m \left(\frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n}\right) k - \lambda_n c$ from monitoring the signal. The market maker adds signals to his "monitored list" from greatest to least expected net benefit until the marginal expected net benefit of adding the next signal is less than or equal to zero. The market maker therefore monitors all signals for which:

$$\lambda_m \left(\frac{\lambda_n q_n}{\lambda_m + \lambda_n q_n}\right) k - \lambda_n c > 0 \tag{2.1}$$

I denote the set of *monitored* signals Ω_{s^*}. Condition 2.1 determines monitoring intensity (the number of monitored signals). As a result of monitoring signals, executing trades, and updating quotes, the market maker generates messaging activity (messaging includes order entry, cancellation, and amendment messages) at an expected rate of Q messages per unit time:

$$Q = \sum_{n \in \Omega_{s^*}} 2\lambda_n q_n + 2\lambda_m \tag{2.2}$$

The first term, $\sum_{n\in\Omega_{s^*}} 2\lambda_n q_n$, is due to quote updates in response to relevant information arrivals on monitored signals, and the second term, $2\lambda_m$, is due to reposting liquidity after being hit by a market order. Both terms are multiplied by two reflecting the fact that after observing useful information or being hit by a market order, the market maker updates his view of the fundamental value and thus adjusts both bid and ask quotes (adjusting the price and / or volume of those quotes).

Recognizing that the expected number of trades per unit time is just the market order arrival intensity, λ_m, the expected OTTR is given by:

$$OTTR = \frac{\sum_{n \in \Omega_{s^*}} 2\lambda_n q_n + 2\lambda_m}{\lambda_m} \tag{2.3}$$

2.3.2 Model with fragmented markets

If the number of markets increases from 1 to M, the single (representative) market maker posts liquidity across multiple venues. The aggregate market order arrival rate, λ_m, is assumed to remain the same as in the single-market case, just split across multiple venues. The overall quoting activity of the market maker consists of two components: (a) quote updates resulting from relevant information received by monitoring signals, $2M \sum_{n \in \Omega_{s^*}} \lambda_n q_n$ (market maker updates quotes on all M markets in response to monitored signals), and (b) reposting liquidity / revising quotes on all markets after getting a fill on market orders, $2M\lambda_m$ (market order arrivals constitute useful signals, from the market maker's viewpoint). Note that market fragmentation does not affect the signal monitoring decision of the market maker, who chooses the set of signals to monitor in the same manner as in a single-market case. For a given security, the resulting expected OTTR across venues is therefore:

$$OTTR = \frac{2M(\sum_{n \in \Omega_{s^*}} \lambda_n q_n + \lambda_m)}{\lambda_m} \tag{2.4}$$

Consider the OTTR of individual markets $j = 1 \dots M$. The market share of trading volume (market orders) for each individual market j is ρ_j. The market maker updates his quotes on market j in response to the following:

(a) every time a piece of relevant information is received from the monitored signals, which results in the number of quote updates $2 \sum_{n \in \Omega_{s^*}} \lambda_n q_n$

(b) after being hit by a market order on market j, which results in the number of quote updates $2\rho_j \lambda_m$

(c) after being hit by a market order on any other market besides market j, which results in the number of quote updates $2(1 - \rho_j)\lambda_m$, since market order arrivals constitute useful signals.

Then, the OTTR for market j is:

$$OTTR_j = \frac{2\sum_{n\in\Omega_s^*}\lambda_n q_n + 2\rho_j\lambda_m + 2(1-\rho_j)\lambda_m}{\lambda_m\rho_j} \tag{2.5}$$

Simplifying Eq. 2.5:

$$OTTR_j = \frac{2\sum_{n\in\Omega_s^*}\lambda_n q_n + 2\lambda_m}{\lambda_m\rho_j} \tag{2.6}$$

2.3.3 Propositions

I now derive propositions about the relation between OTTRs, monitoring intensity and fragmentation. First, I establish the link between OTTRs and fragmentation (Proposition 2.1). Second, I show how OTTRs are related to market shares (Proposition 2.2). Third, I relate OTTR to all the model parameters (Propositions 2.3 – 2.6). I provide the economic intuition for these propositions below, and the proofs in Appendix 2.1. Corresponding empirical hypotheses are presented in Table 2.1.

Proposition 2.1. *The OTTR for a given security increases with the extent of fragmentation of the security's trading across multiple trading venues.*

The intuition is simple. As markets fragment, the market maker posts quotes on several exchanges rather than one, hence every time the market maker updates his quotes (due to information arrival or a trade) it requires order messages be sent to each of the exchanges, driving OTTRs up. Although the model makes a number of simplifying assumptions, such as considering OTTRs of a single market maker that posts bids and asks across all available trading venues, a similar scaling effect (higher OTTR with higher fragmentation) will occur, if the market maker quotes across only a few trading venues or only on one side (bid or ask). In such a scenario, OTTRs would still increase with fragmentation, but at a lower rate than implied by the model.

Empirically, I should observe higher OTTRs for securities with more fragmented trading (Hypothesis 1a). Similarly, from the trading venue perspective, I expect higher OTTRs on exchanges that trade more fragmented securities (Hypothesis 1b).

Proposition 2.2. *The OTTR for a given trading venue is inversely related to its market share.*

Table 2.1: Summary of propositions and empirical hypotheses

Propositions and empirical hypotheses	Empirical support
Proposition 2.1. The OTTR for a given security increases with the extent of fragmentation of the security's trading across multiple trading venues.	
Hypothesis 1a. *OTTRs are higher for securities with more fragmented trading.*	yes
Hypothesis 1b. *OTTRs are higher for exchanges that trade more fragmented securities.*	yes
Proposition 2.2. The OTTR for a given trading venue is inversely related to its market share.	
Hypothesis 2. *OTTRs are higher for markets with lower market shares.*	yes
Proposition 2.3. The OTTR for a given security increases with the quality of signals available for monitoring.	
Hypothesis 3a. *ETFs have higher OTTRs compared to the common stocks.*	yes
Hypothesis 3b. *Securities with higher absolute correlation with the broad market index have higher OTTRs.*	yes
Hypothesis 3c. *OTTRs are higher for securities with higher price-to-tick ratios.*	yes
Hypothesis 3d. *OTTRs are higher on markets with taker-maker fee structures.*	yes
Proposition 2.4. The OTTR for a given security increases with lower monitoring costs.	
Hypothesis 4a. *OTTRs are higher for stocks with higher market capitalization.*	yes
Hypothesis 4b. *OTTRs increase over time, as data processing speeds increase.*	yes
Proposition 2.5. The OTTR for a given security increases with picking-off cost.	
Hypothesis 5a. *OTTRs are higher on days with higher market volatility.*	yes
Hypothesis 5b. *OTTRs are higher for more volatile stocks.*	yes
Proposition 2.6. The OTTR for a given security decreases with the trading frequency, holding the monitoring intensity constant.	
Hypothesis 6. *OTTRs are inversely related to trading volumes.*	yes

This effect occurs, because if a market maker posts quotes on all trading venues and updates them all at the same time, the messaging activity (the OTTR numerator)

will be approximately the same across venues, but the number of trades (the OTTR denominator) will be different. Therefore, empirically, I expect higher OTTRs for trading venues that have lower market shares (Hypothesis 2).

Proposition 2.3. *The OTTR for a given security increases with the quality of signals available for monitoring.*

Monitoring intensity and OTTR are closely related, because the market maker posts quotes as a result of his monitoring. Recall that monitoring intensity in the model is endogenous, and arises from the market-maker's cost-benefit analysis. For example, if the market maker faces higher quality signals (keeping other things constant), he enjoys higher marginal benefit of monitoring and therefore monitors more signals and posts more frequent order updates. This in turn means higher OTTRs.

Empirically, I should observe higher OTTRs in securities with high-quality signals. For example, exchange-traded funds (ETFs) should have higher OTTRs than stocks (Hypothesis 3a), as ETFs follow a predefined index and typically have easily observable high-quality signals (e.g., related trading instruments on the same index, including futures contracts, options, underlying stocks etc.). Similarly, securities that are highly correlated with the market index (including both highly positive or highly negative correlations) can be seen as having high-quality signals available (i.e., a simple data feed of market index updates would provide a fairly precise signal for the value of security that is highly correlated with the market). Therefore, I should observe higher OTTRs in securities that are highly correlated with the broad market index such as S&P500 (Hypothesis 3b).

The tick size is also likely to affect how often a market maker updates their quotes. This effect can be understood through the signal quality parameter. Recall the signal quality is the probability that when there is a change in the signal, the market maker will want to update his quotes. With a very large tick size, there has to be a very large change in the security value before the market maker would want to repost quotes at the next tick. On the contrary, with a very fine pricing grid, even very small changes in security value will warrant quote updates. Therefore, I expect higher OTTRs for stocks with a smaller percentage tick size, i.e., a higher price-to-tick ratio (Hypothesis 3c).

For similar reasons, I also expect higher OTTRs in markets with a finer pricing grid, such as taker-maker markets (Hypothesis 3d).

Proposition 2.4. *The OTTR for a given security increases with lower monitoring costs.*

When the monitoring cost per signal is lower, the market maker has an incentive to monitor more signals, which results in higher OTTRs. Therefore, in the cross-section of stocks, I expect higher OTTRs for stocks that have more low-cost (or free) signals, such as large stocks that have rich public information widely available to investors (Hypothesis 4a). Also, I should observe an upward trend in OTTRs through time, as data processing costs have decreased significantly in the last two decades (Hypothesis 4b).

Proposition 2.5. *The OTTR for a given security increases with picking-off cost.*

When faced with a higher cost of trading at stale prices, the market maker has stronger incentives to monitor more signals to minimize the costs of being hit by market orders without having updated quotes. Therefore, higher picking-off costs lead to higher monitoring intensity and higher OTTRs.

The cost of being picked off is larger when there are greater changes in the value of the underlying security. For example, if in a given period of time, the security value changes by 10 basis points (bps) and the market maker fails to adjust quotes before getting hit by a market order, they incur a loss of about 10 bps. Yet if the value changes by 100 bps and the market maker is hit before they manage to update quotes, their loss is around 100 bps. Therefore, picking-off costs (and thus OTTRs) are expected to be higher in more volatile securities and more volatile times (Hypotheses 5a and 5b).

Proposition 2.6. *The OTTR for a given security decreases with the trading frequency, holding the monitoring intensity constant.*

Holding market maker's monitoring intensity fixed, a higher rate of market order arrivals decreases OTTRs, as every trade is associated with fewer quote updates on average. Therefore, I expect lower OTTRs in stocks with higher trading volumes (Hypothesis 6).

2.4 Empirical analysis

2.4.1 Data and descriptive statistics

The sample starts on January 1, 1998 and ends on December 31, 2018, covering a period in which OTTRs experienced substantial growth. I analyze a representative cross-section of stocks and ETFs, chosen using random sampling with stratification by size. To construct this sample, I obtain the full cross-section of US stocks as of December 2018 from CRSP, sort them by size (market capitalization) and pick every 20th stock. The resulting sample contains 241 stocks in 2018. Next, I take those 241 market caps, and trace each one back through time at the rate of GDP deflator, so that I have the same size distribution in each year, but adjusted for inflation. Then, in each year (1998 – 2017), I select the 241 stocks that most closely correspond to the 241 inflation-adjusted sizes (via sampling without replacement).

This approach constructs a sample that in every year is representative of today's market with respect to the distribution of size, and hence allows me to identify changes in OTTRs through time that happen for reasons (e.g., changes in technology or market structure) other than changes in the composition of the market. I know from prior studies that the composition of stocks in equity markets has changed substantially through time towards fewer, but considerably larger listed companies (e.g., Doige et al., 2017). The sampling approach allows me to control for this tendency in the time series.

I construct the sample of ETFs following a similar approach as that of stocks, but allowing for changes in size (assets under management, AUM) over time. Given that ETF markets are relatively young (the first US ETF was launched in 1993), and have been growing rapidly, it is not possible to take today's distribution of AUM and find matching ETFs in 1998. I therefore sample ETFs by sorting them by AUM in each year, forming ten bins of ETFs by size in each cross-section, and selecting the median ETF (by AUM) from each bin. As a result, the sample covers about 10% of the ETFs available in each year of the sample, and those ETFs are representative of the AUM distribution in that year.

I use the Thomson Reuters Tick History (TRTH) database for order book data (orders and trades) and the Center for Research in Security Prices (CRSP) data for stock characteristics. The data cover 14 US trading venues: NYSE American,

NASDAQ Boston, NSX, Direct Edge X, Direct Edge A, Investors Exchange, CME Chicago, NYSE, NASDAQ, NYSE Arca, NASDAQ Philadelphia, OTC Markets, BATS-Z, and BATS-Y.

I aggregate the intraday trade and quote data to daily observations, with stock- and exchange-level granularity. I compute the OTTR as the number of quote updates (changes to bid or ask prices or volumes at the best bid and ask prices) divided by the number of trades.[2] Table 2.2 provides detailed definitions of all the variables used in further regression analysis.

Table 2.3 reports descriptive statistics. The dataset contains 1,210,225 stock-day observations, and 215,748 ETF-day observations. At the exchange-day granularity, I have 48,772 observations. The sample period includes 5,284 trading days. I winsorize the OTTR variable in the stock-day dataset at the 99th percentile to reduce the effects of outliers. The winsorization is done separately for the stock-day and ETF-day datasets. The exchange-day OTTR is a simple average of winsorized OTTRs from ETF-day and stock-day datasets.

An average stock in the sample has daily OTTR of 16.87, daily volume of 0.99 million shares, and trades on 5.44 markets on an average day between 1998 and 2018. The average market capitalization is $5.7 billion, average absolute correlation with the S&P500 index is 0.39, and average range of daily high-low prices is 4.06%. ETFs have an order of magnitude higher OTTRs (349.41 on average), trade on slightly fewer markets (4.76 on average), but in comparable volumes (1 million shares daily), experience lower daily volatility (1.04% average high-low range, similar to daily S&P500 volatility of 1.35%), and have a higher absolute correlation with the S&P500 index (0.67) due to diversification of idiosyncratic risk. The average ETF in the sample has assets under management of $1.85 billion. ETFs also have lower average tick-to-price ratios, compared to stocks (4 bps for ETFs vs 23 bps for stocks). Note that tick-to-price ratios reflect the tick size changes in the US markets following the introduction of decimalization (effective

[2]The numerator of this measure sums the number of times in a given day the best bid changes either in price or quantity and the number of times the best ask changes either in price or quantity. Each order entry, amendment, and cancellation at the best bid or ask will trigger a change to the price or quantity at the best bid and ask and therefore contribute to the numerator. The measure does not consider order messages beyond the best quotes. In unreported analysis I use SEC MIDAS data that includes order cancellations and amendments at all levels and find that the two versions of OTTR (that computed from SEC MIDAS data and that from TRTH data) are very closely correlated, although different in their magnitude.

Table 2.2: Variable definitions

Variable name	Variable definition
Stock-day observations	
OTTR	Number of order entry, amend, and cancel messages at the best quotes, divided by the number of trades. Winsorized at 1% level.
Frag1	Number of trading venues where a security has at least one trade on a given day.
Frag2	One minus Herfindahl-Hirschman index calculated from dollar volumes. Herfindahl-Hirschman index is the sum of squared market shares of all exchanges trading a given security.
Frag3	One minus Herfindahl-Hirschman index calculated from number of trades. Herfindahl-Hirschman index is the sum of squared market shares of all exchanges trading a given security.
ETF Dummy	Takes the value of one for exchange-traded funds, and zero otherwise.
AbsCorrelS&P	Absolute value of 22-day correlation with S&P500 index. Calculated using daily returns.
TickToPrice	Tick size divided by the closing price.
HighLowVolat	Daily stock volatility measure, computed as high minus low price, divided by the average of daily high and low prices.
MktCap	Market capitalization in thousand USD. For ETFs, MktCap is AUM (as reported in CRSP).
Volume	Number of shares traded.
Exchange-day observations	
OTTR	Simple average of the OTTR of stocks and ETFs traded on a given exchange, computed from the winsorized (at 1%) stock-day OTTR.
NumMkts	Number of trading venues with at least one trade for an average security traded on a given exchange.
MktShareStocks	Exchange's market share in stock trading, computed from dollar volumes.
MktShareETFs	Exchange's market share in ETF trading, computed from dollar volumes.
Taker Dummy	Exchange's market share in ETF trading, computed from dollar volumes.
Time series observations	
HighLowVolatMkt	Daily market volatility measure, computed as high minus low S&P500 level, divided by the average of daily high and low.
VIX	Level of VIX volatility index.
t	Time trend variable incremented by one each trading day.

Table 2.3: Descriptive statistics

This table reports descriptive statistics for the variables used in regression analysis. The sample contains 1,210,225 stock-day observations, 215,748 ETF-day observations, 48,772 exchange-day observations, and 5,284 time-series observations. $OTTR$ is the order-to-trade ratio. $Frag1$ is a measure of fragmentation based on number of markets. $Frag2$ ($Frag3$) is a measure of fragmentation based on Herfindahl-Hirschman index using dollar volume market shares (number of trades market shares). $AbsCorrelS\&P$ is the absolute correlation of the security returns with the S&P500 index. $HighLowVolat$ is a security-level volatility measure, computed from the high/low price range. $HighLowVolatMkt$ is a market-level volatility measure, computed as the S&P500 index volatility. The sample includes 241 stocks and 20 ETFs from 1998 to 2018.

	Mean	StdDev	25th pctl	50th pctl	75th pctl
Stock-day observations					
OTTR	16.88	35.28	4.00	7.33	13.66
Frag1	5.45	3.22	3.00	5.00	8.00
Frag2	0.47	0.26	0.26	0.53	0.70
Frag3	0.53	0.26	0.33	0.61	0.75
AbsCorrelS&P	0.39	0.24	0.18	0.37	0.58
TickToPrice, %	0.23	0.89	0.03	0.07	0.21
HighLowVolat, %	4.06	4.06	1.79	2.95	4.95
MktCap, $ bn	5.70	24.15	0.11	0.54	2.36
Volume, mln	0.99	3.84	0.02	0.15	0.62
ETF-day observations					
OTTR	349.42	744.69	23.96	82.62	300.19
Frag1	4.76	2.81	2.00	4.00	7.00
Frag2	0.50	0.27	0.34	0.57	0.72
Frag3	0.54	0.26	0.43	0.61	0.75
AbsCorrelS&P	0.67	0.28	0.49	0.77	0.89
TickToPrice, %	0.04	0.07	0.01	0.02	0.04
HighLowVolat, %	1.04	2.66	0.37	0.68	1.21
MktCap, $ bn	1.85	8.26	0.04	0.19	0.89
Volume, mln	1.00	7.42	0.01	0.03	0.18
Exchange-day observations					
OTTR	70.16	48.57	34.69	60.44	98.01
NumMkts	6.53	2.74	3.57	7.50	8.84
MktShareStocks	0.09	0.14	0.01	0.05	0.13
MktShareETFs	0.11	0.17	0.00	0.03	0.14
Time series observations					
HighLowVolatMkt, %	1.35	1.00	0.71	1.09	1.68
VIX	20.21	8.50	13.89	18.55	24.05

as of April 9, 2001), as well as the tick size changes related to the SEC Tick Size Pilot 2016 – 2018.[3]

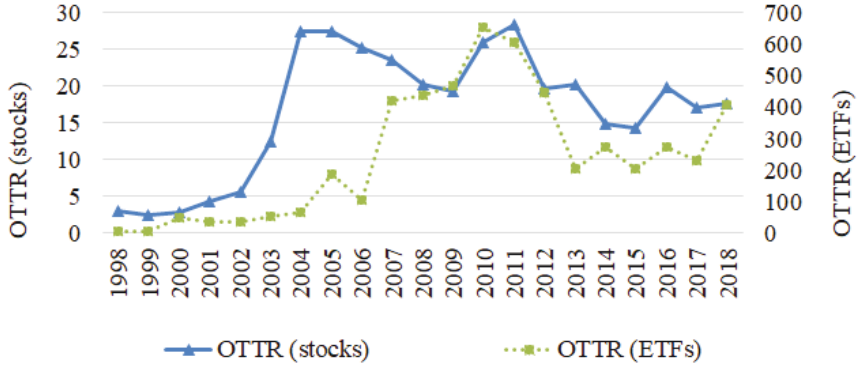
An average trading venue in the sample has a daily OTTR of 70.16 across all securities it trades (both stocks and ETFs). The average exchange-day OTTR is higher than the average stock-day OTTR, because high ETF OTTRs drive the mean OTTR up. The average market shares of venues are 9.79% in stocks and 10.77% in ETFs. I distinguish between the "taker-marker" markets (Edge-A, Bats-Y and NASDAQ Boston) and the "maker-taker" markets (the rest). "Maker-taker" is a trading fee structure that charges a higher fee to "liquidity takers" (i.e., those submitting market orders) than "liquidity makers" (i.e., those posting limit orders), with that latter sometimes being rewarded with a rebate for limit orders that execute. "Taker-maker" markets do the opposite (i.e., charge limit orders and compensate market orders). The taker-maker fee model allows market participants to trade at a more granular price grid of sub-pennies (net of fees) by providing price improvement relative to the National Best Bid and Offer (NBBO).

2.4.2 Time series trends in OTTRs and concurrent market structure changes

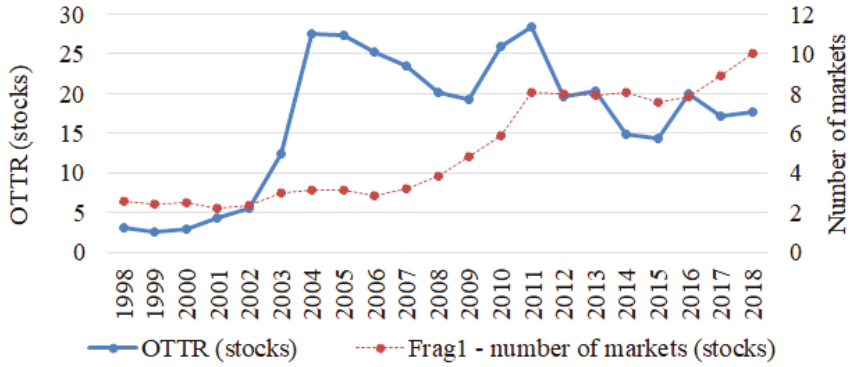
The time series of OTTRs (see Figure 2.1) reflect key market structure changes in US markets, including the introduction of autoquote and the order protection rule (Rule 611 of Regulation National Market System). The first increase in stock OTTRs coincides with the introduction of autoquote in 2003, as NYSE started disseminating order book updates automatically. As discussed in Hendershott, Jones, & Menkveld (2011), the phase-in of autoquote resulted in a 50% increase in message traffic almost immediately and laid the foundation for electronic market making. Autoquote increased the arrival intensity of signals and created demand for fast data feeds, which enabled market makers to drastically reduce their reaction time to new information. In other words, the arrival rates of relevant signals increased, while the marginal cost of monitoring signals decreased, resulting in more frequent quote updates by market makers and a higher OTTR.

[3]The tick size changes relevant to the sample period are as follows. In 1997, the New York stock exchange and NASDAQ reduced the tick size from 1/8th of a dollar to 1/16th. In April 2001, all US exchanges switched to a tick size of one cent for stocks priced above one dollar and 0.01 cents for stocks priced below one dollar. Between October 3, 2016 and September 28, 2018, as part of the SEC Tick Size Pilot, 1,400 stocks had an increase in tick size from $0.01 to $0.05.

Panel A. OTTRs for stocks vs ETFs



Panel B. OTTRs vs fragmentation



Panel C. OTTRs vs correlation with S&P 500, and market volatility

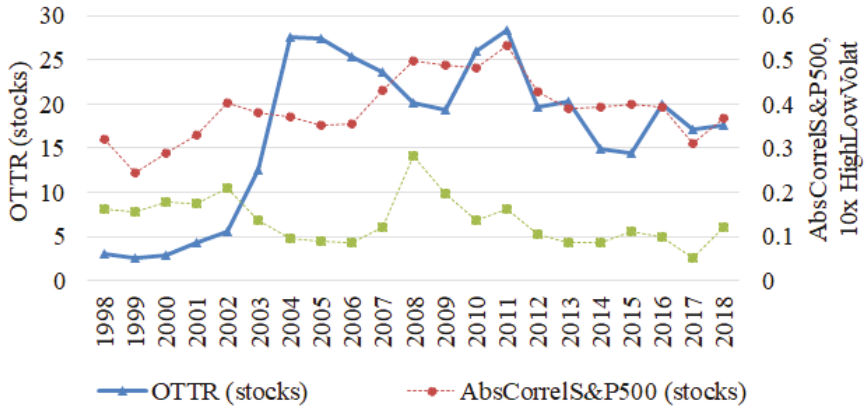


Figure 2.1: Time series of OTTRs and explanatory variables

This figure plots the annual averages of OTTRs and explanatory variables. $Frag1$ is the number of stock exchanges on which the stock is traded. $AbsCorrelS\&P500$ is the absolute correlation of the stock's returns with the S&P 500 index. $HighLowVolat$ is daily stock volatility. 10x $HighLowVolat$ (market) is daily S&P 500 volatility multiplied by 10. The sample includes 241 stocks and 20 ETFs.

OTTRs further increased in 2005, when the SEC enacted the order protection rule

(OPR) of Regulation National Market System (Reg NMS). The rule forces orders to be routed to the trading venue that offers the best quotes. As a result, new trading venues emerged, and trading fragmented across multiple exchanges. In 2001, an average stock in the sample traded on two venues, but by 2008 it traded on four. ETF trading also fragmented around the OPR.

Stock OTTRs remain high during the period 2007–2013. These high OTTR levels coincide with several potential drivers. First, the global financial crisis of 2007–2009 was associated with very high volatility and high correlations of stocks and ETFs with the market index. Both of these factors are positively related to OTTRs, as highlighted by the model. Second, trading continued to become more fragmented, contributing to high OTTRs. Finally, this period also coincides with growth in high-frequency trading (HFT), underpinned by technological advances that decreased the costs of monitoring a large set of high-frequency data feeds.

Overall, the average OTTR levels in the US, at about 25–28 during their peak levels in 2011–2012, are relatively high compared to other less fragmented markets. For example, Australia has an average OTTR of 7.5 during 2011–2012 (ASIC, 2012). One of the key differences between the US and Australian markets is that the former are significantly more fragmented. An average stock trades on 10 venues in the US during 2011–2012, but only on two venues in Australia. For further comparison, Canadian and European markets, which have similar or higher degrees of fragmentation compared to the US, display higher OTTRs: average OTTR in the EU is about 30 in 2012, and in Canada it is 50 (ASIC, 2012).

2.4.3 Cross-sectional and time-series determinants of OTTRs

I test the relation between OTTRs and the drivers implied by the theory model using regressions. Broadly, I ask how the hypothesized factors such as market fragmentation, monitoring intensity, picking-off risk and so on are related to OTTRs through time and in the cross-section of stocks and markets. The empirical design does not explicitly address endogeneity issues. Instead, the regressions examine non-causal relations between observed OTTR levels and hypothesized drivers of

OTTR, using the theory model for guidance. Testing non-causal relations between OTTRs and the drivers of OTTR suggested by the theory model does not necessarily require exogenous variation in explanatory variables.

Table 2.4: Determinants of OTTRs using stock-day observations

This table reports OLS regression results for six regression models with stock-day observations. The dependent variable is $LogOTTR = Ln(1+OTTR)$. Independent variables are in the first column. Variable definitions are in Table 2.2, noting that $LogTickToPrice = Ln(1 + TickToPrice)$, $LogMktCap = Ln(MktCap)$, $LogVIX = Ln(VIX)$. Standard errors are clustered by stock and day. T-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively. The sample includes 241 stocks and 20 ETFs from 1998 to 2018.

	LogOTTR (1)	LogOTTR (2)	LogOTTR (3)	LogOTTR (4)	LogOTTR (5)	LogOTTR (6)
Frag1	0.08***					
	(25.22)					
Frag2		0.31***		0.28***	0.30***	0.29***
		(6.23)		(5.50)	(6.06)	(5.66)
Frag3			0.17***			
			(3.76)			
ETF Dummy	1.44***	1.32***	1.31***	1.33***	1.32***	1.32***
	(24.20)	(22.33)	(22.39)	(22.70)	(22.25)	(22.64)
AbsCorrelS&P	1.13***	1.11***	1.10***	1.16***	1.12***	1.12***
	(20.85)	(20.58)	(20.87)	(21.60)	(20.24)	(20.36)
LogTickToPrice	-14.88**	-12.63*	-11.49**	-11.90*	-12.66*	-11.79*
	(-1.96)	(-1.92)	(-2.17)	(-1.79)	(-1.92)	(-1.79)
HighLowVolatMkt	2.28***	4.35***	4.28***			4.28***
	(2.93)	(6.14)	(6.07)			(5.76)
LogMktCap	0.01	0.04***	0.04***	0.04***	0.04***	0.04***
	(1.57)	(4.26)	(4.42)	(4.53)	(4.10)	(4.62)
LogVolume	-0.29***	-0.28***	-0.27***	-0.28***	-0.28***	-0.28***
	(-38.84)	(-38.21)	(-38.86)	(-36.94)	(-37.88)	(-36.80)
LogVIX					0.08***	
					(3.06)	
HighLowVolat				0.73***		0.43**
				(4.50)		(2.55)
t (time trend)		0.00***	0.00***	0.00***	0.00***	0.00***
		(14.39)	(15.89)	(13.35)	(14.07)	(13.73)
Adjusted R^2	53%	55%	56%	55%	55%	55%
Clustered Std Errors	Stock&Day	Stock&Day	Stock&Day	Stock&Day	Stock&Day	Stock&Day

Tables 2.4 and 2.5 report the regression results. In Table 2.4, I use variables aggregated at the stock-day level and thus focus mainly on stock-level characteristics. In Table 2.5, I use aggregation at the exchange-day level and therefore include market-level variables.

There is a positive relation between OTTR and fragmentation, consistent with the model. For each additional market on which the security is traded ($Frag1$), the OTTR increases by 8%, all else equal (see regression (1) in Table 2.4).[4] Using alternative proxies for fragmentation (e.g., Herfindahl-Hirschman index) and

[4]Note that in discussing these results, I approximate the change in OTTR to the change in (1+OTTR). For example, the precise interpretation would be to say that for each unit increase in $Frag1$, (1+OTTR) increases by 8%, all else equal. However, the difference is negligible for

controlling for the time trend in OTTR confirms this result. So does analysis aggregated at the exchange level (the positive coefficient on $NumMkts$ in regression (1) of Table 2.5). This evidence supports Hypotheses 1a and 1b that OTTRs are higher for securities with more fragmented trading and markets that trade more fragmented securities. The model suggests that this effect arises, because when trading is fragmented, liquidity providers duplicate their quotes across multiple venues, leading to more order activity for the same amount of trading.

The regression results also confirm that securities with high-quality signals have higher OTTRs. On average, OTTRs are 133% higher for ETFs compared to stocks, after controlling for other factors, according to regression (4) in Table 2.4. OTTRs are higher for securities that have stronger correlations with the market index. This evidence supports Hypotheses 3a–3b and is consistent with liquidity providers revising their quotes more often when there is a richer stream of relevant information about the security value. ETFs, as index-tracking securities, have a particularly frequent and relevant set of signals (e.g., underlying stock prices, index futures, and so on), leading to a considerably higher rate of quote revisions and OTTRs. Similar logic applies to securities that are highly correlated with the S&P500 index, and for which S&P500 futures and ETFs can be used as signals about market movements.

The relative tick size also determines how often arriving information will be substantial enough to warrant revising the quotes by an entire tick. To illustrate this, consider two stocks. Stock A is priced at \$50, and stock B at \$5. Say, a tick size is \$0.01, and stock A quotes are \$49.99–\$50, while stock B quotes are \$4.99–\$5. If a piece of news comes out, implying 3 bps improvement in the stock fundamental value, the change in midquote that reflects that change in fundamental value is ($\$49.995 \times 0.0003 = \$0.015$) for stock A, but only ($\$4.995 \times 0.0003 = \$0.0015$) for stock B. Because the change in fundamental value is greater than the minimum tick size for stock A ($\$0.015 > \$0.01$), the market maker in A will update his quotes (shifting the mid-quote from \$49.995 to \$50.005, as the new bid-ask becomes \$50–\$50.01). However, the market maker in stock B will not update quotes, as the value change lies within the bid-ask spread (3 bps improvement translates into \$0.0015 value, which is smaller than full tick size). If the two securities have the same volatility in the fundamental value, the market maker in security A (low

observed OTTR levels, so for convenience of discussion, I phrase the results with respect to effect on OTTR, rather than on (1+OTTR). The rest of the regression results are interpreted in a similar manner.

tick-to-price security) will revise his quotes more often due to the finer pricing grid than the market maker in security B (high tick-to-price security).

Empirically, regression (4) in Table 2.4 suggests that a 1% higher tick-to-price ratio is associated with 11.9% lower OTTRs (in line with Hypothesis 3c). This finding complements evidence in Yao & Ye (2018) that a large relative tick size constrains competition in prices and results in lower OTTRs.

The regression results suggest that OTTRs are higher for large stocks, in line with Hypothesis 4a, and consistent with evidence in O'Hara (2015) and Rosu et al. (2020). As shown in regression model (3) in Table 2.4, a 1% higher market capitalization is associated with a 4% higher OTTR, all else equal. Because large stocks are more widely followed by analysts and well covered in data feeds, the market maker faces lower cost of monitoring for those stocks or alternatively a better availability of relevant and useful signals about the security value. The result of the cheaper and more relevant information is higher monitoring intensity by liquidity providers, and therefore higher OTTRs.

I also find that OTTRs increase over time (positive coefficient on the time trend), as data processing speeds increase and costs come down, supporting Hypothesis 4b. Indeed, HFT market makers have benefited greatly from faster data processing speeds (see the time series of these speeds in Figure 2.2), which has dramatically reduced the cost of monitoring a large set of signals.

The results also indicate a positive relation between picking-off cost and OTTRs. Specifically, as shown in Table 2.4 regression (3), if daily market volatility (high-low range) increases by 1 percentage point, OTTRs increase by 4.28%. As the risk of trading at stale prices is higher in volatile market conditions, OTTRs increase, reflecting markets makers' higher monitoring intensity (in line with Hypothesis 5a). Regression (4) in Table 2.4 confirms a similar (though lower in magnitude) relation between OTTR and individual stock volatility (in line with Hypothesis 5b).

To assess the economic significance and relative importance of the various factors driving OTTR, I standardize the coefficients and plot their relative magnitude in Figure 2.3. The plot quantifies the effect on the OTTR of a one standard deviation increase in each of the explanatory variables. Panel A suggests that shocks to trading volume and systemic volatility (correlation with the S&P500 index) have the largest impact on OTTRs. A one standard deviation increase

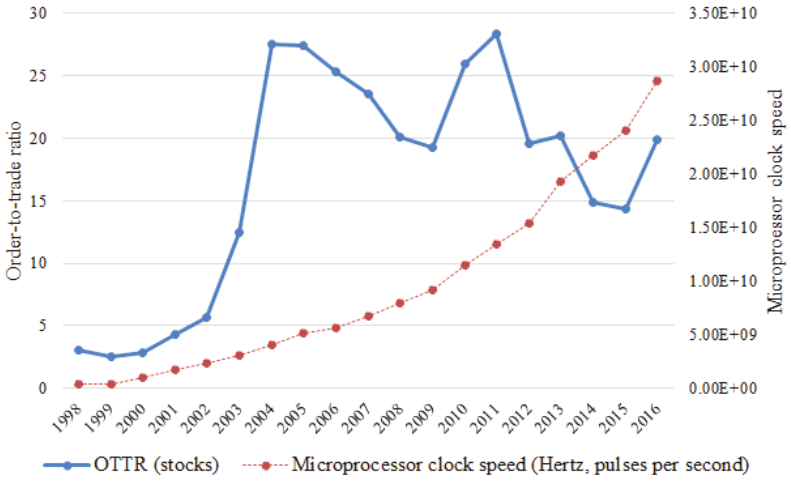


Figure 2.2: Order-to-trade ratios and data processing speed

This figure plots the time series of OTTRs against the microprocessor clock speed measured in hertz (a proxy for technology advancement as in Kurzweil, 2005). The variables are computed as annual averages. The sample includes 241 stocks and 20 ETFs.

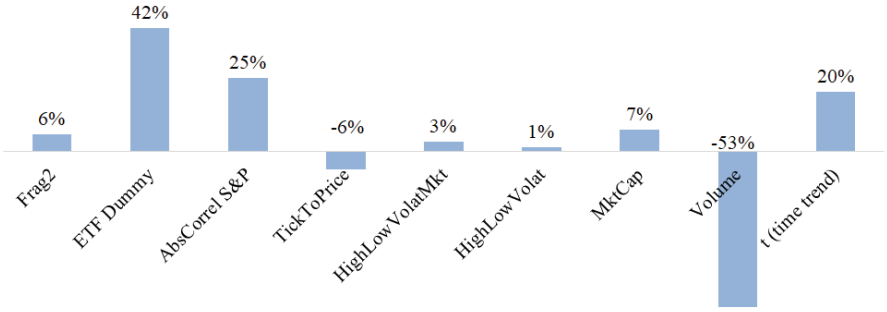
in volume is associated with 53% lower OTTRs, while a one standard deviation increase in correlation with the S&P500 index leads to a 25% higher OTTR.

2.4.4 How OTTRs vary across markets

A third dimension in which OTTRs vary (in addition to time-series and cross-sectional variation) is across markets. The theory model predicts that markets with a lower share of trading volume will have higher OTTRs (see Equation (2.6)). To test this conjecture, I relate the average OTTR observed on a given exchange to this exchange's market share while controlling for other factors, such as fragmentation, volatility, and a time trend.

I find that the average OTTR for a given exchange is inversely related to its share of stock trading volume, consistent with Hypothesis 2. Regression model (2) in Table 2.5 suggests that OTTRs decrease by 2% for each 1% increase in exchange's market share. This result is consistent with market makers executing trades on small venues less frequently, but updating quotes across venues at similar frequency. The estimated coefficient for ETF market share is also negative in Table 2.5, but is not statistically significant.

Panel A. Stock-day observations



Panel B. Exchange-day observations

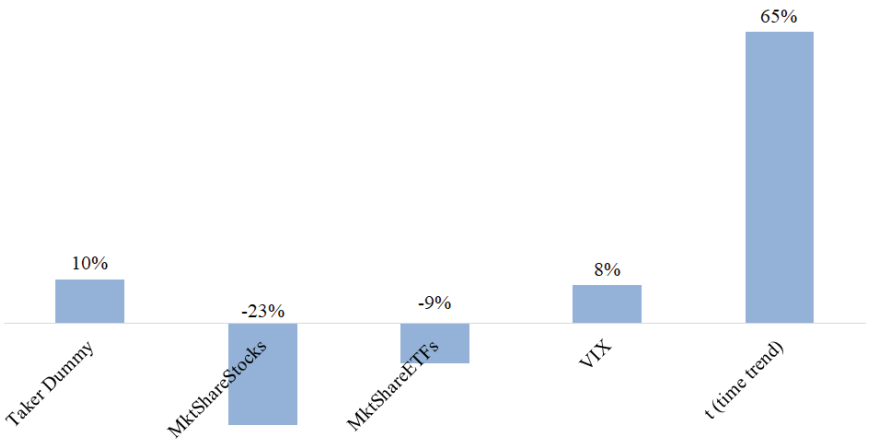


Figure 2.3: Standardized regression coefficients for the drivers of OTTRs

This figure plots the percentage change in $(1 + OTTR)$ associated with a one standard deviation change in the explanatory variables. The explanatory variables are on the horizontal axis. The percentage change in $(1+OTTR)$ is on the vertical axis. For example, one standard deviation increase in the fragmentation measure $(Frag2)$ is associated with 6% increase in $(1 + OTTR)$, holding other factors constant. The sample includes 241 stocks and 20 ETFs from 1998 to 2018.

Differences in exchange fee structures also lead to variation in OTTRs across markets. US venues with taker-maker fee structures effectively allow trades at sub-penny increments. This feature has a similar effect to a smaller tick size: market makers update quotes more often, because the finer pricing grid allows them to respond to small changes in value. The regression results (regression (3) in Table 2.5) suggest that OTTRs are 25% higher on markets with taker-maker fee structures, consistent with Hypothesis 3d.

Table 2.5: Determinants of OTTRs using exchange-day observations

This table reports OLS regression results for five regression models with exchange-day observations. The dependent variable is $LogOTTR = Ln(1 + OTTR)$, where $OTTR$ is a simple average of the OTTRs of stocks and ETFs traded on a given exchange. Independent variables are in the first column. Variable definitions are in Table 2.2, noting that $LogVIX = Ln(VIX)$. Standard errors are clustered by exchange and day. T-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively. The sample includes 241 stocks and 20 ETFs from 1998 to 2018.

	LogOTTR (1)	LogOTTR (2)	LogOTTR (3)	LogOTTR (4)	LogOTTR (5)
NumMkts	0.15***				
	(3.98)				
Taker Dummy	0.39***	0.54***	0.25*	0.25*	0.25*
	(2.78)	(3.23)	(1.71)	(1.71)	(1.71)
MktShareStocks		-2.00***	-1.76***	-1.76***	-1.76***
		(-2.86)	(-3.92)	(-3.91)	(-3.91)
MktShareETFs	-0.36	-0.92	-0.52	-0.51	-0.52
	(-0.51)	(-1.00)	(-0.98)	(-0.96)	(-0.98)
HighLowVolatMkt	0.80	-7.32***	3.35		
	(0.24)	(-2.06)	(0.98)		
LogVIX				0.21	
				(1.23)	
t (time trend)			0.00***	0.00***	0.00***
			(5.31)	(5.22)	(5.17)
Adjusted R^2	28%	20%	38%	38%	38%
Clustered Std Errors	Exchange & Day	Exchange & Day	Exchange & Day	Exchange & Day	Exchange & Day

The economic significance of the exchange-day results is presented in Panel B of Figure 2.3. The time trend in OTTRs is very strong at the exchange-day level: OTTRs for an average trading venue increased at an average rate of 65% per six years (i.e., per 1,516 trading days). Market share is the second most important factor: one standard deviation increase in market share is related to 23% lower OTTRs.

2.4.5 The non-linear effects of fragmentation and market share

Recall that Equation (2.4) of the theory model suggests that OTTRs scale up linearly with the number of markets. The key assumption that generates this prediction is that market makers post quotes on all of the markets that trade a security. The regressions in Table 2.4 confirm that OTTRs are positively related to fragmentation. However, whether the relation is linear or not depends on whether liquidity providers post quotes across all markets or only some. To examine this issue of potential non-linearity in how OTTRs scale up with the number of markets,

I plot the fitted values of OTTRs at different levels of fragmentation and compare them to the theoretical, linear OTTR-fragmentation relation.

For the empirical relation between OTTR and fragmentation, I use fitted OTTR values from a regression, rather than observed OTTR averages for different fragmentation levels. The former approach gives mean OTTR conditional on observing mean levels of other explanatory variables (e.g., tick-to-price, volatility, market cap, correlations with S&P and so on). The fitted OTTR is computed using coefficients from regression model (1) in Table 2.4, with explanatory variables (apart from $Frag1$) fixed at their average levels from Table 2.3.

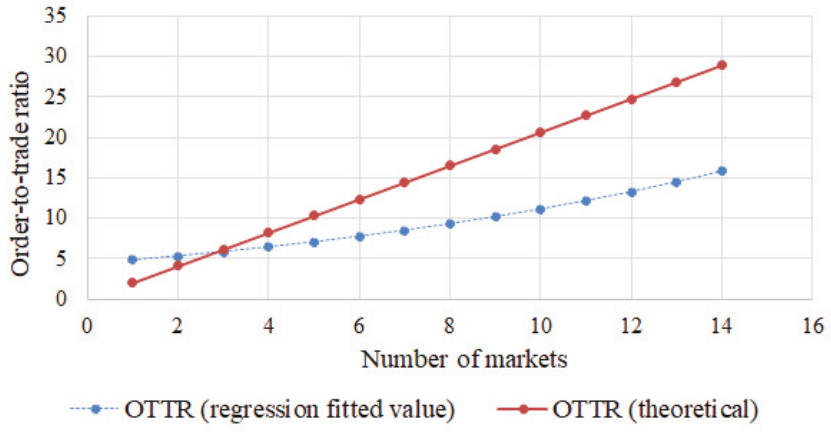
For the theoretical relation between OTTR and fragmentation, I use Equation (2.4) and vary the number of markets from 1 to 14, keeping other parameters fixed. I conservatively assume the monitoring set consisting of one signal (the S&P500 ETF, SPY), and compute the signal quality as the daily proportion of same-direction mid-quote changes in SPY and a given stock, out of the total number of mid-quote changes in SPY (similar to Dobrev & Schaumburg, 2017). Trading (quoting) intensity in a given stock is just the average daily number of trades (quote updates).

The result, in Figure 2.4, shows that for moderate levels of fragmentation (2–4 markets), fitted values of stock-day OTTRs from the regression are broadly in line with the theoretically predicted OTTRs. However, for higher levels of fragmentation (5–14 markets), the empirically observed OTTRs are lower than those implied by the linear theoretical relation. This result is consistent with the notion that market makers post liquidity across several selected venues, rather than across all 14 trading venues, and therefore OTTRs increase with fragmentation, but at a decreasing marginal rate.

Next, I examine the exchange-level relations between OTTR and market share. Recall from Equation (2.6) of the theory model that OTTR on a given market is inversely related to that market's share of trading. The empirical results in Table 2.5 confirm the inverse relation between market share and OTTR, although the regression model does not recover the non-linearity that is suggested by the theory model.

I present the OTTR-market share relation in Figure 2.4. The empirical relation relies on coefficients from regression model (4) in Table 2.5 (I fix the explanatory variables, apart from $MktShareStocks$, at their average levels from Table 2.3).

Panel A. Stock-day observations



Panel B. Exchange-day observations

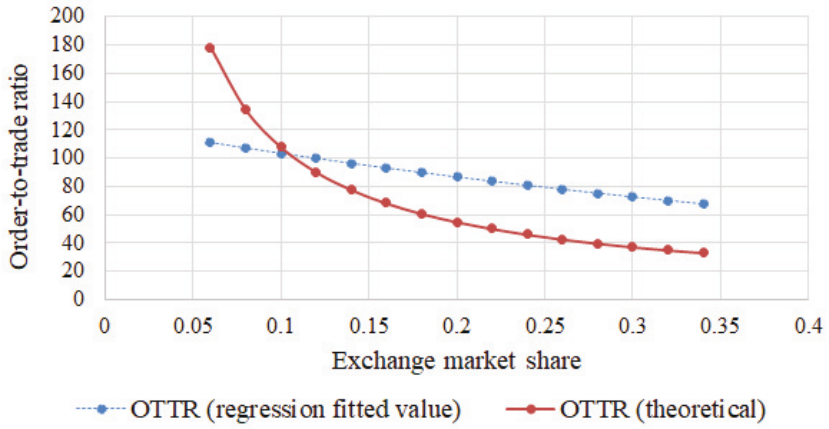


Figure 2.4: The relation between OTTRs, fragmentation, and market shares

Panel A plots stock-day OTTRs for different levels of fragmentation. Panel B plots exchange-day OTTRs for different market shares. The dashed line is the empirically observed relation between OTTRs and fragmentation / market share. The solid line is the theoretical relation.

For the theoretical relation, I use Equation (2.6), and vary market share from 0.2 to 0.32, a range that reflects market shares of US trading venues.

The result in Figure 2.4 shows that for moderate market shares, the regression model recovers OTTR levels that are broadly similar to (or lower than) the theoretically estimated OTTRs. Note that empirically observed market shares are generally lower than the range considered for this exercise. For example, 75th percentile of the market share variable is 13%, just above the 12% market share

at which theoretical OTTRs are aligned with regression fitted values. Note also that exchange-day OTTRs are higher on average than stock-day OTTRs, which is reflected in the scale of y-axis. The greater exchange-level OTTRs result from including ETFs in exchange-day average (recall from descriptive statistics in Table 2.3 that ETFs have an order of magnitude greater OTTRs compared to stocks).

2.5 A benchmark for monitoring OTTRs

One of the reasons for the recent increase in regulatory interest in OTTRs is that abnormally high OTTRs can be signs of illegal trading strategies such as spoofing. But what is an "abnormal" OTTR? How can regulators tell whether a security with twice the OTTR of another security is a manipulation concern or simply a security in which OTTRs are naturally higher? How can regulators tell whether a spike in OTTRs on a particular day reflects suspicious trading or just different market conditions? The challenge is that OTTRs vary substantially through time, across stocks, and across markets for perfectly legitimate reasons.

The theory model provides a simple way to examine whether the OTTR in a given setting is abnormal, by considering the drivers of natural variation in OTTRs. For example, by using Equation (2.4), and the empirical proxies for the theoretical parameters, I can estimate the expected OTTR of a given security in a given market. This provides a benchmark for a "normal" OTTR, against which it is possible to gauge whether an observed OTTR is abnormal and to what extent. A regulator can use such an OTTR benchmark to monitoring markets and identify suspicious trading activity.

To demonstrate this approach, I use the theory model to estimate theoretical OTTRs for 100 random stocks on 20 trading days in November 2018. I then compare the distribution of theoretical OTTRs to the distribution of observed OTTRs for those same stock-days.

I compute empirical OTTRs as the ratio of daily order updates (in both price and quantity), divided by the number of trades. To arrive at theoretical OTTRs, I calculate the following:

$$OTTR_{it} = \frac{2M_{it}(\lambda_{SPY_t} q_{it} + \lambda_{mit})}{\lambda_{mit}} \tag{2.7}$$

where M_{it} is the number of markets trading stock i on day t, λ_{SPY_t} is the number of mid-quote updates in the SPY ETF, which is assumed, conservatively, to be the only signal monitored. The SPY, being one of the most liquid market-tracking securities, is likely to provide a market maker with one of the most relevant indications of changes in market-wide valuations. I estimate the quality of this signal, q_{it}, as the proportion of same-direction mid-quote changes in SPY and the given stock i, out of the total number of mid-quote changes in SPY, similar to Dobrev & Schaumburg (2017). Recall that theoretically, the quality of a signal is the probability that a change in the signal value results in a revised quote for the stock. Finally, λ_{mit} is the number of trades in stock i on day t.

Table 2.6: Empirical OTTRs vs a theoretical benchmark

This table reports descriptive statistics for empirically observed OTTRs of 100 randomly selected US stocks during November 2018 and theoretical OTTRs for these stocks. The empirically observed OTTRs are calculated using daily data on the number of trades and quote messages. Theoretical OTTRs are calculated using Equation (2.7) in the text.

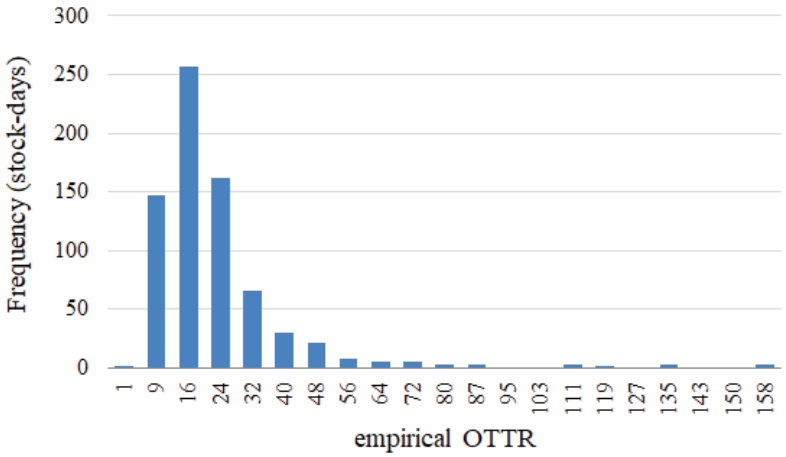
	Mean	StdDev	25th pctl	50th pctl	75th pctl
Empirical OTTR	18.85	17.18	9.57	14.40	22.34
Theoretical OTTR	25.29	5.82	24.18	26.15	28.03
Difference (Empirical - Theoretical)	-6.44	16.38	-15.60	-10.75	-3.29

In Figure 2.5, I plot the empirical distribution of OTTR together with the theoretical benchmark distribution. The figure also shows the distribution of the differences between the empirical and theoretical OTTRs each stock-day. Table 2.6 provides the descriptive statistics for the same two distributions. On average, the observed OTTRs are lower than theoretical OTTRs motivated by market making. However the empirical distribution of observed OTTRs is right-skewed, and for 19% of stock-day observations the empirical OTTRs are higher than the theoretical values. Given I calculated the theoretical OTTRs conservatively by assuming only one signal is monitored, the results suggest that most of the time empirically observed OTTRs are well within levels that are consistent with legitimate market making.

2.6 Conclusions and policy implications

This chapter sheds new light on why OTTRs have grown so rapidly in equity markets and whether they warrant concern. I find that the growth in OTTRs

Panel A. Empirical OTTRs



Panel B. Empirical minus theoretical OTTRs

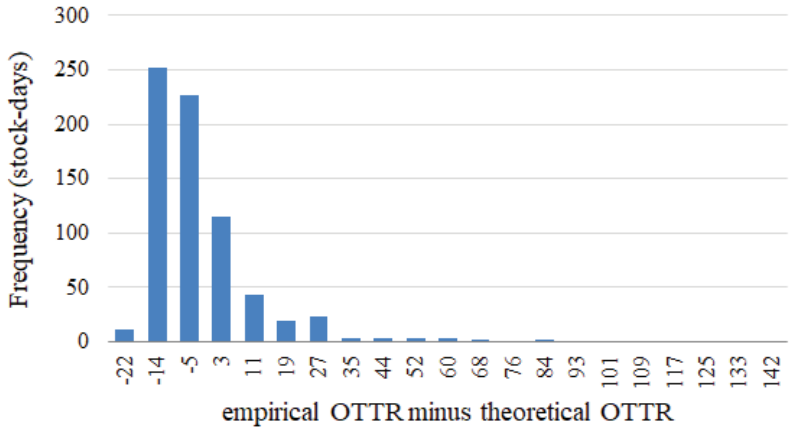


Figure 2.5: Distribution of theoretical vs empirical OTTRs

Panel A plots the distribution of empirically observed OTTRs for 100 randomly selected US stocks during November 2018. Panel B plots the distribution of differences between the empirically observed OTTRs and the OTTRs predicted by the theory model for these same stocks. The vertical axis plots the frequency of observations in stock-days. The horizontal axis plots the values of OTTRs.

through time is largely due to three major changes in US stock markets: (i) the proliferation of trading venues and fragmentation of trading following regulatory changes that encouraged competition between trading venues (e.g., Reg NMS); (ii) automation of quote dissemination, increasing the availability and timeliness of market data; (iii) technological improvements such as increased computational power that increase the ability of market participants to monitor a large set of data feeds. Fragmentation of trading increases OTTRs, because limit orders are

often duplicated across markets, leading to a larger number of quote messages for the same number of trades. Automation of quote dissemination encouraged growth in algorithmic trading, which typically involves more frequent quote updates than manual market making. Finally, by increasing the ability to process data feeds, technological improvements lead to more frequent quote updates as prices or volumes change in other securities.

In light of these drivers, I show that OTTRs in the most recent part of my sample are, in most cases, well within the levels that would be expected to prevail under normal market making. My benchmark for what is expected under normal market making is calculated under conservative assumptions (providing a lower bound on what is a normal OTTR) and accounts for the degree of fragmentation in today's markets and the high frequency of market data. I therefore conclude that the overall levels of OTTRs in the market are consistent with legitimate market making, but occasionally they spike to levels that warrant further investigation.

When investigating whether a given OTTR is abnormal, it is important to account for the substantial cross-sectional and time-series variation in what is a "normal" OTTR. The model of the determinants of OTTRs provides one way to achieve this. The results show that OTTRs tend to be higher in more volatile stocks and market conditions (due to more frequent information arrivals), higher price-to-tick stocks (due to the higher resolution of price increments), lower volume stocks (due to smaller denominator), and in ETFs (because of the availability of frequent and precise signals about an ETF's value). I find that OTTRs are naturally much higher on markets with lower market shares. All of these effects are consistent with a model of market making and therefore should be considered when evaluating whether an OTTR is abnormal, such as in a surveillance system that monitors a market for suspicious activity.

My findings have implications for how regulatory measures that aim to constrain the level of order activity are likely to affect market quality. The model shows that the OTTR emerges endogenously as liquidity providers balance the tradeoff between the costs of monitoring signals and updating their quotes and the costs of having their quotes adversely selected when they have not updated them in response to new information. Adding a fee, a tax, or a limit on order messages raises the costs of keeping quotes up to date. The model implies the outcome is lower monitoring and updating intensity by liquidity providers, which comes at the expense of higher adverse selection risk. Intuitively, if it is more costly to keep

quotes up to date with the latest informational arrivals and changes in market conditions, those quotes will more often be "picked off" and execute when it is unfavorable for the liquidity provider. The ultimate consequence is less liquidity provision (wider spreads and/or less depth) and higher trading costs as liquidity providers recoup the higher adverse selection costs.

The reduction in liquidity due to an order message tax or limit is likely to be the largest when OTTRs would naturally be at high levels, such as periods of high volatility or in low-volume stocks. Such a tax might therefore disproportionately harm liquidity where liquidity is already scarce. Conversely, if OTTR limits or taxes are set to take effect only beyond a certain level, then as long as that level is above the natural OTTR that arises from market making, then the limits/taxes may not impede liquidity provision, and may instead help reduce excessive OTTRs. A limit/tax on OTTRs beyond a particular threshold should take into account that OTTRs will naturally tend to be much higher on exchanges that have lower market shares and for ETFs compared to stocks. With relevant data, future work may distinguish between how illegal activities vs market making impact OTTRs depending on market conditions.

Appendix 2.1. Proofs

Proposition 2.1. The OTTR for a given security increases with the extent of fragmentation of the security's trading across multiple trading venues.

Proof of Proposition 2.1.

Recall the expression for the OTTR for a market maker who provides liquidity in a given stock in M markets: $OTTR = \frac{2M(\sum_{n\in\Omega_{s*}} \lambda_n q_n + \lambda_m)}{\lambda_m}$. Taking the first derivative with respect to the number of markets: $\frac{dOTTR}{dM} = \frac{2(\sum_{n\in\Omega_{s*}} \lambda_n q_n + \lambda_m)}{\lambda_m}$. This expression is strictly positive, because $\frac{2\sum_{n\in\Omega_{s*}} \lambda_n q_n}{\lambda_m} + 2 > 0$, since $\lambda_n q_n \geq 0$ and $\lambda_m > 0$, which means that OTTR is increasing in the number of markets (fragmentation).

Intuitively, for a single market case ($M = 1$), $OTTR \geq 2$ as it takes at least two messages to generate a trade: posting both a bid and an ask quote. If no additional information is obtained from the signals (i.e., signal quality is 0), OTTR=2, which is the case only if $q_n = 0 \; \forall \; n \in \Omega_{s*}$. As $q_n \geq 0$ by construction (signal quality cannot be negative), $OTTR > 2$ for all cases except for $q_n = 0$. As the number of markets increases (e.g., $M = 2$), the OTTR increases, because a market maker now updates quotes twice as often: (a) on two markets rather than one, when a new useful signal n arrives, and (b) on two markets rather than one, when a new market order arrives on at least one market.

Proposition 2.2 The OTTR for a given trading venue is inversely related to its market share.

Proof of Proposition 2.2

Recall the expression for the OTTR for a given trading venue j: $OTTR_j = \frac{2\sum_{n\in\Omega_{s*}} \lambda_n q_n + 2}{\lambda_m \rho_j}$. Taking the first derivative with respect to the market share: $\frac{dOTTR_j}{d\rho_j} = -\frac{2\sum_{n\in\Omega_{s*}} \lambda_n q_n + 2}{\lambda_m \rho_j^2} < 0 \; \forall \rho_j \in (0,1), \lambda_n, q_n, \lambda_m$. The negative first derivative of $OTTR_j$ with respect to market share means that a trading venue's OTTR decreases when its market share increases.

Proposition 2.3. The OTTR for a given security increases with the quality of signals available for monitoring.

Proof of Proposition 2.3.

Recall that a market maker's monitoring intensity is an important driver of the OTTR: $OTTR = \frac{2M(\sum_{n\in\Omega_s*}\lambda_n q_n+\lambda_m)}{\lambda_m}$, where monitoring intensity is the number of *monitored* signals in the set Ω_s^*. As more signals are monitored, the liquidity provider posts proportionally more quote updates in response to those signals, driving the OTTR up. The liquidity provider monitors all signals for which the marginal benefit of monitoring, $\lambda_m\left(\frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n}\right)k$, exceeds the marginal cost, $\lambda_n c$. Because improved signal quality increases the marginal benefit of monitoring without affecting the marginal cost, the liquidity provider will monitor more when he receives better quality signals.

Proposition 2.4. The OTTR for a given security increases with lower monitoring costs.

Proof of Proposition 2.4.

The market maker's cost of processing an information arrival is c. Therefore, his cost of monitoring signal n per unit time is $\lambda_n c$. Marginal net benefit of monitoring (per unit time) is: $\left(\frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n}\right)k - \lambda_n c$.

To see that the monitoring intensity (and therefore the OTTR) increases with a lower cost of monitoring, take the first derivative of marginal net benefit with respect to c: $\frac{d\left(\left(\frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n}\right)k-\lambda_n c\right)}{dc} = -\lambda_n < 0$, as $\lambda_n > 0$ by the properties of Poisson processes (signal intensity, or the number of signal updates per unit time, can only be a positive number). Because a lower c reduces the marginal cost of monitoring without affecting the marginal benefit, the liquidity provider will monitor more signals when his cost of monitoring is lower.

Proposition 2.5. The OTTR for a given security increases with picking-off cost.

Proof of Proposition 2.5.

The market maker incurs the cost k each time he is hit by a market order without having updated his quotes. Taking the derivative of marginal net benefit of monitoring with respect to k: $\frac{d\left(\left(\frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n}\right)k-\lambda_n c\right)}{dk} = \frac{\lambda_n q_n}{\lambda_m+\lambda_n q_n} > 0$. This expression is strictly positive for all non-zero quality signals, hence monitoring intensity (and therefore OTTR) increases with higher picking-off cost.

Proposition 2.6. The OTTR for a given security decreases with the trading frequency, holding the monitoring intensity constant.

Proof of Proposition 2.6.

For a given monitoring intensity, trading frequency only enters the OTTR expression to reflect quote updates in response to executed trades (in the numerator), and the number of executed trades (in the denominator). Taking the first derivative of the OTTR with respect to trading frequency: $\frac{dOTTR}{d\lambda_m} = \frac{-2M \sum_{n \in \Omega_{s*}} \lambda_n q_n}{\lambda_m^2}$. Because $\frac{dOTTR}{d\lambda_m} < 0$ for all parameter values that correspond to non-zero quality signals, OTTR decreases as the trading frequency increases.

Chapter 3

The value of ETF liquidity

"Now you can trade the S&P 500 Index in real time" was the slogan in the newspaper for the first ETF. What kind of nut would do that?

John C. Bogle, founder of the Vanguard Group.

3.1 Introduction

Can two identical baskets of securities trade at different prices? The law of one price says they should not, and yet there are many cases of exchange-traded funds (ETFs) that replicate the same index, but charge different fees (management expense ratios, MERs). Moreover, the fee differentials are persistent, do not decrease through time or with competition, and in fact the higher fee ETF is often the one with *more* trading activity. So, what is it that investors are paying for when they choose a higher cost ETF over a cheaper competitor that tracks the same index? This chapter shows that the answer is liquidity. What is more, I also show that the liquidity clienteles that give rise to the apparent violation of the law of one price are also instrumental to understanding the process by which ETFs compete and the equilibrium that is observed in this rapidly expanding market.

To illustrate the central points of this chapter, consider the MERs of the three ETFs that track the S&P 500 index: State Street's SPY charges 9.4 basis points (bps) per annum, while BlackRock's IVV and Vanguard's VOO charge only 4 bps. I observe a similar situation in most same-index ETFs, whereas highly liquid first movers charge higher management fees compared to their cheaper competitors.

What differentiates the relatively expensive ETFs like SPY from their competitors is the sheer amount of readily accessible liquidity in the secondary market.[1] SPY not only has more assets under management, but also higher turnover, leading to much greater daily trading volume. Investors that have a high need for liquidity, such as those with short holding horizons, will happily trade the SPY, because with short holding horizons, the fee differential becomes negligible while the liquidity does not. Conversely, a long-horizon investor has less need for SPY's liquidity, but will care about the fee differential and is therefore more likely to opt for one of SPY's cheaper competitors. The concentration of high-turnover investors in SPY maintains its high level of trading liquidity. Furthermore, the SPY has no incentive to cut its fee to match its competitors, because its high-turnover clientele is relatively insensitive to the fee and will continue to choose SPY due to its (self-perpetuating) high liquidity.

If faced with a choice of multiple ETFs tracking the same index, which ETF would an investor choose? Intuitively, and as I show formally in a theoretical model, it depends on investment horizon. Short-horizon investors ("high-turnover investors"), find it optimal to choose a more liquid ETF, even if it has a higher fee. Because high-turnover investors trade in and out of positions frequently, annualized trading costs rather than ETF fees constitute a higher proportion of their investment costs. Therefore, high-turnover investors, such as institutions using ETFs for short-term tactical portfolio allocations, are a natural clientele of the highly liquid high-fee ETFs like SPY. This clientele in turn gives greater secondary market turnover and reinforces the higher level of the ETF's liquidity. At the other end of the spectrum, long-term investors such as retail investors using ETFs as buy-and-hold vehicles will be less concerned about liquidity due to their infrequent trading, but relatively more concerned about the fees that create a performance differential in the longer term.

I show that the attraction of short-horizon investors to more liquid ETFs creates a "liquidity begets liquidity" effect. Investors are caught in a form of prisoner's

[1] Arguably, the three ETFs in this example are different in other dimensions, such as tracking performance. Interestingly, despite higher MERs, SPY does not offer better performance in terms of tracking error: in fact, the opposite is the case. In the case of SPY, this adds to the performance drag, resulting in IVV beating SPY by 0.48% over 10 years. Here, SPY ETF is used as an illustration, as it is a rather extreme example in terms of vast accumulated liquidity. In the robustness tests, I show that the results are not driven by S&P 500 ETFs. I also formally control for tracking performance and demonstrate that the sample same-index ETFs are very similar across multiple dimensions (such as legal structure, dividend distribution policies, tax treatment etc.).

dilemma: while it would be optimal for all investors as a group to switch to a cheaper ETF, an individual short-horizon investor has no incentive to individually deviate from the more liquid, higher fee ETF. This effect allows issuers of highly liquid ETFs to extract a rent (via their fee) from the liquidity externalities of their clientele. Liquidity externalities create a strong first-mover advantage among competing ETFs and lead to less than perfectly competitive fee setting. Liquidity externalities also give rise to persistent fee differentials. Thus, the proposed model helps resolve the apparent paradox of same-index ETFs charging vastly different MERs. Ultimately, the results show that the value of market liquidity to an investor is integral to the nature of competition among ETFs and the equilibrium in this market.

I also provide an empirical analysis of interplay between ETF fees, their liquidity, and investor clienteles. The empirical results show the above example of the SPY and its competitors is by no means an isolated case, and liquidity and clienteles play an important role in how same-index ETFs compete, and even more generally, how ETFs on similar indices or investment styles compete. Consistent with the model, for ETFs that track the same index, higher fees (MERs) tend to be associated with more secondary market liquidity: higher dollar volume and narrower relative bid-ask spreads.

I also find empirical evidence of the clientele effect. Higher-fee ETFs tend to have higher turnover (traded dollar volume divided by market capitalization), suggesting a clientele skewed towards shorter-horizon investors. This finding is in line with the industry view that that high ETF turnover is mainly due to institutions trading large positions for short-term tactical allocation, hedging, or rebalancing. The common feature of these institutional traders is that they require substantial liquidity and trade in and out of their ETF positions frequently. Retail traders, on the other hand, are more likely to use ETFs as buy-and-hold investments and therefore have longer holding horizons (Balchunas, 2016).

The model also explains why new low-fee ETF launches, even if based on the same underlying index, do not necessarily drive incumbents to lower their fees. For example, Box, Davis, and Fuller (2018) study the effects of new ETF introductions on the incumbent ETF's liquidity and fees. They find that fees of the incumbent do not decrease as a result of greater competition. The model provides an explanation for this tendency: liquidity externalities prevent investors from switching to lower-fee competitors, thus allowing the incumbent ETF to maintain its higher fee.

Another contribution is that the model characterizes the conditions under which there are likely to be multiple ETFs competing on the same index, as opposed to ETFs differentiating from their competitors by tracking a different index. Same-index competition requires the emergence of liquidity clienteles, which is more likely when there is a high proportion of high-turnover investors and significant differences in the holding horizons between high-turnover and low-turnover investors. Under such conditions, the high-fee ETF finds it more profitable to keep fees high and serve solely the high-turnover clientele instead of lowering their fee to capture both the high-turnover and low-turnover investors. Other factors that make it more likely that multiple ETFs will compete on the same underlying index include a larger combined AUM allocated to a given index and relatively low fixed costs of issuers. These factors matter, because they determine the benefits of economies of scale, which can impede the emergence of multiple competing ETFs.

The findings also help explain the striking concentration of liquidity in a handful of major funds: 50% of ETF dollar volume is concentrated in the top 15 ETFs by traded volume (out of the total of almost 2,000 ETFs listed in the US). This concentration of trading in a handful of ETFs persists despite no shortage of newcomers: a new ETF is launched on average every trading day.[2] By zooming in on indices with multiple competing ETFs, I capture almost half of all equity ETF dollar volumes, and just as my model suggests, those are the ETFs that attract significant institutional trading (i.e., high proportion of high-turnover investors), track major benchmark indices like S&P 500 or Russell 2000, have substantial combined AUM allocated to them, and are held by highly heterogeneous investors: on one side of the spectrum the short-term traders (e.g., for hedging purposes or tactical positions), and on the other — the long-term buy-and-hold investors (e.g., for gaining broad market exposure at low cost).

Finally, I exploit unique features of the ETF market to provide novel measures of the value of market liquidity to the marginal investor. The standard approach in the asset pricing literature to measuring the value of liquidity is very indirect. It involves trying to infer the liquidity premium by measuring average asset returns for securities with different levels of liquidity. In contrast, an implication of my model is that the fee differentials of ETFs that track the same underlying index

[2]According to ETF.com, there were 118 ETF launches in the first half of 2018 (January 1 – June 26), 271 launches in the year 2017, 247 in 2016, 284 in 2015, and 202 in 2014.

directly reveal the value of liquidity. The fee differential reflects the marginal value of the additional liquidity in the more liquid ETF.

Using this insight, I estimate that issuers of highly liquid ETFs can extract 0.51 bps in higher fees (compared to their competitors) per each 1 bp of narrower bid-ask spread. In terms of trading volumes, highly liquid ETFs can extract 1.15 bps higher fees than their competitors that have half as much secondary market trading.

The competition strategies among passive funds are particularly interesting in the light of recent developments in the competitive landscape. In July and August 2018, two big fund managers announced slashing their fees to zero: Fidelity did so for two of their index funds' MERs, and Vanguard for their brokerage platform fees, allowing investors to trade ETFs for free.[3] Just as my model predicts, passive fund issuers can choose to compete at either extreme: by setting fees close to zero while being relatively illiquid (as is the case for Fidelity's index funds), or by charging above-zero MERs while being highly liquid (as is the case for Vanguard's ETFs that you can trade at zero cost).

In the broader context of investments, ETFs have become an increasingly popular investment vehicle. In terms of money invested globally, they exceeded the $5 trillion mark in 2018. In the year 2017 alone, ETFs saw $460 billion of new inflows, which amounts to $1.8 billion inflows on an average working day. ETFs account for over 30% of dollar volume traded in the US stock markets. Seven out of ten most actively traded US securities in 2017 were ETFs rather than stocks (Financial Times, 2017). SPY alone is responsible for around one-third of all ETF traded dollar volume. It is also the most frequently traded security in the world, trading over 20 times a second. It is therefore important to understand the drivers of investors' and issuers' behavior in this rapidly growing market.

This chapter proceeds as follows. Section 3.2 reviews relevant literature, and Section 3.3 describes key institutional details. Section 3.4 develops a model of ETF competition, Section 3.5 derives welfare implications, and Section 3.6 presents the empirical analysis. Section 3.7 concludes the chapter.

[3]See the news releases from CNBC (2019), and Business Insider (2018).

3.2 Literature review

ETFs – index-tracking funds that trade like common stocks – represent the shift to low-cost algorithm-controlled asset allocation. Exchange-traded funds are relatively new: the first ones (Toronto 35 Index Participation Units, TIPs 35) started trading in 1990 on Toronto Stock Exchange. However, SPY (S&P 500 Unit Investment Trust), launched in 1993 by Amex Stock Exchange and State Street Global Advisors, is widely known as the first ETF.

The model developed in this chapter estimates how much ETF liquidity is worth to investors. It speaks to several strands of literature. The first includes studies on ETFs (see Madhavan (2016) for a survey) and fund management more broadly. The second is the liquidity clientele literature pioneered by Amihud and Mendelson (1986) and followed by liquidity-adjusted asset pricing studies. The third is the market fragmentation vs consolidation literature, which considers liquidity-related network externalities.

3.2.1 What are the effects of ETFs?

The ETF literature has mostly focused on various effects of ETFs. For example, a number of studies investigate how ETFs affect market fragility by propagating market-wide demand shocks (Malamud, 2016; Ben-David, Franzoni, & Moussawi, 2018; Krause, Ehsani, & Lien, 2014; Chinco & Fos, 2019; Bhatcharya & O'Hara, 2018).

Multiple studies also explore the effects of ETFs on individual securities. For example, Dannhauser (2017), Madhavan (2016), Lettau & Madhavan (2018), Madhavan & Sobczyk (2016), Glosten, Nallareddy, & Zou (2016), Wermers & Xue (2015), Marshall, Nguyen, & Visaltanachoi (2013), and Li & Zhu (2016) argue that ETFs improve price discovery in the underlying securities. In contrast, Hamm (2014), Da & Shive (2018), Israeli, Lee, & Sridharan (2017) and Agarwal, Hanuona, Moussawi, & Stahel (2018) find that ETFs may harm informational efficiency in underlying securities, causing the latter to incorporate market-wide news rather than idiosyncratic news.

3.2.2 How do investment funds compete?

Little attention has been paid to how ETFs compete and why there is a considerable heterogeneity in the ETF investor base. The existing studies suggest that liquidity is one of ETFs' most attractive features compared to unlisted funds (Madhavan, 2016). Furthermore, as industry practitioners point out, "Many institutional investors won't touch an ETF with volume less than $100 million a day or that isn't the most liquid ETF in the category" (Balchunas, 2016). The model in this chapter recognizes the dominance of institutional traders in the most liquid ETFs, which is consistent with Huang, O'Hara, & Zhong (2019), Li & Zhu (2016), Easley, Michayluk, O'Hara, & Putnins (2020), and Xu, Yin, & Zhao (2019). Unlike earlier studies, I do not zoom in on specific institutional uses of ETFs (e.g., hedging industry-specific risk, circumventing short-selling constraints or employing market timing strategies). Instead, I show that because institutional uses for ETFs lead to short holding horizons, institutions form a natural clientele for highly liquid ETFs. That allows most liquid ETFs to charge higher fees in equilibrium.

This chapter is also related to studies of mutual fund clienteles. It highlights an important difference between ETFs and traditional mutual funds. Namely, in mutual funds, investor turnover is an undesirable feature of a fund, as redemptions and inflows create negative externalities on other investors. For example, Edelen (1999) shows that open-end fund returns are penalized, because fund managers act as uninformed traders in portfolio assets when fund inflows or redemptions occur. I show that ETFs are quite different in two important regards. Firstly, existing ETF investors stand to benefit, not suffer, from higher investor turnover (i.e., more ETF trading in the secondary market), as with more trading, liquidity costs are lower for all investors. Secondly, ETF managers set fees in response to the prevailing liquidity clienteles, rather than choosing the fees to entice a specific type of clientele (e.g., set high load fees to entice long-horizon clientele and minimize redemptions as in Chordia, 1996). Therefore, ETFs and unlisted mutual funds operate and compete in rather different ways.

In a case study of S&P 500 index funds, Hortacsu & Syverson (2004) investigate the factors driving significant dispersion in fund fees. They suggest that search costs and product differentiation play important roles in these funds' ability to charge different fees, despite providing identical index exposure. My study is different in that I consider major index ETFs, which, unlike mutual funds, tend to

be dominated by institutional traders. For example, Agapova (2011) shows that fund flows into ETFs are dominated by institutional investors with higher trading needs, compared to the conventional index funds. To these institutional investors, search and product differentiation are arguably less important than liquidity costs incurred through portfolio rebalancing, portfolio completion, and hedging (Balchunas, 2016).

3.2.3 How much is liquidity worth?

Investors increasingly opt for passive funds tracking a benchmark at low cost instead of paying an active manager for trying to beat the benchmark (Easley, Michayluk, O'Hara & Putnins, 2020). The race to the bottom in passive fund fees means that investors benefit from competition. ETFs, which are a subset of passive funds, offer quite competitive fees (MERs). However, it is telling that first zero-fee passive funds were launched by Fidelity in July 2018, while zero-fee ETFs are yet to emerge. ETFs, unlike traditional open-ended passive funds, are liquid, and their liquidity affects issuers' ability to charge higher MERs in equilibrium. This chapter shows theoretically why the ETF market is prone to oligopolistic fee-setting behavior, despite a new ETF being launched on average every day. The empirical analysis in this chapter illustrates that the theory holds: between two ETFs tracking the same index, the more liquid one charges higher MERs.

The proposed model of ETF liquidity is in the spirit of Amihud and Mendelson (1986). The latter show that long-horizon investors earn higher returns by holding less liquid stocks (while short-horizon investors sacrifice some return by paying higher execution costs for immediate liquidity). A similar mechanism is at play in my model: long-horizon investors opt for low-liquidity low-fee ETFs (while short-horizon investors sacrifice some return by paying higher MERs for greater liquidity). Compared to Amihud & Mendelson (1986), my model differs in several important regards. Liquidity is endogenous and a function of investor choices, whereas in Amihud & Mendelson (1986), the bid-ask spread is exogenously given. Also, ETF issuers are aware of the liquidity/fee tradeoffs that investors face, and set fees to strategically capture a particular clientele.

I measure liquidity similarly to early inventory control models in market microstructure: Garman (1976), Stoll (1978), and Amihud & Mendelson (1980).

Also, the endogenous effect of "liquidity begetting liquidity" in equilibrium is similar to that in Vayanos & Wang (2007), Duffie, Garleanu, & Pedersen (2005), and Mahanti, Nashikkar, Subrahmanyam, Chacko, & Mallik (2008). Previous papers mostly focus on over-the-counter (OTC) markets, particularly bond markets, and consider search costs of investors rather than investment fees. They also do not model the fee-setting dynamics between investors and issuers.

Multiple empirical studies document the inverse cross-sectional relation between asset returns and liquidity, after controlling for risk (Brennan & Subrahmanyam, 1996; Amihud, 2002). Subsequent literature shows that investors demand compensation for liquidity risk (Hasbrouck & Seppi, 2001; Chordia, Roll, & Subrahmanyam, 2000; Huberman & Halka, 2001). I differ from these studies in taking a more direct approach of inferring liquidity premia from MERs of ETFs following identical underlying portfolios. In that respect, my approach is more similar to the bond market papers that compare yield differentials between different instruments of similar coupon and maturity (Amihud & Mendelson, 1991; Krishnamurthy, 2002; Longstaff, Mithal, & Neis, 2005; Friewald, Jankowitsch, & Subrahmanyam, 2012).[4] However, in case of OTC-traded instruments like bonds, credit default swaps etc., liquidity is scarce and search costs are high, hence investors sacrifice yield to avoid extreme illiquidity (Duffie, Garleanu & Pedersen, 2005). However, in case of ETFs, investors are on the opposite side of the liquidity spectrum: they accept higher MER to access extremely liquid securities.

The common thread between the above studies and mine is that liquidity differentials can in fact explain the apparent law of one price (LOOP) violations. While many papers point to limits to arbitrage as one of the major explanations for why such violations can persist (e.g., Shleifer & Vishny, 1997), my results highlight another driver: differences in market liquidity. Differences in liquidity can lead to apparent violations of LOOP not because illiquidity creates limits to arbitrage, but because investors value liquidity. Therefore, two assets with identical cash flows can have persistently different prices, if they have different liquidity clienteles. This tendency is what drives the fee differentials among competing ETFs. Consequently, the LOOP principle could be expanded to a broader concept of liquidity-adjusted LOOP (LALOOP).

[4]See Holden, Jacobsen, & Subrahmanyam (2014) for a more detailed literature review on liquidity premia.

3.2.4 Why do liquidity externalities arise?

Broadly, the model of ETF liquidity is based on the classic microeconomic models of non-cooperating agents, with the resulting Nash equilibrium (Nash, 1951). What differentiates this model from many other settings is that the investors ("consumers") care not only about fees ("prices") but also about a second dimension: liquidity. What makes this setting interesting is that liquidity is not a static product feature, but a product of endogenous investor choices. Fee setting by ETF issuers affects investor choices, which affects liquidity, which then effects fee setting, and so forth. This model is therefore closest to microeconomic models of competition for goods that have network externalities (e.g., Katz and Shapiro, 1985; Economides, 1996), which in this case are liquidity externalities. However, my model differs from standard network externalities models by having clienteles form due to heterogeneity in investment horizons, which is an important and realistic feature.

In considering liquidity externalities, this study is indirectly related to the vast literature on fragmentation of trading across competing trading venues (see Gomber, Sagade, Theissem, Weber, & Westheide (2017) for a more detailed review of consolidation vs competition literature). Another parallel is with the models of fragmented trading, in which multiple equilibria are possible (Pagano, 1989). There are certain parallels between the fragmentation of AUM across same-index ETFs and the fragmentation of trading volumes across multiple exchanges. For example, trading venues compete on speed, fees, and fee structures, which leads to clientele effects in fragmented markets (Foucault, Kadan, & Kandel, 2005; Yao & Ye, 2018). Similarly to Foucault et al. (2005), I model the high-turnover (impatient) ETF traders behaving strategically in choosing the cost-minimizing ETF (trading venue). However, the trading venue liquidity, unlike ETF liquidity, is not driven by investors' holding horizons. Hence, the nature of liquidity advantage in the ETF market is different: sufficient investor heterogeneity is a necessary condition for multiple ETFs in the same index, which is not the case for competition among trading venues.

The fragmentation literature also investigates the welfare effects of competition. For example, Mendelson (1987) models the trade-off between market thinness in case of fragmentation and high order communication costs in case of consolidation.

Baldauf & Mollner (2016) model the trade-off between benefits from lower bid-ask spreads and the costs from cross-venue arbitrage, Colliard & Faucault (2012) suggest that competition is beneficial due to lower trading fees, Pagnotta & Philippon (2018) argue that product differentiation benefits heterogeneous investors. In short, these studies propose different sources of costs / benefits from having multiple providers of the same underlying product (i.e., liquidity of the trading venue). My study also quantifies the costs / benefits from having multiple ETFs providing liquidity in the same index. The model explicitly accounts for welfare transfers between investors and issuers, and the overall dead weight loss to society due to oligopolistic fee-setting.

3.3 Institutional details

3.3.1 Creation-redemption process

The three key players in the ETF ecosystem are issuers, authorized participants (APs), and market makers. The issuer (like Black Rock, State Street or Vanguard) interacts with authorized participants (well-capitalized self-clearing broker-dealers like Barclays, Citibank or Credit Suisse) in the primary market by creating and redeeming ETF shares. Creation happens when an AP delivers a pre-specified creation basket of underlying securities and exchanges them for ETF units.[5] Redemption happens when an AP delivers ETF units, and exchanges them for underlying shares. The ETF creation/redemption unit usually consists of a multiple of 50,000 ETF shares.[6] ETF units are priced based on their net asset value (NAV) rather than the price at which an ETF is trading on the secondary market. In

[5]Depending on the regulatory environment and the contractual agreement between the AP and the issuer, the AP can deliver cash rather than a basket of securities to be exchanged for ETF units. In some cases, the physical creation-redemption basket is based on the subsample of portfolio securities, with the remainder of (typically illiquid) portfolio securities being settled in cash. When the creation / redemption is settled in cash, the ETF basket is forward priced using the closing prices of the same day closing auction. Then, the issuer buys / sells the relevant constituent stocks in the closing auction.

[6]Most existing ETFs operate under a single-basket creation-redemption mechanism: the basket of securities an AP delivers to the ETF issuer when creating ETF units is the same as the basket of securities he receives back from the ETF issuer when redeeming ETF units. However, the recent patent granted to Black Rock envisions a multi-basket ETF structure under which the creation basket may be different from the redemption basket. This structure is targeted at ETFs holding illiquid securities. The patent filing is available here: https://patents.google.com/patent/US8131632B2/en

general, NAV=(Assets - Liabilities)/ number of ETF shares outstanding, where Assets include cash and all securities that an ETF holds (valued at closing prices).

Each ETF typically has multiple APs, which are designated by the issuer to perform creations and redemptions in a given ETF. APs creating and redeeming ETF units is what keeps the ETF price within the arbitrage bands around NAV. This is achieved through the arbitrage mechanism. If the ETF price (net of creation-redemption fee) is above NAV, APs have an incentive to create new ETF units while delivering underlying portfolio stocks to the issuer. If the ETF price (net of creation-redemption fee) is below NAV, APs do the opposite: redeem ETF units and sell the underlying stocks.

The creation-redemption fee is what the ETF issuer charges the AP every time they create or redeem shares. Most issuers charge a fixed fee per creation-redemption event, no matter how many ETF units an AP is creating or redeeming. Large creations therefore benefit from economies of scale. Arbitrage bands (and therefore ETF's tracking difference relative to NAV) are tighter for ETFs with greater trading volumes than for their more thinly traded counterparts. Ultimately, the creation-redemption fee gets passed on to the end investor in the form of bid-ask spread.

Market makers provide bid and ask quotes for ETFs in the secondary market, allowing end investors to access ETF liquidity. They manage their ETF inventory by interacting with APs, who create and redeem ETF shares on behalf of market makers. APs and market makers may or may not be the same firm, but their roles are distinct: APs' merely execute the creation-redemption, while market makers actively decide on whether to create or redeem and how much. For example, when a market maker has excess ETF inventory, he will approach an AP to redeem ETF shares, if the cost of redemption is lower than the cost of carrying his inventory position in the ETF.

3.3.2 Liquidity

ETF liquidity has several layers: (i) secondary market trading, which accounts for 90% of ETF volumes,[7] according to Investment Company Fact Book (2019),

[7]Industry practitioners refer to secondary market trading as "on-screen liquidity" of the ETF. It captures on-exchange ETF volumes.

(ii) primary market trading, represented by off-exchange activity, whereas large institutional investors interact directly with APs and have ETF shares created or redeemed on-demand,[8] and (iii) trading in derivatives on the underlying index or on the ETF itself.

The primary market trading through the creation-redemption process via AP represents the ETF implied liquidity. Implied liquidity in an ETF does not depend on how much the ETF trades on the secondary market, but rather —- how much the ETF constituents trade. Because an AP has to buy / sell underlying securities every time he creates / redeems ETF shares, the liquidity of underlying portfolio is what matters for implied liquidity. Note that across same-index ETFs, implied liquidity is the same, as the constituent stocks of the underlying index are identical.

In addition to the liquidity of underlying basket, the ETF liquidity pool encompasses related derivatives (Abner, 2013). For example, SPY liquidity pool would encompass S&P 500 futures, S&P e-mini futures, SPY options etc. These derivative securities on the ETF / underlying index create an additional layer of ETF liquidity via two distinct channels: (i) they make it cheaper for market makers to hedge their ETF inventory positions, and (ii) they increase trading volumes, as more hedging and arbitrage activity takes place between counterparties in the ETF vs derivatives markets.

ETF options represent another layer of ETF liquidity. Options are written on ETFs that are already highly liquid. Among same-index ETFs, it is typically the first mover ETF that accumulates the most liquidity and has options written on it. To the extent that ETF options are actively traded and have substantial open interest, they have a positive effect on ETF liquidity.

3.3.3 Regulatory Structure

US equity ETFs are regulated by SEC as registered investment companies under the 1940 Investment Company Act (Balchunas, 2016). In practice, this means

[8]For example, when an institution seeks to execute an order of comparable size as an ETF creation/redemption unit (e.g., 50,000 ETF shares), they can directly approach an AP and have the ETF shares created or redeemed without sending an order to the exchange. If an order is not in multiples of ETF units (i.e., 50,000 ETF shares or more), an AP can match multiple orders on the opposite sides of the trade, or combine orders on one side to achieve the size of creation-redemption unit. Bloomberg Tradebook advises institutional investors to resort to block trading, if an ETF order exceeds 1% of average daily volume.

that ETF issuers are subject to the same regulations as mutual funds. Similarly to mutual funds, ETFs should report the marked-to-market net asset value (NAV) at the end of each trading day (ICI, 2019).

Until recently, ETF issuers had to apply for exemptive relief under 1940 Investment Company Act when launching a new ETF. In September 2019, SEC simplified the ETF approval process by adopting a new rule (rule 6c-11 under 1940 Investment Company Act) that eliminates the need for exemptive relief (SEC, 2019). Effective December 23, 2019, the rule makes ETF approval timelines shorter, and ETF launches less costly.

There are two main ETF structures under the 1940 Investment Company Act: unit investment trusts (UITs) and open-end investment funds. The UITs cannot lend out portfolio securities or hold derivatives, do not have a board, and can only reinvest dividends quarterly rather than daily. This results in wider tracking difference for UITs, relative to underlying index. UITs also cannot earn extra income from security lending. These differences make open-end fund structure more popular among ETF issuers.

3.3.4 Tax

The tax treatment of ETFs is relevant at two levels: (i) tax that an ETF pays, and (ii) tax that an investor pays. ETFs are tax-efficient, because unlike a mutual fund, an ETF fund does not face a tax event every time an individual investor transacts in his ETF shares. Because ETFs trade just like stocks, individual investors' actions do not affect taxation of the fund.

If ETFs incur any capital gains from the daily management of the fund, they pass them on to investors, which means investors would incur a tax event. However, because ETFs passively track a predefined index, any capital gains from fund management are rare. In some cases, they may occur in funds using derivatives, but rarely in ETFs that invest in a portfolio of stocks.

The tax that an investor pays applies the same way to ETFs as to stocks. As a general rule, US investors pay 39.6% capital gains tax if they sell their ETF at a gain within a year, and 20%, if they were holding it for more than a year. Another possible taxable event is dividend distributions by the fund. ETFs pay out full

dividends from securities held in their funds. The only difference between funds is the frequency of such dividend distributions, and it is stated in their prospectuses.

3.4 A model of ETF competition

3.4.1 Model primitives

Asset. Consider an economy where a single equity index can be traded. The index tracks a basket of stocks, and pays off a stochastic dividend $\tilde{v}$, where $\tilde{v}$ is normally distributed with mean $\mu \geq 1$ and standard deviation $\sigma > 0$. The risk-free rate is normalized to zero.

Agents. Four types of agents are present in the economy:

(i) two exchange-traded funds that may choose to track the index;

(ii) a continuum of investors;

(iii) a competitive authorized participant for each ETF, acting as an ETF market maker and creating/redeeming ETF units for underlying stocks;

(iv) dealers in the underlying stocks.

There are infinitely many agents in (ii) and (iv), such that dealers and investors do not individually impact the equilibrium outcomes.

At most two risk-neutral exchange-traded **funds** (ETFs) exist in the economy, denoted **L** and **F**. In particular, funds launch sequentially such that the "leader" ETF **L** enters the market before the "follower" ETF **F**. Funds incur a marginal cost $c > 0$ for each unit of assets under management and time, as well as a fixed cost $\Gamma \geq 0$ per unit of time. Importantly, funds have the option not to launch if they expect negative profits (i.e., if the costs are large enough). Upon entering the market, ETFs set management fees f_L and f_F, respectively. The fees are measured as a fraction of AUM per unit time.

Funds accumulate assets under management (AUM) from a unit continuum of **investors**, who choose between the two ETFs. Each investor has a stochastic private value $\tilde{\theta}_t \in \{\theta, 0\}$ per unit of time for holding Q units of the fund (e.g., for

diversification, hedging, or deferred consumption reasons). At $t = 0$, investors are equally likely to have a positive private value (and hold an index fund) or have no private value (and not hold an index fund). Collectively, all investors own $\frac{Q}{2}$ fund units in expectation at any point in time, which is the expected combined AUM in ETFs tracking the index. I assume investors cannot replicate the underlying index directly without incurring prohibitive costs (such as price impact or time costs).

Each investor i trades in and out of her ETF position as the private value switches between zero and θ at exponentially distributed times with rate λ_i. The mechanism is similar to the one in Pagnotta & Philippon (2018). Alternatively, investor i has an expected holding period λ_i^{-1}. Further, the arrival rates are uniformly distributed, that is $\lambda_i \sim \text{Uniform}\,[\Lambda - \xi, \Lambda + \xi]$. Each investor is aware of the dispersion ξ, but does not observe the average holding period Λ^{-1}. It follows that investor i's best estimate of the population mean Λ is her own arrival rate λ_i.

Each fund has its own competitive **authorized participant**, or **AP**,who acts as a market maker in the respective ETF's secondary market and who can create/redeem ETF units in exchange for the underlying stocks. For risk-management purposes, each authorized participant has an inventory constraint of $+/-Q$ (either a long or a short position).[9] Once an AP takes the other side of an investor's trade, they can mean-revert their position through one of two mechanisms. First, another investor might arrive in the ETF market looking to trade in the opposite direction, such that the **AP** effectively intermediates the matching of a buyer and a seller on the ETF secondary market. Second, if no such match is found before a period of time has elapsed, which occurs with Poisson intensity η, the AP must engage in the creation-redemption process with the ETF issuer. This process involves the AP buying or selling stocks on the underlying stock market and then creating/redeeming units of the ETF.

Finally, there are a large number of competitive **dealers** in the underlying stock market, with mean-variance preferences and risk-aversion γ. To trade a quantity q of the underlying basket of stocks (where $q \geq 0$ implies the dealers buy the asset), dealers quote a competitive price as in Kyle (1985):

$$\pi\,(q) = \mu + \frac{\gamma}{2}\sigma^2 q \tag{3.1}$$

[9] In effect, this means each AP facilitates trading for up to one buyer and one seller.

I summarise the model notation in Table 3.1:

Table 3.1: Model notation

Exogenous parameters and their interpretation.	
Parameter	Definition
μ, σ	Expected return and variance, respectively, of the index (ETF basket).
c	ETF marginal cost for each unit of assets under management.
Γ	ETF fixed operating cost.
θ	Investor private value for the ETF basket.
Q	Trade quantity of individual investor.
λ_i	Investor i's arrival rate (inverse expected holding horizon).
η	Rate at which the AP unloads inventory on the underlying market.
Λ, ξ	Average and half-range of investor arrival rates.
Endogenous quantities and their interpretation.	
Variable	Definition
f_k, $k \in \{L, F\}$	Management fee for fund k.
$\pi(q)$	Supply schedule of dealers in the underlying securities market.
$p(q)$	Supply schedule of ETF authorized participant.
λ_k, $k \in \{L, F\}$	Aggregate investor arrival rate in fund k.
w_L, w_F	Market shares of fund H and L, respectively.
Φ	Ratio between leader's expected profit upon deterring and accommodating entry.

Timing. Figure 3.1 summarizes the sequence of events at each time t. Funds L and F decide whether to enter the market at $t = -2$ and $t = -1$, respectively. Upon entering, each fund sets a management fee.

At $t = 0$, investors observe management fees and allocate cash to at most one exchange-traded fund. At any point following $t = 0$, there is continuous trading: investors switch in and out of their positions by trading with **AP** (at random times, e.g., $\tilde{\tau}_0$) and authorized participants engage in creation/redemption with the underlying market (also at random times, e.g., $\tilde{\tau}_1$).

Equilibrium. I am looking for stable subgame-perfect Nash equilibria in pure and mixed strategies.

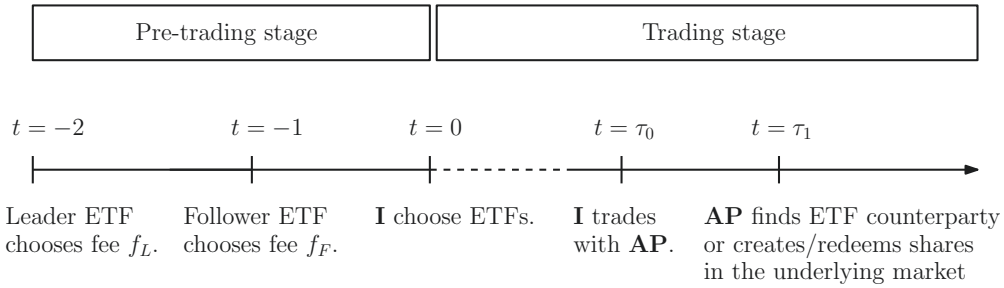


Figure 3.1: Model timing

Definition 1. An *equilibrium* of the game consists of (i) ETF entry decision and management fees at $t=-2$ (for fund **L**) and $t=-1$ (for fund **F**), (ii) investors' choice of funds at $t=0$, (iii) authorized participant's demand schedule on the ETF market, $p(q)$ and (iv) buy and sell prices quoted by the dealer on the underlying market, $\pi(q)$, where q stands in for traded quantity.

3.4.2 Equilibrium

3.4.2.1 Competitive market-making in the ETF market.

Let λ_k denote the aggregate intensity with which round-trip trades are completed in the secondary market for ETF k, where $k \in \{L, F\}$. Since Poisson processes are additive, it follows that λ_k is a sum of arrival intensities of all investors who choose ETF k, conditional on the trade direction: The $(2\xi)^{-1}$ term corresponds to the density of arrival rates across investors, and the first $\frac{1}{2}$ term reflects that I am considering only half of the potential market (either only buy orders or only sell orders). Consequently, λ_k can be interpreted as the Poisson intensity of the **AP** matching a buyer with a seller in ETF k.

I focus on linear equilibria. Let **AP** post a linear price schedule $p(q) = a + bq$. Perfect competition, together with the ability to post prices in real time, imply that the **AP's** expected profit on each marginal round-trip trade is zero, that is

$$\begin{aligned}\mathbb{E}\text{Profit}_{AP} = &\frac{\lambda_k}{\eta+\lambda_k}\underbrace{[a+bQ-(a-bQ)]}_{=2b\times Q} \\ &+\frac{\eta}{\eta+\lambda_k}\left[a+bQ-\left(\mu+\frac{\gamma}{2}\sigma^2 Q\right)\right], \quad k\in\{L,F\}\end{aligned} \tag{3.2}$$

Consider a representative trade in ETF k. With probability $\frac{\lambda_k}{\eta+\lambda_k}$, the **AP** matches a buyer and seller in the ETF market before turning to the underlying stock dealer, and obtains a round-trip profit of $2b \times Q$. Conversely, with probability $\frac{\eta}{\eta+\lambda_k}$, the authorized participant does not find a match on the ETF market before turning to the underlying market. In this case, he engages in the creation-redemption process and buys Q units of stock from the dealer at the price $\pi(Q)$ as defined in equation (3.1).

Since the competitive condition (3.2) should hold for any value of Q, I identify a and b by grouping the terms and find: $a = \mu$ and $b = \frac{\eta}{\eta+2\lambda_k}\frac{\gamma}{2}\sigma^2$. Therefore, the competitive price posted by the **AP** in ETF k is:

$$p_k(q) = \mu + \frac{\eta}{\eta + 2\lambda_k}\frac{\gamma}{2}\sigma^2 q \tag{3.3}$$

I note that the competitive **AP** price is adjusted for the expected price impact on the underlying stock market, if a natural counterparty is not found in the ETF market. The round trip cost for an investor equals twice the price impact for trading Q, that is

$$\text{Round-trip cost} = \frac{2\eta}{\eta + 2\lambda_k}\frac{\gamma}{2}\sigma^2 Q \tag{3.4}$$

Equation (3.4) implies that ETF bid-ask spreads depend on two components: (i) secondary market turnover in the ETF market, which is a function of investors' trading frequency, and (ii) liquidity in the underlying stock market (which is a function of dealers' risk aversion, trade size and fundamental volatility of the asset). The creation-redemption cost, although not explicitly modelled, belongs to the second category, as it is a fixed cost that the AP incurs every time he creates / redeems ETF shares. Note that the round trip cost for investors decreases in λ_k, that is ETFs with higher investor turnover tend to have lower trading costs, because in such ETFs, an AP is better able to match buyers and sellers rather than having to resort to creation-redemption.

3.4.2.2 Investors' ETF selection

Investor i derives private value θ from holding ETF k, and incurs the fee f_k (both θ and f_k are expressed per unit time). If the investor's trading intensity per unit of time is λ_i, driven by the exogenous changes in her private value, then her expected holding period is ${\lambda_i}^{-1}$. The investor also incurs a round-trip transaction cost given

in (3.4), as she establishes her ETF position Q, and then exits that position at the end of the holding period. Therefore, the investor's expected profit from investing in the ETF is:

$$\mathbb{E}\text{Profit}_i = \frac{1}{\lambda_i}(\theta - f_k)\,Q - \frac{2\eta}{\eta + 2\lambda_k}\frac{\gamma}{2}\sigma^2 Q \tag{3.5}$$

If both ETFs launch before time $t = 0$, investors trade off (i) fund management fees and (ii) liquidity (round-trip cost of trading) for the two ETFs. In turn, each fund k's liquidity depends on the frequency of creation/redemption and on the mass of investors that allocate money to it – that is, investors face a coordination problem.

To solve for the equilibrium ETF selection, I apply global game techniques as in Morris & Shin (2006) and Argenziano (2008). I conjecture, and verify afterwards, that there exists a threshold $\bar{\lambda}$ such that investors with $\lambda_i > \bar{\lambda}$ choose the ETF L and investors with $\lambda_i \leq \bar{\lambda}$ choose the ETF F.

Since investors do not know Λ, each investor uses her own arrival rate as an unbiased signal of the average inverse holding period, since $\mathbb{E}\Lambda = \lambda_i$. The marginal investor, with arrival rate $\bar{\lambda}$, is indifferent between the two ETFs. In equilibrium, since the marginal investor holds the correct "threshold belief", she estimates the aggregate trading intensities across the two ETFs as

$$\lambda_L = \frac{1}{2}\int_{\bar{\lambda}}^{\bar{\lambda}+\xi} \lambda \frac{1}{2\xi} \mathrm{d}\lambda = \frac{1}{8}\left(2\bar{\lambda} + \xi\right) \tag{3.6}$$

$$\lambda_F = \frac{1}{2}\int_{\bar{\lambda}-\xi}^{\bar{\lambda}} \lambda \frac{1}{2\xi} \mathrm{d}\lambda = \frac{1}{8}\left(2\bar{\lambda} - \xi\right) \tag{3.7}$$

The marginal investor believes that all other investors with a holding period lower (higher) than hers invest in ETF L (F, respectively). The larger the heterogeneity between investor trading intensities, ξ, the larger the difference between aggregate liquidity levels in the two ETFs. For the marginal investor with intensity $\bar{\lambda}$ to be indifferent between the two ETFs, the total cost of trading ETF L needs to be equal to the total cost of trading ETF F:

$$-\frac{1}{\bar{\lambda}} f_L Q - \frac{\eta}{\eta + 2\lambda_L}\gamma\sigma^2 Q = -\frac{1}{\bar{\lambda}} f_F Q - \frac{\eta}{\eta + 2\lambda_F}\gamma\sigma^2 Q \tag{3.8}$$

Rearranging terms, I obtain

$$\begin{aligned}\frac{1}{\bar{\lambda}}\left(f_L - f_F\right) &= \gamma\sigma^2\left(\frac{\eta}{\eta + 2\lambda_F} - \frac{\eta}{\eta + 2\lambda_L}\right) \\ &\approx \gamma\sigma^2\left[\left(1 - \frac{2\lambda_F}{\eta}\right) - \left(1 - \frac{2\lambda_L}{\eta}\right)\right], \end{aligned} \tag{3.9}$$

where the last step is a first-order Taylor series approximation of the fraction $\frac{\eta}{\eta + 2\lambda_k}$, $k \in \{L, F\}$. This final linearization step ensures tractability of the solution and is equivalent to assuming that η is "large enough" relative to λ_k. Lemma 1 obtains immediately by solving for $\bar{\lambda}$ in equation (3.9).

Lemma 1. (Investor ETF selection.) In equilibrium, conditional on ETF management fees, all investors with $\lambda_i > \bar{\lambda}$ choose the ETF L and all investors with $\lambda_i \leq \bar{\lambda}$ choose the ETF F, where

$$\bar{\lambda} = \frac{2\eta\left(f_L - f_F\right)}{\gamma\xi\sigma^2} \tag{3.10}$$

From Lemma 1, if the fee differential $f_L - f_F$ increases, then $\bar{\lambda}$ decreases. Intuitively, fewer investors choose the high-fee ETF, if the cost differential is larger. I make the conjecture, which I verify in equilibrium, that the leader ETF L is able to charge a higher management fee due its first-mover advantage.

3.4.2.3 Follower ETF fee-setting at $t = -1$

To understand the fee setting decisions, I work backwards starting from the follower's fee setting behavior given the leader's fee, before turning to the leader's fee setting behavior anticipating the follower's response.

From Lemma 1, all investors with a trading intensity between $\bar{\lambda}$ and $\Lambda + \xi$ (i.e., short-term investors) choose ETF L and, conversely, all investors with $\lambda_i \in \left[\Lambda - \xi, \bar{\lambda}\right]$ choose ETF F. The market shares of the two funds are w_L and w_F,

respectively, where

$$w_L = \frac{1}{2\xi}\left(\Lambda + \xi - \underbrace{\frac{2\eta\left(f_L - f_F\right)}{\gamma\xi\sigma^2}}_{\bar{\lambda}}\right) \text{ and}$$
$$w_F = \frac{1}{2\xi}\left(\frac{2\eta\left(f_L - f_F\right)}{\gamma\xi\sigma^2} - \left(\Lambda - \xi\right)\right) \tag{3.11}$$

However, ETFs only receive fees throughout the time period when investors hold the fund. At each point in time, since investors switch in and out of their positions at equal rates, each individual investor owns the ETF with probability one half. Therefore, the steady state average holding for fund k is $\frac{1}{2}w_k$, where w_k is defined in (3.11).

At $t = -1$ the follower ETF F decides whether or not to enter the market and, conditional on entering, sets a management fee f_F taking the leader's fee f_L as given. The fund receives inflows of Q for each investor it attracts during her holding period, and earns a profit margin of $f_F - c$ for each unit of time and assets under management. Finally, fund F incurs a fixed cost Γ per unit of time. I solve the follower ETF's problem:

$$\mathbb{E}\text{Profit}_F = \max_{f_F} Q\frac{1}{4\xi}\left(\frac{2\eta\left(f_L - f_F\right)}{\gamma\xi\sigma^2} - \left(\Lambda - \xi\right)\right)\left(f_F - c\right) - \Gamma, \tag{3.12}$$

which yields the following price reaction function to the leader's management fee:

$$f_F\left(f_L\right) = \frac{c + f_L}{2} - \frac{\gamma\xi\sigma^2\left(\Lambda - \xi\right)}{4\eta} \tag{3.13}$$

I note that fund F enters the market if and only if it can earn positive expected profit conditional on choosing the optimal fee, that is if

$$\mathbb{E}\text{Profit}_F = \frac{Q\left(2c\eta + \gamma\xi\sigma^2(\Lambda - \xi) - 2f_L\eta\right)^2}{32\gamma\eta\xi^2\sigma^2} - \Gamma > 0 \tag{3.14}$$

3.4.2.4 Leader fee-setting and entry deterrence

At $t = -2$, the leader ETF L sets its management fee taking as given the followers' reaction function $f_F\left(f_L\right)$ in equation (3.13). Importantly, the fee level is

strategically chosen to either accommodate or deter follower entry, depending on which option maximizes the leader's expected profit. I analyze the two options separately, and provide a condition for successful entry deterrance.

Leader accommodates entry. First, consider the scenario in which the incumbent ETF accommodates the follower's entry. That is, L solves

$$\mathbb{E}\text{Profit}_L = \max_{f_L} Q \frac{1}{4\xi} \left(\Lambda + \xi - \frac{2\eta \left(f_L - f_F \left(f_L \right) \right)}{\gamma \xi \sigma^2} \right) \left(f_L - c \right) - \Gamma \tag{3.15}$$

Solving equation (3.15) for the optimal leader fee, and then substituting it into the follower's reaction function (3.13), I obtain the equilibrium fees

$$f_L^\star = c + \frac{\gamma \left(\Lambda + 3\xi \right) \xi \sigma^2}{4\eta} > c \tag{3.16}$$

$$f_F^\star = c + \frac{\gamma \left(5\xi - \Lambda \right) \xi \sigma^2}{4\eta} \tag{3.17}$$

Note that the leader fee, f_L, is always above the marginal cost c, whereas the follower only generates a postive margin, if investor heterogeneity is large enough, that is if $\xi > \frac{\Lambda}{5}$. Further, I verify that $f_L \geq f_F$, as postulated before since $\Lambda > \xi$.

The leader fund's economic rents are higher when investors trade more (i.e. higher Λ) and when investors' trading intensities are very heterogeneous (i.e. higher ξ). Intuitively, the source of ETF L's economic rents lies in investors' coordination problem. For a high trading intensity investor, it is more costly to choose ETF F due to its high liquidity costs. However, F's liquidity is endogenously low due to the absence of high-turnover investors. It would be optimal for high-intensity investors as a group to switch to ETF F, as in that case, ETF F would become liquid, and investors would enjoy both lower management fee and lower round-trip transaction costs. However, there are infinitely many high-intensity investors, meaning they cannot coordinate. An individual high-turnover investor, however, does not benefit from switching to the low-fee ETF, because liquidity costs in ETF F would be larger than her savings from the lower management fees.

The low-fee, follower fund caters to investors with lower trading intensities, and therefore its economic rents are only positive, if there are enough of those investors in the spectrum $\lambda_i \in \left[\Lambda - \xi, \bar{\lambda} \right)$. The higher the dispersion of trading intensities, ξ, the more likely that ETF F breaks even.

It follows from equation (3.11) that the equilibrium market shares are

$$w_L^\star = \frac{1}{2\xi}\left(\Lambda + \xi - \frac{2\eta\left(f_L^\star - f_F^\star\right)}{\gamma\xi\sigma^2}\right) = \frac{3}{8} + \frac{\Lambda}{8\xi} > \frac{1}{2} \text{ and}$$
$$w_F^\star = \frac{1}{2\xi}\left(\frac{2\eta\left(f_L^\star - f_F^\star\right)}{\gamma\xi\sigma^2} - (\Lambda - \xi)\right) = \frac{5}{8} - \frac{\Lambda}{8\xi} < \frac{1}{2} \tag{3.18}$$

that is, the leader obtains a higher market share than the follower. Finally, I can evaluate the expected profits for the two funds at the equilibrium fee,

$$\mathbb{E}\text{Profit}_L = \frac{Q}{64\eta}\gamma\left(\Lambda + 3\xi\right)^2\sigma^2 - \Gamma \text{ and}$$
$$\mathbb{E}\text{Profit}_F = \frac{Q}{64\eta}\gamma\left(\Lambda - 5\xi\right)^2\sigma^2 - \Gamma \tag{3.19}$$

Leader deters entry. Second, consider a scenario where the incumbent ETF deters the follower's entry. To achieve this goal, ETF **L** sets the management fee such that the expected profit of the follower in (3.14) is zero, that is

$$f_L^{\text{deter}} = c + \frac{\gamma\xi\sigma^2(\Lambda - \xi)}{2\eta} + \frac{2\sqrt{2\gamma\Gamma}\xi\sigma}{\sqrt{\eta Q}} > c \tag{3.20}$$

Conditional on successful deterrence, the leader enjoys a market share of 100% and earns an expected profit of

$$\begin{aligned}\mathbb{E}\text{Profit}_L^{\text{deter}} &= \frac{Q}{2}\left(f_L^{\text{deter}} - c\right) - \Gamma \\ &= \frac{Q}{2}\left(\frac{\gamma\xi\sigma^2(\Lambda - \xi)}{2\eta} + \frac{2\sqrt{2\eta\Gamma}\xi\sigma}{\sqrt{\gamma Q}}\right) - \Gamma\end{aligned} \tag{3.21}$$

The leader has incentives to deter entry if and only if the expected profit in equation (3.21) is larger than the expected profit in equation (3.19). Lemma 2 formalizes the condition and provides further comparative statics on the two-fund equilibrium condition.

Lemma 2. The ETF leader accommodates the follower entry, and consequently a two-ETF economy obtains in equilibrium, if and only if $\Phi \leq 1$ where I define the profit ratio Φ as

$$\Phi \equiv \frac{16\xi\left(\Lambda - \xi + \frac{4\sqrt{2\eta\Gamma}}{\sqrt{\gamma Q}\sigma}\right)}{(\Lambda + 3\xi)^2} \tag{3.22}$$

It follows that Φ increases in the fixed cost Γ and decreases in the AUM parameter Q. Further, if $\Phi < 1$ and $5\xi > \Lambda$, the profit ratio Φ decreases in investor heterogeneity ξ.

Proposition 3.1 summarizes the equilibrium of the ETF competition game. I consider the economically interesting case where $5\xi > \Lambda$ such that the follower obtains positive market share in a two-fund equilibrium. If, conversely, $5\xi \leq \Lambda$, the leader ETF is an uncontested monopolist. Proofs are in Appendix 3.2.

Proposition 3.1. *(Two-fund equilibrium). A two-fund equilibrium obtains if*

$$\Phi \equiv \frac{16\xi\left(\Lambda - \xi + \frac{4\sqrt{2\eta\Gamma}}{\sqrt{\gamma Q}\sigma}\right)}{(\Lambda + 3\xi)^2} \leq 1 \tag{3.23}$$

that is, if the leader ETF accommodates the follower's entry. If the condition in Lemma 2 is true, then the following strategies constitute an equilibrium of the trading game:

(i) At $t = -2$ and $t = -1$, respectively, the two ETFs enter the market and post management fees

$$f_L^\star = c + \frac{\gamma\left(\Lambda + 3\xi\right)\xi\sigma^2}{4\eta} > c$$
$$f_F^\star = c + \frac{\gamma\left(5\xi - \Lambda\right)\xi\sigma^2}{4\eta}$$

(ii) At $t = 0$, investors choose funds following the threshold strategy defined in Lemma 1.

(iii) In the trading stage, authorized participants in ETF $k \in \{F, L\}$ post demand schedules

$$p_k\left(q\right) = \mu + \frac{\eta}{\eta + 2\lambda_k}\frac{\gamma}{2}\sigma^2 q \tag{3.24}$$

(iv) In the trading stage, dealers in the underlying market post demand schedules

$$\pi\left(q\right) = \mu + \frac{\gamma}{2}\sigma^2 q. \tag{3.25}$$

If $\Phi > 1$, then the leader ETF sets the deterrence fee f_L^{deter} in equation (3.20) and serves the entire market.

In equilibrium, ETF liquidity clienteles emerge endogenously as an outcome of the coordination game between investors. That is, all investors decide simultaneously between the two funds and the first-mover advantage in the fee-setting is enough to establish one of the ETFs as the more liquid one. In reality, exogenous factors would strengthen this equilibrium effect and "cement in" the advantage of the first fund at the market. For example, the first ETF to launch would have already established a clientele before its competitor arrival: the existing clientele reinforces network effects and helps the first ETF to be more liquid. Further, the incumbent ETF might have attracted more investor attention through advertisement or brand recognition, helping it attract its "sticky" investor clientele.

Figure 3.2 illustrates the condition for successful ETF entry. In particular, a two-ETF index is more likely, if either (i) there is more heterogeneity in investor holding periods ξ or (ii) the assets under management in the ETF industry (Q) are larger.

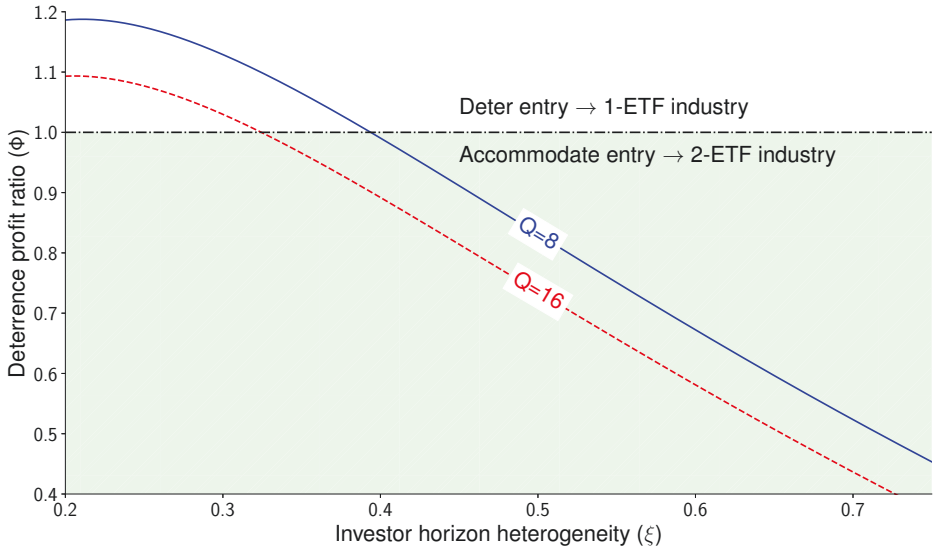


Figure 3.2: ETF entry, assets under management, and investor heterogeneity

This figure illustrates the deterrence profit ratio, Φ from Lemma 2, as a function of holding period heterogeneity and index assets under management (AUM). The leader ETF accommodates entry of a second ETF if and only if $\Phi \leq 1$. Parameter values: $c = 0.5$, $\Lambda = 1$, $\gamma = 4$, $\sigma = 5$, $\Gamma = 0.01$, $\eta = 2.5$.

3.4.3 Comparative statics and predictions

I first focus on the two-fund equilibrium to obtain predictions about the cross-section of funds (Predictions 1 to 4). Next, I turn to the conditions for a two-fund

equilibrium to exist (Predictions 5 and 6).

Prediction 1. Exchange-traded funds with higher management fees enjoy a higher market share and attract short-term investors. The market share differentials increase in the level of investor heterogeneity.

The L ETF obtains a "premium" market share (i.e., higher than one-half) as it manages to attract more investors. The lower the holding period heterogeneity ξ, the more investors cluster to the L ETF due to network effects. Intuitively, a high ratio $\frac{\Lambda}{\xi}$ implies that investors are concentrated in a relatively narrow spectrum of high trading intensities. If this is the case, ETF L captures a high market share, as the liquidity advantage is too dear to high-turnover investors, making them a sticky clientele. Conversely, a low $\frac{\Lambda}{\xi}$ implies that investors are dispersed across a wide range of trading intensities, that is, there is a higher pool of low-turnover investors who are willing to switch to the cheaper fund F.

Prediction 2. Exchange-traded funds with high management fees have a higher turnover than ETFs with low management fees.

Turnover in the two funds can be proxied by the aggregate investor arrival rate, λ_L and λ_F. Evaluated at the equilibrium fees, I find that the high-fee fund has a higher turnover than the low-fee fund, and the turnover difference increases in holding period heterogeneity. That is,

$$\begin{aligned} \lambda_F &= \frac{(5\xi - \Lambda)(7\Lambda - 3\xi)}{128\xi} \text{ and} \\ \lambda_L &= \frac{(\Lambda + 3\xi)(7\Lambda + 5\xi)}{128\xi}, \end{aligned} \tag{3.26}$$

with $\lambda_L - \lambda_F > 0$. The intuition follows from investors' optimal choice of ETFs. High-turnover investors find it optimal to choose ETF L, as they weigh liquidity savings more than the higher management fee. In equilibrium, ETF L has high turnover, reflecting an investor clientele that trades often, while ETF F has low turnover, reflecting the investor clientele that trades less actively.

Prediction 3. High-fee ETFs have higher liquidity than low-fee ETFs.

Liquidity in the two funds can be proxied by the costs of round-trip trade (i.e., absolute bid-ask spread). From equation (3.4), the round-trip cost is negatively correlated with the aggregate investor arrival rate, $\lambda_k, k \in \{L, F\}$.

Prediction 4. Exchange-traded funds with high management fees enjoy higher profits than ETFs with low management fees, and the profit differential increases in the amount of holding-period heterogeneity, ξ.

I note that, since $\xi > 0$, it is always the case that $\mathbb{E}\text{Profit}_L > \mathbb{E}\text{Profit}_F$.

Prediction 5. A two-ETF economy obtains only if there is enough heterogeneity in investors' holding periods.

A two-ETF economy only emerges if both funds have positive market share in equilibrium. From equation (3.18), this is only the case if $\xi > \frac{\Lambda}{5}$. Also, from Lemma 2, the entry condition for the follower fund is relaxed if investor heterogeneity increases (i.e., Φ decreases in ξ).

Prediction 6. A two-ETF economy obtains only if the combined assets under management of the index funds are large relative to the fixed costs.

From Lemma 2, I note that the profit ratio Φ increases in Γ and decreases in Q. Therefore, the two-ETF economy is more likely for a large Q and a small Γ. Intuitively, the size of the ETF industry needs to be large enough to generate economies of scale sufficient to cover the fixed cost. Otherwise, the follower fund chooses not to enter the market.

Figure 3.3 illustrates the model comparative statics as discussed above. In particular, note that in a two-ETF economy, each fund will have a lower turnover than under a monopoly that aggregates all investors, that is

$$\lambda_{\text{monopoly}} = \frac{1}{2}\int_{\Lambda-\xi}^{\Lambda+\xi} \lambda \frac{1}{2\xi} \mathrm{d}\lambda = \frac{\Lambda}{2} \tag{3.27}$$

3.5 Welfare implications

In this section, I assess the welfare implications of the oligopolistic, 2-fund equilibrium in Proposition 3.1. First, I note that both authorized participants and dealers in the model are competitive and earn zero expected profits. Consequently, I can define welfare as the sum of investor and fund expected utilities.

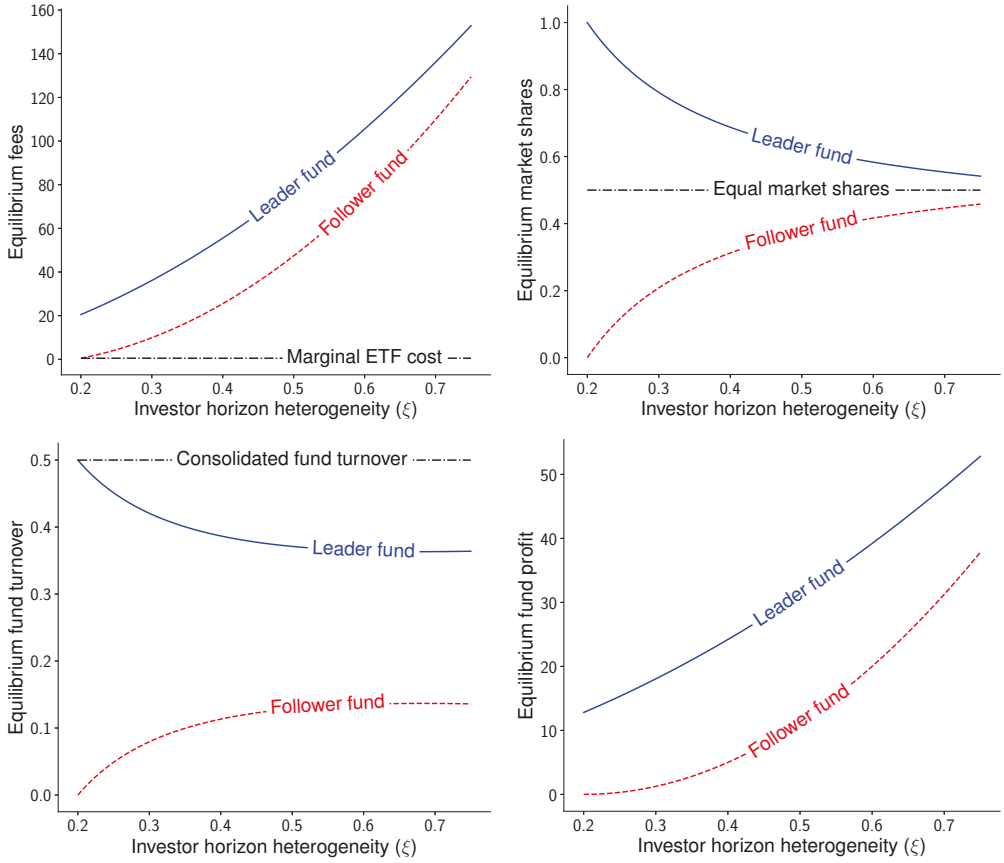


Figure 3.3: Equilibrium comparative statics

This figure illustrates the equilibrium outcomes as a function of the heterogeneity in investment horizons (ξ). In particular, I plot equilibrium fees (upper left), market shares of the two funds (upper right), fund turnover (lower left), and fund profit (lower right). Parameter values: $c = 0.5$, $\Lambda = 1$, $\gamma = 4$, $\sigma = 5$, $\Gamma = 0$, $\eta = 2.5$, $q = 4$.

3.5.1 Welfare benchmark

A natural benchmark for the welfare analysis is an economy with a single and competitive ETF. In such an economy, all investors allocate their wealth to a single fund. The one-fund benchmark is intuitive as it eliminates the coordination problem across investors, and allows for the largest possible trading network. Furthermore, the benchmark economy is free of imperfect competition considerations: the unique fund is competitive and earns zero profit in expectation.

First, I solve for the competitive fee (f_B) posted by a fund aggregating the entire universe of investors. The unique fund attracts the entire universe of investors, each of whom is equally likely to hold or not to hold the ETF at any given point in time. The management fee that sets the expected profit of the single ETF to

zero is given by

$$\begin{aligned}\mathbb{E}\text{Profit}_{\text{Single ETF}} &\equiv \frac{Q}{2}\left(f_B - c\right) - \Gamma = 0 \Rightarrow \\ &\Rightarrow f_B = c + 2\frac{\Gamma}{Q}\end{aligned} \tag{3.28}$$

Note that the benchmark fee exceeds the marginal cost c, since the ETF needs to also cover the fixed cost Γ. Since the ETF earns zero expected profit, welfare in the benchmark economy corresponds to the aggregate investors utility. I compute the Poisson rate of roundtrip trades for a typical investor in the single ETF, that is,

$$\lambda_B = \frac{1}{2}\int_{\Lambda-\xi}^{\Lambda+\xi} \lambda_i \frac{1}{2\xi} \mathrm{d}\lambda_i = \frac{1}{2}\Lambda \tag{3.29}$$

Finally, I aggregate investor utility per unit of time across the distribution of holding horizons. From equation (3.5), the investors' expected utility per unit of time is

$$\mathbb{E}\text{Utility}_{\mathbf{I}}^{\text{time unit}} = \frac{1}{2}\left(\theta - f\right) + \frac{\lambda_i}{2}\frac{2\eta}{\eta + 2\lambda_k}\frac{\gamma}{2}\sigma^2 Q \tag{3.30}$$

At all times, investors are equally likely to hold the ETF or not, and accumulate an expected private value net of fees of $\frac{1}{2}\left(\theta - f\right)$. Further, an investor with switching intensity λ_i completes an expected number of $\frac{1}{2}\lambda_i$ round-trips per unit of time and incurs the corresponding cost. I evaluate the investors' expected utility in equation (3.30) at the management fee in (3.28) and the turnover in equation (3.29) to obtain the benchmark welfare:

$$\begin{aligned}\text{Welfare}_{\text{B}} &= Q\int_{\Lambda-\xi}^{\Lambda+\xi}\left[\frac{1}{2}\left(\theta - f_B\right) - \frac{\eta}{\eta+\Lambda}\frac{\lambda_i}{2}\gamma\sigma^2\right]\frac{1}{2\xi}\mathrm{d}\lambda_i \\ &= \frac{Q}{2}\left[\theta - f_B - \frac{\eta\Lambda}{\eta+\Lambda}\gamma\sigma^2\right] \\ &= \frac{Q}{2}\left[\left(\theta - c\right) - \frac{\eta\Lambda}{\eta+\Lambda}\gamma\sigma^2\right] - \Gamma\end{aligned}$$

3.5.2 Equilibrium welfare

From equation (3.30), I aggregate the equilibrium investors' utility per unit of time for fund $k \in \{L, F\}$ by integrating over the set of investors choosing fund k, Ω_k:

$$\begin{aligned}\mathbb{E}U_{\text{Investors},k} &= Q\int_{\lambda\in\Omega_k}\frac{1}{2}\left(\theta - f_k\right)\frac{1}{2\xi}\mathrm{d}\lambda - Q\gamma\sigma^2\int_{\lambda\in\Omega_k}\frac{\eta}{\eta+2\lambda_k}\frac{\lambda}{2}\frac{1}{2\xi}\mathrm{d}\lambda \\ &= \frac{Q}{2}w_k\left(\theta - f_k\right) - \frac{Q}{2}\gamma\sigma^2\int_{\lambda\in\Omega_k}\frac{\eta}{\eta+2\lambda_k}\lambda\frac{1}{2\xi}\mathrm{d}\lambda. \end{aligned} \tag{3.31}$$

The expected profit of fund k is equal to its profit margin per time-investor, aggregated across holding periods and the cross-section of investors, minus the fixed costs:

$$\mathbb{E}\text{Profit}_k = \frac{Q}{2}w_k\left(f_k - c\right) - \Gamma. \tag{3.32}$$

For each fund k, I can compute a "partial welfare" measure by summing up the expected investor utility in (3.31) and fund utility in (3.32), that is

$$\text{Welfare}^k_{\text{eqm}} = \underbrace{\frac{Q}{2}w_k\left(\theta - c\right)}_{\text{gains from trade}} - \underbrace{\frac{Q}{2}\gamma\sigma^2\int_{\lambda\in\Omega_k}\frac{\eta}{\eta+2\lambda_k}\lambda\frac{1}{2\xi}\mathrm{d}\lambda}_{\text{liquidity cost}} - \underbrace{\Gamma}_{\substack{\text{fixed}\\\text{costs}}} \tag{3.33}$$

I first note that management fees "wash out" in the welfare computation, as they represent transfers from investors to the fund. The partial welfare in equation (3.33) consists of three components: (i) the gains from trade, proportional to the difference between investors' private value for the ETF and the marginal cost of the ETF, (ii) liquidity costs incurred when the authorized participant turns to the underlying stock market and performs creation / redemption, and (iii) fixed costs of running the ETF.

The equilibrium welfare obtains through aggregating the quantity in (3.33) for both funds and their respective investor sets, that is:

$$\text{Welfare}_{\text{eqm}} = \frac{Q}{2}\left[\left(\theta - c\right) - \frac{\gamma\sigma^2}{2\xi}\left(\int_{\bar{\lambda}}^{\Lambda+\xi}\frac{\eta\lambda_i}{\eta+2\lambda_L}\mathrm{d}\lambda_i + \int_{\Lambda-\xi}^{\bar{\lambda}}\frac{\eta\lambda_i}{\eta+2\lambda_F}\mathrm{d}\lambda_i\right)\right] - 2\Gamma \tag{3.34}$$

I compare the equilibrium welfare with the benchmark to obtain a measure for the welfare loss in the oligopolistic equilibrium

$$\Delta\text{Welfare} = \Gamma + \underbrace{\frac{Q}{2}\gamma\sigma^2\left[\frac{1}{2\xi}\left(\int_{\bar{\lambda}}^{\Lambda+\xi}\frac{\eta\lambda_i}{\eta+2\lambda_L}\mathrm{d}\lambda_i + \int_{\Lambda-\xi}^{\bar{\lambda}}\frac{\eta\lambda_i}{\eta+2\lambda_F}\mathrm{d}\lambda_i\right) - \frac{\eta\Lambda}{\eta+\Lambda}\right]}_{\text{Network inefficiencies}} \tag{3.35}$$

Two channels reduce welfare in equilibrium. First, the entry of a second ETF generates an additional cost. Second, and more importantly, there are network inefficiencies due to splitting up liquidity. Each fund aggregates only a fraction of investors, reducing the probability of finding a counterparty on the ETF market and, in turn, increasing the probability of costly creation/redemption by the authorized participants.

From Figure 3.4, the welfare loss from an oligopolistic ETF market increases in horizon heterogeneity. A higher dispersion in investor horizons enhances the value of having a single "liquidity pool" concentrating everyone. Further, a higher intensity of creation-redemption activity in the ETF industry (i.e., a higher η) augments the transaction costs for authorized participants, which are evenutually passed on to the investors who are not matched on the ETF market.

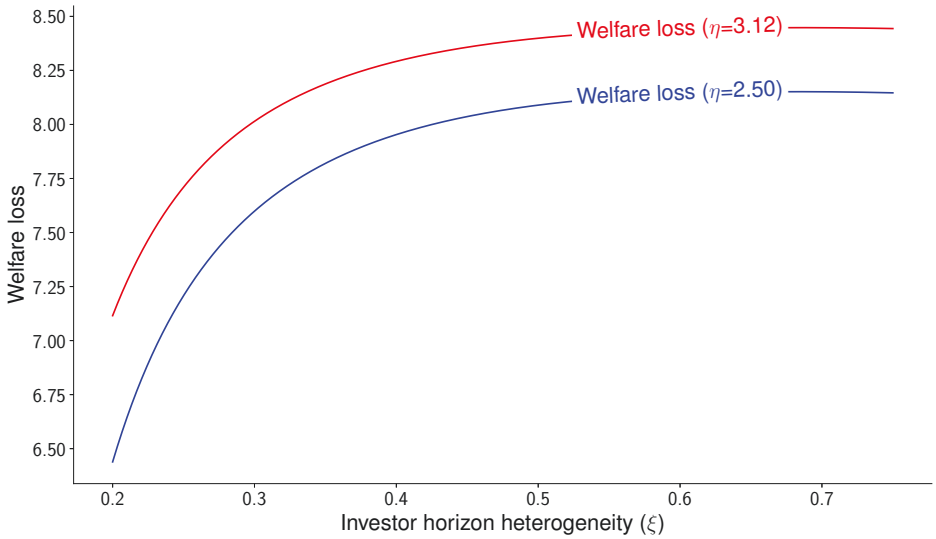


Figure 3.4: Welfare loss in the oligpolistic equilibrium

This figure illustrates the welfare loss in equation (3.35), as a function of holding period heterogeneity and the intensity of creation-redemption activity. Parameter values: $c = 0.5$, $\Lambda = 1$, $\gamma = 4$, $\sigma = 5$, $\Gamma = 0.01$, $\theta = 4$, $q = 4$.

3.6 Empirical analysis

3.6.1 Data and descriptive statistics

I obtain daily data from ETF Global, which covers the full universe of US-domiciled ETFs. Daily ETF spreads and prices are from the Center for Research in Security Prices (CRSP) database, maintained by Wharton Research Data Services (WRDS). I restrict the sample to equity ETFs traded on US markets, excluding ETNs (exchange-traded notes), leveraged or inverse ETFs, and ETFs that are hedged versions of the original fund. The full sample includes 1,035 equity ETFs traded in the US in 2017. Since the model predictions concern static equilibrium relations, I test them in the cross-section of ETFs. Hence, all variables are ETF-level averages.

To test the model predictions, I identify indices tracked by multiple ETFs. The resulting sample contains 60 ETFs based on 24 indices. Each ETF in this sample shares the underlying index with at least one other ETF. See Appendix 3.1 for the list of ETFs and their characteristics.

I cross-validate the list of ETFs by using two alternative approaches. Firstly, I manually check each ETF on ETF.com, an online provider of ETF statistics that allows me to identify ETFs with the same portfolio exposure. Secondly, I compare my list with that in the appendix of Box, Davies & Fuller (2018), who use Morningstar data in their analysis. These checks confirm that my sample correctly identifies ETF pairs (or trios) that invest in the same portfolio of stocks.

Table 3.2 provides descriptive statistics. The full sample covers 1,035 ETFs with a combined AUM of $2.26 trillion, and combined daily value traded of $52.78 billion. The subsample with multiple ETFs per index accounts for 36% of the total AUM and 47% of total daily dollar volume of all equity ETFs in the sample. That is, 60 ETFs that have competitors tracking the same index account for almost half of daily volumes and over one third of AUM of 1,035 equity ETFs in the sample. This highlights how liquidity and AUM concentrates in several major ETFs in this market.

The descriptive statistics in Table 3.2 distinguishes between multiple-ETF indices (Panel A) and one-ETF indices (Panel B). In Panel A (Panel B), the average ETF charges annual management fees of 22.67 bps (50.13 bps). The difference

Table 3.2: Descriptive statistics

This table reports descriptive statistics for the variables used in regression analysis. 60 ETFs in Panel A are based on 24 indices, and cover total AUM of $823.58 billion and total daily dollar volume of $24.79 billion. 975 ETFs in Panel B are based on 975 indices, and cover total AUM of $1,445.42 billion and total daily dollar volume of $27.98 billion. All variables are calculated from the daily frequency data for the year 2017. MER is net expense ratio per annum, relative spread is absolute bid-ask spread divided by midpoint, turnover is annualized percentage ratio of daily dollar volume divided by assets under management (AUM).

	Mean	StdDev	25th pctl	50th pctl	75th pctl
Panel A. Indices with multiple ETFs per index					
MER, bps	22.67	11.80	15.11	20.00	26.66
Relative Spread, bps	6.19	4.92	3.45	4.88	6.68
Turnover, %	331.21	283.01	191.47	244.11	328.47
Number of Constituents	854.52	716.24	338.46	590.67	1,176.98
AUM, $bn	34.32	85.85	4.79	8.70	34.34
Daily $Volume, $mln	1,033.00	3,801.70	318.00	895.00	200.10
Panel B. Indices with one ETF per index					
MER, bps	50.13	40.25	35.00	48.00	62.20
Relative Spread, bps	29.02	66.83	6.13	14.37	31.81
Turnover, %	534.35	1,458.76	171.21	289.54	495.36
Number of Constituents	276.50	532.47	44.16	100.49	276.48
AUM, $bn	1.48	5.44	0.02	0.12	0.69
Daily $Volume, $mln	28.80	177.90	0.20	0.90	4.60

in relative spreads is even wider. Investors pay 6.19 bps on average for a round-trip transaction in ETFs in multiple-ETF indices, compared to 29.02 bps, if an index is tracked by one ETF only. Panel A ETFs also tend to be broader (854 constituents, compared to 276 in Panel B), have significantly larger AUM ($34.32 billion, compared to 1.48 billion), and widely larger daily dollar volumes ($1.03 billion vs $0.03 billion).

3.6.2 OLS regression results

Do high-fee ETFs have higher turnover and lower spreads, compared to same-index competitors? From Figure 3.5, the answer is yes. The top right graph plots ETF turnover against fees. To compare ETFs within an index, I subtract the index-level mean from turnover and fee of each ETF. The graphs show that for same-index

ETFs, higher turnover (and higher volumes) are associated with higher fees. In line with the model Prediction 2, the ETF with high-turnover investor clientele charges higher fees than the low-turnover competitor. Similarly, narrower spreads are associated with higher fees for same-index ETFs, in line with Prediction 3. Higher investor turnover makes the ETF more liquid, implying lower round-trip transaction costs, and that allows issuers to charge higher fees in equilibrium.

Also in line with the model intuition, first-mover ETFs (in dark blue in Figure 3.5) tend to be the most liquid ones and have the highest fees among same-index competitors. Using data in Appendix 3.1, in 79% of cases the first mover ETF in a given index is the most liquid one among competitors (i.e., has the narrowest spreads / the highest dollar volume), in 75% of cases it has the highest fees, and in 79% of cases the highest AUM.

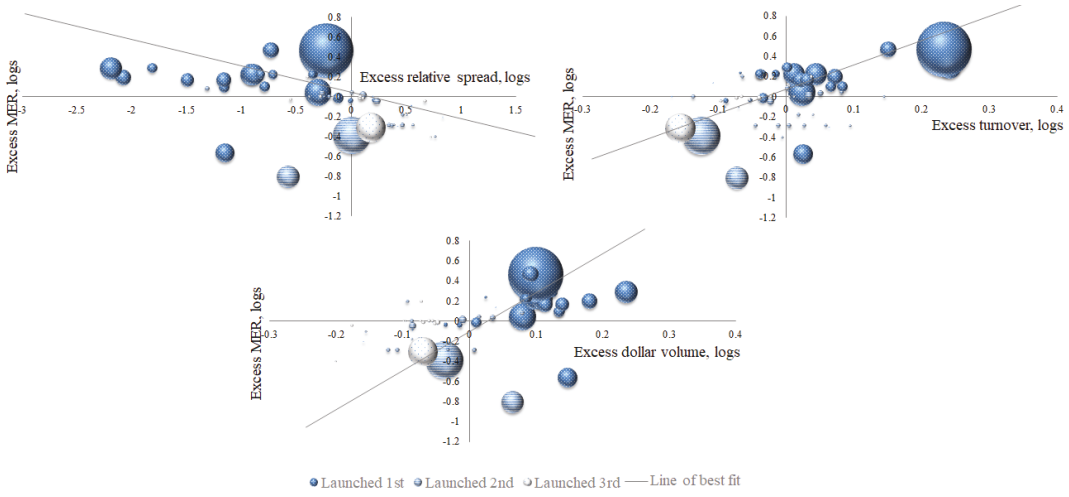


Figure 3.5: Management expense ratios vs liquidity of same-index ETFs

This figure plots demeaned MERs against demeaned liquidity measures on a log-scale. Demeaned MER is the percentage difference between this ETF's MER and the index-level average MER. The log transformation is Ln(1+% Excess MER) and Ln(1+% Excess Liquidity Measure). The bubble size is proportional to assets under management (AUM) of a given ETF. The sample contains 60 same-index ETFs based on 24 unique indices. The variables are calculated from the daily frequency data and averaged per ETF over the year 2017. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is annualized percentage ratio of daily dollar volume divided by assets under management (AUM).

How much do high-fee ETFs charge for their premium liquidity? To quantify this effect, I run cross-sectional OLS regressions with index fixed effects. The baseline regression model is as follows:

$$MER_i = \alpha + \beta_1 LIQ_i + \beta_2 TrackingError_i + \mu_{IND_i} + \varepsilon_i \tag{3.36}$$

Where LIQ_i is relative spread, log-turnover or log-dollar volume of ETF i (depending on the model specification), MER_i is the net expense ratio of ETF i, $TrackingError_i$ is the tracking error of ETF i, and μ_{IND_i} is the index fixed effect. There are 23 index dummies (the omitted dummy is S&P MidCap 400 Value Index), and each index has multiple ETFs tracking it. Index fixed effects capture within-index variation, akin to that plotted in Figure 3.5. In other words, regressions with index fixed effects ask: "How much extra fee are investors paying, on average, for an extra unit of liquidity in ETFs with identical holdings?" I control for tracking error to rule out other factors that could be different between ETFs that track the same underlying index.[10] I report the regression results in Table 3.3.

From Table 3.3 (Panel A), ETF investors pay 0.51 bps higher fee for 1 bps narrower relative spread, conditional on index exposure. This result is statistically significant at 1% level, and economically meaningful. It suggests that investors value ETF liquidity. In line with Prediction 3 from the theory model, lower costs of round-trip transaction are associated with higher fees in same-index ETFs.

For convenience of interpretation, I run regressions for turnovers and dollar volumes using log-transformed variables. The results suggest that doubling ETF turnover is associated with 24% increase in fees, and doubling dollar volume is associated with 1.15 bps increase in fees. In line with Prediction 2, high-turnover ETFs are charging higher fees, as they cater to liquidity-sensitive investors who trade a lot.

ETFs on the same index might have different tracking errors. If investors care about how reliable an ETF is in replicating the index, they might pay higher fees for ETFs with lower tracking errors. Panel A regression models (2), (4) and (6) confirm that is indeed the case. However, the economic magnitude of tracking error coefficients is low. Tracking error is six times less important than spreads in commanding ETF fee discounts, two times less important than turnover, and over ten times less important than dollar volume.

[10] In the Robustness Checks section, I show that same-index ETFs in my sample are also identical in other dimensions (including legal structure, tax treatment, dividend distribution policy etc.).

Table 3.3: Cross-sectional regressions

This table reports results for six models cross-sectional OLS regressions. The dependent variable in ETF fee (MER) in basis points, and independent variables are reported in the first column. Panel A reports the results for simple OLS regressions. Panel B reports the results for AUM-weighted least squares regressions. The sample contains 60 same index ETFs based on 24 indices. All variables are calculated from the daily frequency data and averaged per ETF over the year 2017. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is annualized percentage ratio of daily dollar volume divided by assets under management (AUM), tracking error is the annualized standard deviation of the difference in daily returns between an ETF and its benchmark index. T-statistics are reported in parentheses. ***,**,* indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	MER (1)	MER (2)	LogMER (3)	LogMER (4)	MER (5)	MER (6)
Panel A. Ordinary least squares regressions						
Intercept	23.12^{***} (8.37)	22.97^{***} (8.62)	2.06^{***} (4.98)	1.83^{***} (4.82)	1.29 (0.19)	2.28 (0.36)
Relative Spread	-0.55^{***} (-4.01)	-0.51^{***} (-3.83)				
Log Turnover			0.19^{**} (2.42)	0.24^{***} (3.21)		
Log Dollar Volume					1.21^{***} (3.12)	1.15^{***} (3.11)
Tracking Error		-0.08^{*} (-1.92)				
Log Tracking Error				-0.12^{***} (-3.00)		-0.09^{**} (-2.13)
Adjusted R^2	85%	86%	73%	78%	83%	84%
Fixed effects	Index	Index	Index	Index	Index	Index
Panel B. AUM-weighted least squares regressions						
Intercept	26.38^{***} (4.37)	26.44^{***} (4.33)	1.46^{***} (5.58)	1.44^{***} (5.54)	-2.07 (-0.31)	-1.76 (-0.26)
Relative Spread	-0.69 (-1.11)	-0.66 (-1.03)				
Log Turnover			0.37^{***} (14.66)	0.37^{***} (14.70)		
Log Dollar Volume					1.57^{***} (5.01)	1.56^{***} (4.95)
Tracking Error		-0.25 (-0.49)				
Log Tracking Error				-0.15 (-1.29)		-0.26^{**} (-0.68)
Adjusted R^2	74%	73%	93%	93%	84%	84%
Fixed effects	Index	Index	Index	Index	Index	Index

Panel A regression results quantify the trade-offs between fees and liquidity for an average ETF. However, because sample ETFs are different in size, it's worthwhile to also make a comparison per average dollar invested. In Panel B, I do so using AUM-weighted least squares regressions. For an average dollar of AUM, fees are 37% higher for each doubling of turnover, and 1.56 bps higher for each doubling of dollar volume. So per-dollar effects are more economically significant with respect to dollar volumes and turnovers. That is, high-AUM ETFs are able to extract

higher fees (compared to low-AUM ETFs), per unit of turnover or dollar volume. However, that is not the case for the spread-MER relation, if evaluated on a per-dollar basis. That is not surprising, as high-AUM ETFs like SPY and IVV tend to be ultra-liquid, with spreads constrained to one tick (or even zero, if midpoint crosses are allowed). Therefore, spreads cannot be lowered further in these ETFs.

3.6.3 Probit regression results

Not many indices are tracked by multiple ETFs. The model predicts that indices with multiple ETFs emerge only when investors in a given index have sufficiently different trading intensities (Prediction 5) and when the combined AUM in a given index is sufficiently high (Prediction 6). I test these predictions in a probit model for US equity ETFs.

Table 3.4 estimates the probability of observing an index with multiple ETFs. In line with the theory model Prediction 6, high AUM is a strong predictor of obtaining a multi-ETF index. Recall that high AUM makes it more likely that the low-fee ETF achieves sufficient economies of scale to break even and survive in equilibrium. The higher the total AUM of ETFs following the index, the easier it is to cover fixed costs, even if the follower fund has lower market share and charges lower fees.

Some indices are more conducive to forming ETF investor clienteles than others. In particular, major indices such as S&P, Russell and MSCI attract the high-turnover clientele of institutional investors, who trade index derivatives, as well as ETFs. These investors tend to trade intensively, as they use ETFs for hedging, arbitrage, or short-term tactical allocations. On the other end of the spectrum are buy and hold investors, who rely on broad-market ETFs for asset allocation reasons. Because S&P, Russell and MSCI are likely to appeal to investors from both ends of the spectrum, it is likely that investor heterogeneity in these indices is high.

Table 3.4 results suggest that ETFs following S&P-, Russell- and MSCI-branded indices are more likely to have competitors in the same index. That is consistent with greater investor heterogeneity in multi-ETF indices. As the model (Prediction 5) suggests, multiple ETFs per index are more likely to emerge when investor

clienteles are sufficiently different. This heterogeneity allows the high-fee issuer to keep his fees high and serve the fee-insensitive high-turnover clientele.

Probit regressions also control for other ETF characteristics, such as dollar volumes, spreads, number of constituents, and ETF issuer. The results are as expected: ETFs with direct competitors in the same index are more widely traded, have narrower spreads, are more likely to be issued by the top 3 issuers (State Street, Black Rock or Vanguard), and track indices with more constituents.

Because single-ETF indices significantly outnumber multi-ETF indices, as a robustness check, I run a similar probit regression for a narrower sample. Instead of using all 975 single-ETF indices, I select 150 ETFs from the single-ETF sample at random. Then, I re-run the probit model on data for 150 single-ETF indices + 60 multi-ETF indices. The results (in Panel B of Table 3.4) are similar across the two sampling approaches.

3.6.4 Robustness checks

The regression results for liquidity-fee trade-offs (in Table 3.3) assume that same-index ETFs are very similar across all dimensions, but fees and secondary market liquidity. One can argue that those ETFs, even if tracking the same underlying index, might still differ in other characteristics. For example, ETFs might differ in terms of tracking errors, primary market liquidity, availability of (and open interest in) ETF options, legal structure, dividend reinvestment policy, and tax treatment.

ETFs in this analysis have different tracking errors. I control for these differences in baseline results reported in Table 3.3. For ETFs within the same index, tracking errors are negatively related to ETF fees, but the effect is two to ten times weaker than the liquidity-fee relation. Tracking error is a standard deviation of tracking difference, hence it captures the lack of reliability in ETF's tracking performance relative to underlying index. Alternatively, ETF investors might be interested in ETF premium or discount, which is the difference between the index return and the ETF return, net of fees. Model (2) in Table 3.5 shows that the baseline results are robust to controlling for ETF premium / discount.

ETFs also differ in their primary market liquidity, which is a function of underlying stocks' liquidity and the creation-redemption fee. As underlying stocks are the

Table 3.4: Probit regressions

This table reports results for probit regressions. The dependent variable is the probability of observing multiple ETFs in a given index. Independent variables are reported in the first column. Panel A sample contains 24 indices with multiple ETFs and 975 indices with one ETF. Panel B sample contains 24 indices with multiple ETFs and 150 randomly selected indices with one ETF. All variables are calculated from the daily frequency data and averaged per index over the year 2017. Dollar volume and AUM are in $ billion, number of constituents is in hundreds, relative spread is in basis points, major index dummy applies to MSCI, S&P and Russel, top 3 ETF issuers are Vanguard, BlackRock and State Street. Chi-squared statistics are reported in parentheses. ***,** ,* indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	Prob (1)	Prob (2)	Prob (3)	Prob (4)
Panel A. All equity ETFs				
Intercept	-2.87^{***} (127.11)	-2.87^{***} (126.87)	-2.37^{***} (68.87)	-2.32^{***} (73.69)
Dollar Volume		-0.07 (0.06)	0.44^{**} (4.21)	0.42^{**} (3.83)
Relative Spread			-0.03^{**} (5.32)	-0.03^{**} (4.10)
Major Index Dummy	0.56^{**} (3.60)	0.56^{**} (3.62)	0.56^{**} (4.32)	
AUM	0.04^{***} (23.76)	0.04^{***} (17.22)		
Top3 Issuer Dummy	0.51^{*} (3.12)	0.51^{***} (3.07)	0.56^{**} (4.09)	0.82^{***} (10.24)
Number of Constituents				0.02^{**} (4.94)
Panel B. Randomly selected equity ETFs				
Intercept	-2.18^{***} (46.38)	-2.18^{***} (46.19)	-1.61^{***} (15.83)	-1.48^{**} (14.46)
Dollar Volume		0.72 (0.28)	2.26^{*} (3.17)	2.31^{**} (3.46)
Relative Spread				-0.03 (2.40)
Major Index Dummy	0.55 (2.11)	0.54 (2.04)	0.54 (2.19)	
AUM	0.04^{***} (9.43)	0.03^{**} (4.42)		
Top3 Issuer Dummy	0.75^{**} (3.87)	0.74^{**} (3.67)	0.59 (2.44)	0.87^{***} (6.03)
Number of Constituents				0.01 (0.17)

same for same-index ETFs, the only difference might arise from the fee that issuers charge for every creation and redemption performed by authorized participants. Creation-redemption fees are typically passed on to the end investor in the form of bid-ask spread. Therefore, I substitute the spread measure for creation-redemption fee in model (3) of Table 3.5.[11] The coefficient on creation-redemption fee is

[11]I calculate creation-redemption cost in the following manner: $Cost_{CR} = \frac{c}{NOSH*P}$, where c is the fixed dollar cost per creation-redemption unit, P is the market price of one ETF share, $NOSH$ is the number of ETF shares in each ETF unit. This essentially measures a fraction (or bps) cost of creating / redeeming ETF shares, standardized per one ETF share (as creation/redemption unit sizes differ across ETFs). Creation-redemption costs are from Bloomberg.

similar in magnitude to that on bid-ask spread in baseline regressions: 1 bps higher creation-redemption cost is associated with 0.4 bps lower MER (not statistically significant). In the baseline regression (Table 3.3), 1 bps higher spread is associated with 0.51 bps lower MER (statistically significant). In other words, creation-redemption cost in the primary market is a weak proxy for bid-ask spread charged by market makers in the secondary market.

Table 3.5: Robustness checks

This table reports the results for five cross-sectional regression models with index fixed effects. The dependent variable is ETF fee (MER) in bps. The independent variables are in the first column. The sample contains 60 ETFs based on 24 indices. All variables are calculated from the daily frequency data and averaged per ETF over the year 2017. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is annualized percentage ratio of daily dollar volume divided by assets under management (AUM), tracking error is the standard deviation of tracking difference, tracking difference (or premium / discount) is the difference in daily returns between an ETF and its benchmark index, creation unit cost is the ratio of creation fee to the value of ETF creation basket, D_{UIT} is the dummy for unit investment trust. T-statistics are reported in parentheses. ***,** ,* indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	MER (1)	MER (2)	MER (3)	LogMER (4)	MER (5)
Intercept	23.05*** (8.71)	23.99*** (10.06)	20.80*** (6.45)	2.65*** (4.80)	20.63*** (4.15)
Relative Spread	−0.54*** (-3.05)	−0.48*** (-2.88)			
Log Turnover				0.14 (1.60)	
Log Dollar Volume					1.19*** (3.89)
Tracking Error	0.00*** (4.60)		0.00 (-0.42)		
Log Tracking Error				-0.06 (-1.51)	-0.41 (-0.65)
D_{UIT}	5.14*** (28.76)	5.16*** (29.05)	5.64*** (17.34)	0.56*** (2.43)	1.85 (1.09)
Premium / Discount		-0.33 (-1.43)			
Creation Unit Cost			-0.40 (-1.47)		
Adjusted R^2	91%	91%	88%	85%	90%
Fixed effects	Index	Index	Index	Index	Index

In a broad sense, ETF liquidity pool encompasses not only primary and secondary market liquidity, but also the liquidity of related instruments, such as index futures, index options, and ETF options. The availability of index-based derivatives is the same for same-index ETFs, but the availability of ETF options is not. Appendix 3.1 reports the options open interest for the sample ETFs. From the Appendix, the most liquid ETF per category is also the one with ETF options available.

Options are launched on ETFs that are most widely traded. This further reinforces ETF liquidity, and is consistent with the model logic of self-perpetuating liquidity effects.

There are two legal structures that apply to the sample of same-index ETFs: 58 of them are open-end investment companies, and two (SPY and MDY) are unit investment trusts (UITs). ETFs that are structured as UIT reinvest dividends quarterly rather than daily, and therefore have wider tracking errors. I introduce a UIT dummy variable in robustness checks, and the results do not materially change, compared to baseline regressions in Table 3.3. Apart from UIT ETFs, all other ETFs in the sample reinvest dividends daily. Investors in these ETFs also face identical tax treatment, and the same dividend distribution policy.

Overall, the results are robust to considering multiple sources of potential differences across ETFs that track the same index. The empirical tests are in line with the model predictions.

3.7 Conclusions

This chapter highlights the role of liquidity in ETF competition. As I show in the theory model, ETFs compete not only on fees, but also on the second dimension, liquidity. High-turnover investors value liquidity and pay higher fees to access it. This gives rise to liquidity clienteles in equilibrium: the high-fee ETF serves high-turnover investors, and the low-fee ETF serves low-turnover investors. Self-perpetuating liquidity cycles (also termed "liquidity begetting liquidity") provide a monopolistic advantage to the high-fee ETF.

The empirical analysis of US multi-ETF indices confirms the key model predictions. High-fee ETFs tend to have higher turnover and narrower spreads, compared to same-index competitors. Also, multiple ETFs per index tend to emerge when combined assets under management are high, and investors have very different trading intensities.

I quantify the value of ETF liquidity. Investors require an average fee discount of 0.51 bps to buy an ETF with 1 bps higher spread, compared to a peer ETF tracking the same index. In terms of volumes, if an ETF doubles its daily volume traded, it can charge 1.15 bps higher fees.

The results imply that liquidity can be a "double-edged sword" in affecting investor welfare. On one hand, highly liquid markets are beneficial to allocating resources efficiently. On the other hand, liquidity gives rise to monopolistic rents paid by liquidity-sensitive investors.

Appendix 3.1. List of same-index ETFs

This appendix provides the list of ETFs that share the same index benchmark with at least one other ETF. The sample contains 60 same-index US-domiciled ETFs based on 24 unique indices. The ETF characteristics are averages for the year 2017, constructed using daily data. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is the annualized percentage ratio of daily dollar volume divided by assets under management (AUM). Tickers with * changed to self-indexing mid-way through the sample period.

Index benchmark	ETF issuer	Ticker	Inception
FTSE Emerging NR USD	Charles Schwab	SCHE	2010/01/14
FTSE Emerging NR USD	Vanguard	VWO	2005/03/04
JPX-Nikkei 400 Net Total Return Index	Deutsche Bank	JPN	2015/06/24
JPX-Nikkei 400 Net Total Return Index	Blackrock	JPXN	2001/10/23
MSCI ACWI Ex USA NR USD	Blackrock	ACWX	2008/03/26
MSCI ACWI Ex USA NR USD	State Street	CWI	2007/01/10
MSCI ACWI Low Carbon Target Index	Blackrock	CRBN	2014/12/08
MSCI ACWI Low Carbon Target Index	State Street	LOWC	2014/11/25
MSCI EAFE 100% Hedged NR USD	Deutsche Bank	DBEF	2011/06/09
MSCI EAFE 100% Hedged NR USD	Blackrock	HEFA	2014/01/31
MSCI Japan 100% Hedged NR USD	Deutsche Bank	DBJP	2011/06/09
MSCI Japan 100% Hedged NR USD	Blackrock	HEWJ	2014/01/31
NASDAQ 100 Equal Weighted TR USD	First Trust	QQEW	2006/04/19
NASDAQ 100 Equal Weighted TR USD	Direxion	QQQE	2012/03/21
Russell 1000 Growth TR USD	Blackrock	IWF	2000/05/22
Russell 1000 Growth TR USD	Vanguard	VONG	2010/09/20
Russell 1000 TR USD	Blackrock	IWB	2000/05/15
Russell 1000 TR USD	State Street	ONEK*	2005/11/08
Russell 1000 TR USD	Vanguard	VONE	2010/09/20
Russell 1000 Value TR USD	Blackrock	IWD	2000/05/22
Russell 1000 Value TR USD	Vanguard	VONV	2010/09/20
Russell 2000 Growth TR USD	Blackrock	IWO	2000/07/24
Russell 2000 Growth TR USD	Vanguard	VTWG	2010/09/20
Russell 2000 TR USD	Blackrock	IWM	2000/05/22
Russell 2000 TR USD	State Street	TWOK*	2013/07/08
Russell 2000 TR USD	Vanguard	VTWO	2010/09/20
Russell 2000 Value TR USD	Blackrock	IWN	2000/07/24
Russell 2000 Value TR USD	Vanguard	VTWV	2010/09/20

Index benchmark	ETF issuer	Ticker	Inception
Russell 3000 TR USD	Blackrock	IWV	2000/05/22
Russell 3000 TR USD	State Street	THRK*	2000/10/04
Russell 3000 TR USD	Vanguard	VTHR	2010/09/20
S&P 500 Growth TR	Blackrock	IVW	2000/05/22
S&P 500 Growth TR	State Street	SPYG	2000/09/25
S&P 500 Growth TR	Vanguard	VOOG	2010/09/07
S&P 500 TR USD	Blackrock	IVV	2000/05/15
S&P 500 TR USD	State Street	SPY	1993/01/22
S&P 500 TR USD	Vanguard	VOO	2010/09/07
S&P 500 Value TR USD	Blackrock	IVE	2000/05/22
S&P 500 Value TR USD	State Street	SPYV	2000/09/25
S&P 500 Value TR USD	Vanguard	VOOV	2010/09/07
S&P Global Infrastructure TR USD	State Street	GII	2007/01/25
S&P Global Infrastructure TR USD	Blackrock	IGF	2007/12/10
S&P MidCap 400 Growth TR	Blackrock	IJK	2000/07/24
S&P MidCap 400 Growth TR	Vanguard	IVOG	2010/09/07
S&P MidCap 400 Growth TR	State Street	MDYG	2005/11/08
S&P MidCap 400 TR	Blackrock	IJH	2000/05/22
S&P MidCap 400 TR	Vanguard	IVOO	2010/09/07
S&P MidCap 400 TR	State Street	MDY	1995/05/04
S&P MidCap 400 Value TR USD	Blackrock	IJJ	2000/07/24
S&P MidCap 400 Value TR USD	Vanguard	IVOV	2010/09/07
S&P MidCap 400 Value TR USD	State Street	MDYV	2005/11/08
S&P SmallCap 600 Growth TR	Blackrock	IJT	2000/07/24
S&P SmallCap 600 Growth TR	State Street	SLYG	2000/09/25
S&P SmallCap 600 Growth TR	Vanguard	VIOG	2010/09/07
S&P SmallCap 600 TR USD	Blackrock	IJR	2000/05/22
S&P SmallCap 600 TR USD	State Street	SLY	2005/11/08
S&P SmallCap 600 TR USD	Vanguard	VIOO	2010/09/07
S&P SmallCap 600 Value TR	Blackrock	IJS	2000/07/24
S&P SmallCap 600 Value TR	State Street	SLYV	2000/09/25
S&P SmallCap 600 Value TR	Vanguard	VIOV	2010/09/07

Ticker	MER, bps	Rel spread, bps	Dol volume, $ mln	AUM, $ mln	Turnover, %
SCHE	13.00	4.11	21.91	3404.86	161.72
VWO	14.22	2.40	456.54	55862.66	205.71
JPN	35.67	17.83	0.09	12.83	184.94
JPXN	48.00	12.55	0.25	88.26	70.19
ACWX	32.09	2.67	24.82	2393.42	261.13
CWI	30.00	4.29	7.77	1295.34	150.90
CRBN	20.00	11.54	0.94	403.20	58.68
LOWC	20.00	17.98	0.11	140.78	19.61
DBEF	35.00	3.39	48.11	7919.25	153.04
HEFA	35.94	4.27	33.49	4075.72	218.31
DBJP	45.00	3.40	19.63	1896.35	260.98
HEWJ	49.00	3.44	32.11	1098.26	734.91
QQEW	60.00	4.82	2.78	506.10	138.18
QQQE	35.00	15.57	1.36	124.43	275.71
IWF	20.00	0.93	181.09	35900.41	127.05
VONG	12.00	3.72	5.26	1162.83	113.69
IWB	15.00	1.00	118.77	18490.70	161.76
ONEK*	10.00	18.39	0.56	137.28	102.48
VONE	12.00	4.60	2.51	717.25	88.37
IWD	20.00	0.92	230.89	36874.94	157.76
VONV	12.00	3.60	4.35	1067.79	102.48
IWO	24.58	1.81	95.36	8078.46	297.28
VTWG	20.00	6.18	0.92	193.21	120.04
IWM	20.00	0.71	3651.68	38912.53	2365.30
TWOK*	10.00	13.04	1.51	204.92	185.24
VTWO	15.00	5.17	8.53	1050.01	204.53
IWN	24.58	1.41	111.87	8766.73	321.59
VTWV	20.00	7.54	1.06	173.43	154.20
IWV	20.00	1.15	28.34	7686.42	92.86
THRK*	10.00	15.21	1.49	416.72	89.92
VTHR	15.00	4.82	1.29	343.30	94.48
IVW	18.00	1.14	91.79	17775.55	130.00
SPYG	12.72	5.86	3.91	769.41	128.47
VOOG	15.00	3.98	6.02	1544.69	98.02

Ticker	MER, bps	Rel spread, bps	Dol volume, $ mln	AUM, $ mln	Turnover, %
IVV	4.00	0.51	834.05	112581.34	186.46
SPY	9.40	0.40	17012.97	240662.22	1780.40
VOO	4.32	0.61	434.58	68728.50	159.18
IVE	18.00	1.16	83.42	13687.09	153.51
SPYV	12.72	8.60	2.10	330.83	160.68
VOOV	15.00	5.75	2.70	726.10	93.46
GII	40.00	38.18	0.85	142.82	149.67
IGF	47.42	5.94	9.75	1617.15	151.72
IJK	25.00	2.25	23.19	6607.17	88.41
IVOG	20.00	3.47	2.59	685.03	95.17
MDYG	15.00	8.06	3.70	665.06	139.74
IJH	7.00	0.80	214.32	39471.90	136.73
IVOO	15.00	2.79	3.49	721.24	121.88
MDY	25.00	0.69	359.18	19412.60	466.23
IJJ	25.00	2.50	21.94	5821.33	94.95
IVOV	20.00	6.23	2.15	651.70	82.98
MDYV	15.00	8.44	3.26	457.36	179.29
IJT	25.00	4.08	19.30	4391.47	110.71
SLYG	15.00	8.12	5.84	1209.82	121.60
VIOG	20.00	5.01	1.25	249.79	125.54
IJR	7.00	1.64	198.20	29655.62	168.49
SLY	15.00	10.44	3.31	765.07	109.02
VIOO	15.00	3.49	4.55	635.85	179.86
IJS	25.00	4.01	21.62	4779.00	114.00
SLYV	15.00	8.33	6.12	977.80	157.62
VIOV	20.00	4.86	1.26	215.83	146.98

Ticker	Structure	Creation unit cost, %	1 Yr avg premium or discount, %	1 Yr NAV tracking error
SCHE	Open-End Investment Company	0.23%	0.11%	0.55
VWO	Open-End Investment Company	0.11%	0.02%	4.46
JPN	Open-End Investment Company	0.20%	0.03%	1.18
JPXN	Open-End Investment Company	0.03%	-0.04%	1.17
ACWX	Open-End Investment Company	0.13%	-0.03%	0.25
CWI	Open-End Investment Company	0.17%	0.03%	0.56
CRBN	Open-End Investment Company	0.09%	0.05%	0.24
LOWC	Open-End Investment Company	0.09%	-0.04%	0.13
DBEF	Open-End Investment Company	0.08%	-0.08%	0.11
HEFA	Open-End Investment Company	0.01%	-0.01%	3.97
DBJP	Open-End Investment Company	0.09%	-0.09%	0.18
HEWJ	Open-End Investment Company	0.01%	0.00%	8.16
QQEW	Open-End Investment Company	0.03%	0.00%	0.05
QQQE	Open-End Investment Company	0.02%	0.02%	0.06
IWF	Open-End Investment Company	0.02%	0.00%	0.04
VONG	Open-End Investment Company	0.02%	0.02%	0.02
IWB	Open-End Investment Company	0.04%	0.01%	0.03
ONEK*	Open-End Investment Company	0.03%	0.01%	0.05
VONE	Open-End Investment Company	0.02%	0.01%	0.03
IWD	Open-End Investment Company	0.03%	0.00%	0.05
VONV	Open-End Investment Company	0.02%	0.01%	0.05
IWO	Open-End Investment Company	0.03%	0.00%	0.03
VTWG	Open-End Investment Company	0.07%	0.02%	0.06
IWM	Open-End Investment Company	0.04%	0.00%	0.04
TWOK*	Open-End Investment Company	0.04%	0.02%	0.20
VTWO	Open-End Investment Company	0.04%	0.02%	0.05
IWN	Open-End Investment Company	0.05%	0.01%	0.05
VTWV	Open-End Investment Company	0.08%	0.04%	0.07
IWV	Open-End Investment Company	0.04%	0.00%	0.03
THRK*	Open-End Investment Company	0.03%	0.02%	0.13
VTHR	Open-End Investment Company	0.07%	0.04%	0.03
IVW	Open-End Investment Company	0.01%	0.01%	0.09
SPYG	Open-End Investment Company	0.02%	0.01%	0.04
VOOG	Open-End Investment Company	0.02%	0.00%	0.02

Ticker	Structure	Creation unit cost, %	1 Yr avg premium or discount, %	1 Yr NAV tracking error
IVV	Open-End Investment Company	0.01%	0.00%	0.01
SPY	Unit Investment Trust (UIT)	0.02%	0.00%	0.06
VOO	Open-End Investment Company	0.01%	0.00%	0.02
IVE	Open-End Investment Company	0.02%	0.01%	0.03
SPYV	Open-End Investment Company	0.03%	0.01%	0.06
VOOV	Open-End Investment Company	0.02%	0.01%	0.03
GII	Open-End Investment Company	0.04%	0.04%	0.26
IGF	Open-End Investment Company	0.04%	0.00%	0.24
IJK	Open-End Investment Company	0.01%	0.00%	0.05
IVOG	Open-End Investment Company	0.02%	0.00%	0.03
MDYG	Open-End Investment Company	0.01%	0.03%	0.04
IJH	Open-End Investment Company	0.01%	0.01%	0.03
IVOO	Open-End Investment Company	0.02%	0.01%	0.03
MDY	Unit Investment Trust (UIT)	0.04%	0.00%	0.05
IJJ	Open-End Investment Company	0.01%	0.01%	0.02
IVOV	Open-End Investment Company	0.02%	0.03%	0.04
MDYV	Open-End Investment Company	0.02%	0.03%	0.05
IJT	Open-End Investment Company	0.01%	0.01%	0.09
SLYG	Open-End Investment Company	0.01%	0.05%	0.07
VIOG	Open-End Investment Company	0.01%	0.04%	0.04
IJR	Open-End Investment Company	0.04%	0.02%	0.09
SLY	Open-End Investment Company	0.05%	0.08%	0.04
VIOO	Open-End Investment Company	0.02%	0.03%	0.04
IJS	Open-End Investment Company	0.02%	0.01%	0.08
SLYV	Open-End Investment Company	0.02%	0.06%	0.08
VIOV	Open-End Investment Company	0.02%	0.04%	0.05

Appendix 3.2. Proofs

Proof of Lemma 2.

The ratio Φ follows immediately from comparing equations (3.19) and (3.21). It is immediate to show that Φ increases in Γ and decreases in Q. I can find a fixed cost threshold $\bar{\Gamma}$ such that $\Phi\left(\bar{\Gamma}\right) = 1$ and the leader deters entry for any $\Gamma > \bar{\Gamma}$, that is,

$$\bar{\Gamma} = \frac{\gamma Q \sigma^2 (\Lambda - 5\xi)^4}{8192 \eta \xi^2}. \tag{A3.1.1}$$

To see that Φ decreases in ξ, I compute the first order condition,

$$\frac{\partial \Phi}{\partial \xi} = \frac{16\left(\Lambda(\Lambda - 5\xi) + \frac{4\sqrt{2}\sqrt{\Gamma}\sqrt{\eta}(\Lambda - 3\xi)}{\sqrt{\gamma}\sqrt{Q}\sigma}\right)}{(\Lambda + 3\xi)^3}. \tag{A3.1.2}$$

I note that the first order condition increases in Γ itself. I evaluate the first-order condition at $\bar{\Gamma}$,

$$\frac{\partial \Phi}{\partial \xi}\left(\bar{\Gamma}\right) = \frac{\Lambda^2 - 25\xi^2}{\xi(\Lambda + 3\xi)^2} < 0, \tag{A3.1.3}$$

for any $\Lambda < 5\xi$. From the monotonicity of the partial derivative, it follows that

$$\frac{\partial \Phi}{\partial \xi} < 0, \forall \Gamma < \bar{\Gamma}. \tag{A3.1.4}$$

Consequently, for any combination of parameters for which the leader would accommodate entry, an increase in investor heteroegeneity relaxes the condition for a two-fund equilibrium.

Proof of Proposition 3.1

The condition $\Phi \leq 1$ follows immediately from Lemma 2, which compares the leader's profit from accommodating or deterring entry.

The equilibrium fees follow from funds' maximization problem as in equation (3.16). Investors' behavior is an immediate consequence of Lemma 1. The derivation of the uniqueness conditions follows the standard global games method in Morris & Shin (2006) and Argenziano (2008). The competitive price schedules set by the authorized participant and the underlying dealer are such that they earn zero expected profits in equilibrium.

It remains to be shown that whenever the leader chooses to accommodate the follower, that is whenever $\Phi \leq 1$, the follower optimally enters the market – i.e., it earns positive expected profit.

From equation (3.19), the follower earns positive expected profit in a two-fund equilibrium if and only if $\Gamma < \underline{\Gamma}$ where

$$\underline{\Gamma} = \frac{\gamma Q \sigma^2 \left(\Lambda - 5\xi\right)^2}{128\eta}. \tag{A3.1.5}$$

It remains to be shown that $\underline{\Gamma} > \bar{\Gamma}$, where $\bar{\Gamma}$ is defined in (A3.1.1), such that the follower always enters the market when the leader accommodates entry. Equivalently, I can show

$$\frac{\underline{\Gamma}}{\bar{\Gamma}} = \frac{64\xi^2}{(\Lambda - 5\xi)^2} > 1. \tag{A3.1.6}$$

The condition in equation (A3.1.6) is equivalent to showing that

$$39\xi^2 - \Lambda^2 + 10\Lambda\xi > 0, \tag{A3.1.7}$$

which is true if the follower can obtain positive market share, i.e., if $5\xi > \Lambda$. This concludes the proof.

Chapter 4

Measuring information and noise in sequential trading

Noise makes financial markets possible, but also makes them imperfect.

Fischer Black (1986) "Noise" [p. 530].

4.1 Introduction

Prices on the consolidated tape are the key input for security analysis. However, observed prices are necessarily imperfect, as they contain noise (Black, 1986). How do we make informed decisions using noisy prices? One solution is to discern the signal-to-noise ratio in the observed price series. This chapter develops a novel methodology to evaluate the proportion of information and noise in sequential phases of trading or sequential markets. Applications of this new method include analyzing price discovery in securities (e.g., futures and forex) that trade in non-overlapping time zones, analyzing the effectiveness of different market phases in aggregating information (e.g., the opening mechanism vs the closing mechanism vs continuous trading vs overnight trading), and analyzing intraday or intrasecond patterns in price discovery.

Information shares, following the approach developed by Hasbrouck (1995), have served as a workhorse for empirical analysis of price discovery. However, this approach requires concurrent trading of one security in multiple parallel markets or multiple related securities. The present study develops the sequential market

analogue of information shares, based on a permanent-temporary variance decomposition. Unlike previous studies of sequential price discovery (e.g., Wang & Yang, 2011), this method explicitly accounts for noise in prices and allows it to vary across the different markets or phases of trading. Therefore, the proposed approach recovers both information shares and noise shares for each trading phase.

The information-to-noise ratio, which can be computed using the proposed method, is especially important for benchmark prices. For example, the official open and close prices are used to value trillions of dollars worth of investments. Understanding the extent to which the benchmark is contaminated by noise is important. For example, the performance of $11.5 trillion worth of mutual fund investments is evaluated using net asset values that are based on closing prices and benchmarked against indices that are also computed from closing prices. At the same time, $3.8 trillion in passive funds prefer to trade at closing price to minimize their tracking errors against the index.[1] Derivative securities, with notional value of $595 trillion, typically use opening or closing prices to settle positions.[2] Academic research in economics and finance relies on opening and closing prices to compute daily returns for asset pricing and corporate finance studies. Furthermore, finance professionals rely on the official open and close in corporate valuations and capital budgeting. Therefore, both academics and practitioners would benefit from understanding the mix of information and noise in prices at specific points in time when benchmarks are set. They would also benefit from understanding how that mix compares to other points in time that could serve as alternative benchmarks. The novel method developed in the present study enables this analysis.

This chapter considers a structural model of price formation, in which the observed price in each trading phase is the sum of the efficient price and a pricing error (noise). The efficient price follows a random walk process, and the pricing errors are stochastic. I allow the variance of efficient price innovations to differ in each of the trading phases or sequential markets. I interpret the relative amount of efficient price variation in a given trading phase as that phase's contribution to price discovery, which is conceptually similar to how Hasbrouck's (1995) information shares attribute price discovery to different parallel markets.

I also allow the variance of pricing errors — the level of noise — to vary in the different phases of trading. This realistic feature of my approach sets it apart from

[1] The data are from Morningstar Direct, as of August 2019.
[2] The data are from the Bank for International Settlements, as of June 2018.

existing models such as Wang & Yang (2011), in which noise is assumed constant. Ample indirect evidence suggests noise has considerable variation through time and across different trading mechanisms, and the proposed approach allows that variation to be captured and incorporated into the estimates. Explicitly accounting for the variation in noise through time and in different trading phases not only removes a bias from the estimated information shares for each phase, but also makes it possible to examine the dynamic properties of noise and information-to-noise ratios. Empirical implementation of the proposed method involves a reinterpretation of Hasbrouck (1995) vector autoregressive system (VAR), such that instead of each equation in the VAR being a relation between returns of related securities, it is a relation between returns from sequential phases of the trading day. Building on this approach, I calculate information shares, noise shares, and information-to-noise ratios for each of the trading phases or sequential markets.

I use Monte Carlo simulations to test the ability of the proposed approach to recover information and noise and correctly attribute it to the different sequential markets. The simulations show that, at least for the simulated structural model of price formation, the approach produces unbiased and fairly precise estimates. I then illustrate the application of the new method by taking intraday data from US stock markets and estimating the intraday patterns in price discovery and noise. The results show that the vast majority of the day's price discovery occurs in the opening auction, which aggregates and impounds all of the overnight information. The open and the period immediately following the open does have elevated noise as well, but not as elevated as the amount of information and therefore the information-to-noise ratio is the highest in the morning. In contrast, the close is associated with a small increase in information, but no discernable increase in the information-to-noise ratio, as noise also increases somewhat at the close.

This chapter contributes a new method to the empirical toolbox that can be used for analyzing price discovery. It is therefore related to the existing literature on empirical measurement of price discovery (see Baillie et al. 2002; Yan and Zivot 2010; Putnins 2013; Narayan and Smyth 2015; and Hasbrouck, 2019 for recent overviews of the literature). But it differs from existing studies by modelling a different setting (sequential rather than parallel markets) and explicitly capturing variation in the level of noise across the sequential markets. The present study is also related to literature on intraday / overnight volatility patterns (Harris, 1986; French & Roll, 1986; Cushing & Madhavan; 2000), and price informativeness

(Campbell & Shiller, 1988; Morck, Yeung, & Yu, 2000; Hasbrouck, 1993; and Brogaard, Nguyen, Putnins, & Wu, 2019).

This chapter proceeds as follows. Section 4.2 reviews relevant literature, and Section 4.3 develops a model for separating information and noise in sequential markets. Section 4.4 provides Monte Carlo simulation evidence, and Section 4.5 illustrates the model application by recovering intraday patterns in information and noise. Section 4.6 concludes the chapter and provides directions for future research.

4.2 Literature review

Lehmann (2002) emphasizes two key aspects of price discovery: (i) efficient and (ii) timely incorporation of new information into prices. The "efficiency" dimension reflects relative avoidance of noise in the price series, while the "timeliness" dimension reflects which price series is the first to incorporate the new information (Putnins, 2013). This chapter speaks to the "efficiency" dimension by developing a novel method for estimating price discovery in sequential trading phases. Unlike existing studies, this method explicitly accounts for noise.

Standard price discovery settings consider the same security trading simultaneously in different markets (e.g., a home exchange and a cross-listing exchange) or through different instruments (e.g., options vs futures). The empirical approaches date back to Hasbrouck (1995) Information Shares (IS) or Gonzalo-Granger (1995) Component Shares (CS). Both of these approaches estimate a combination of "efficiency" and "timeliness" of incorporating new information. Putnins (2013) combines IS and CS into Information Leadership Share (ILS) that measures price discovery purely based on the "timeliness" dimension. ILS identifies the market (or trading instrument) that "moves first" in response to new information about the fundamental value of an asset. Crucially, ILS disregards the relative amount of information incorporated in each market or the relative amount of noise that accompanies any price movements in response to that information.

The price discovery method developed in this study considers the same security traded sequentially in non-overlapping markets or trading phases. This setting is scarcely covered in existing literature. One paper that estimates price discovery in consecutive non-overlapping trading phases is Wang & Yang (2011). But the

method proposed in this chapter relaxes the assumptions in Wang & Yang (2011) and allows for unequal noise variances in different trading phases. It also explicitly estimates the noise share of each phase. This is an important and realistic feature, as different investor clienteles that trade in e.g., continuous session vs the closing auction are likely to introduce different levels of noise into prices.

4.2.1 Price discovery in one security and many markets

Hasbrouck (1995) proposes to measure contributions to price discovery from different markets that trade the same security at the same time. His method relies on cointegrated time series modelled via VECM (vector error correction models). The information share (IS) of a price series is calculated as the proportion of the variance in the common efficient price that is explained by innovations in that price series. The method attributes information shares correctly, if innovations across markets are uncorrelated. If, however, innovations are correlated, one can estimate the upper and lower bounds on IS of each price series.

Gonzalo & Granger (1995) take a somewhat different approach to the permanent-temporary decomposition of returns. In their approach, the common efficient price is a linear combination of all efficient prices from different price discovery venues. The Component Share (CS) of a price series is the normalized weight in the linear combination of prices that form the common efficient price. Baillie et al. (2002) note that IS and CS produce consistent results, if variance of the price series is similar across markets. Putnins (2013) demonstrates through Monte Carlo simulations that the reason for this is noise in price series. Because price series have different levels of noise, and both IS and CS assign different weights to "relative avoidance of noise", they show different outcomes with respect to price discovery shares.

Subsequent methodology studies (including de Jong, 2002; Baillie et al, 2002; Yan & Zivot, 2010; Putnins, 2013) help to interpret different price discovery metrics. Baillie et. al (2002) show how IS and CS measures are related through the VECM error correction vector. They also show that IS and CS provide similar results, if VECM residuals are uncorrelated. Yan & Zivot (2010) analytically combine IS and CS into a single measure that cancels out noise in the time series. Putnins (2013) complements Yan & Zivot (2010) with Monte Carlo simulations of price series that differ across two dimensions: (i) the amount of noise and (ii) the speed

of adjustment to new information. The resulting Yan-Zivot-Putnins ILS measure captures the relative speed of impounding new information in two series with different amounts of noise.

Multiple empirical studies investigate contributions to price discovery of different stock exchanges trading cross-listed stocks (e.g., Hasbrouck, 1995; Harris et al., 1993; Pascual et al., 2006; Frijns et al., 2010; Chen et al., 2013), stock options compared to stocks (Chakravarty et al., 2004; Muravyev et al., 2013), futures compared to stocks (Booth et al., 1999; Covrig et al., 2004; Cabrera et al., 2009), credit default swaps compared to bonds and stocks (Forte & Pena, 2009), screen-based compared to open outcry trading (Ates & Wang, 2005), different contract maturities (Fricke & Menkhoff, 2011), quotes compared to trade prices (Cao et al., 2009), and order flow from different types of traders (Fong & Zurbruegg, 2003; Kurov & Lasser, 2004; Anand & Subrahmanyam, 2008; Anand et al., 2011).

4.2.2 Price informativeness in one security and one market

Price informativeness literature studies how stock prices change in response to information. Decomposing stock return variances makes it possible to attribute information to different sources. For example, two canonical approaches decompose return variances into (i) cash flow- and discount rate- related components (Campbell & Shiller, 1988) and (ii) market-wide and firm-specific components (Morck, Yeung, & Yu, 2000). The limitation of these approaches lies in using observed prices without accounting for noise that arises from market imperfections. But accounting for noise is important, as recent studies (e.g., Asparouhova et al., 2013) show that noise in asset prices can significantly bias inference in cross-sectional return regressions.

Brogaard, Nguyen, Putnins, & Wu (2019) explicitly account for noise in return decomposition method that builds on Campbell (1991) and (Morck, Yeung, & Yu, 2000). Brogaard et al. (2019) distinguish between (i) noise in stock prices, (ii) market-wide information, (iii) firm-specific private information, (iv) firm-specific public information, and (v) discount rate. Their method builds on the Hasbrouck (1991*a*) VAR model to decompose the returns into permanent ("information") and temporary ("noise") components, which is similar to the empirical method used in this chapter.

In broad terms, noise is any deviation between the observed price and efficient price of an asset (Hasbrouck, 1993). By definition, this deviation reverts to zero in the long term, as the observed price converges to the efficient price. On short horizons (i.e., daily or weekly), noise can arise from microstructure frictions such as the bid-ask spread or discrete price grid (Ball & Chordia, 2001), temporary price pressures resulting from nonsynchronous trading (Hendershott & Menkveld, 2014), and generally imperfect liquidity supply in response to order imbalances (Grossman & Miller, 1988; Admati & Pfleiderer, 1991; Bertsimas & Lo, 1998). On short as well as long horizons (i.e., also beyond weekly), noise can arise from liquidity traders (Black, 1986) or irrational traders, in the presence of limits to arbitrage (Barberis & Thaler, 2003).

The focus in this study is on noise generated by microstructure-related frictions, rather than by irrational trading or limits to arbitrage. The information-noise decomposition relies on Hasbrouck's (1993) VAR approach. The underlying assumption is that the long term efficient price response to one unit increase in noise is zero. In simple terms, noise does not move prices in the long run, while information does. However, noise can have economically large effects on prices: in US stocks, daily observed prices deviate by an average of 0.49% from efficient prices (Hendershott & Menkveld, 2014) and such deviations persist for 1.8 days. Removing noise allows for clean inference with respect to information content of prices. For example, Hagstromer & Menkveld (2017) combine Hasbrouck (1991*b*) information-noise decomposition with Hasbrouck (1995) information shares to examine directional information linkages between markets.

Consistent with the presence of microstructure-related noise, prior studies find evidence of return reversals on weekly (Lehmann, 1990) and monthly (Jegadeesh, 1990) horizons. Nagel (2012) shows that short-term return reversals proxy for liquidity provision profits. Bogousslavsky (2016) models infrequent rebalancing events that can give rise to observed patterns in return autocorrelations. Chinco & Fos (2019) show that ETF rebalancing cascades lead to large statistically unpredictable demand shocks, which make stock prices noisier. The present study offers an empirical methodology to examine the effects of rebalancing events (for example), on relative noisiness of prices.

4.2.3 Intraday / overnight return patterns

Literature on intraday and overnight return patterns provides evidence of overnight reversals (Branch & Ma, 2012), overnight (but not intraday) equity risk premia (Cliff, Cooper, & Gulen, 2008; Hendershott, Livdan, & Rosch, 2019; Lou, Polk, & Skouras, 2019), and persistent hourly return patterns (Heston, Korajczhyk, & Sadka, 2010). Branch & Ma (2012) find a significant negative autocorrelation between overnight and intraday returns. Cliff, Cooper & Gulen (2008) find that equity risk premium is only present in overnight returns, while intraday returns are close to zero or negative. Lou, Polk, & Skouras (2019) decompose daily returns into overnight and intraday components and find that almost all abnormal returns on momentum and reversals strategies happen overnight. They attribute these patterns to investor clienteles. Hendershott, Livdan, & Rosch (2019) show that CAPM holds for overnight returns, but not for open-to-close returns.

Volatility also differs across trading phases: realized stock volatility tends to follow a U-shaped pattern during the trading day (McInish & Wood, 1990; Lockwood & Linn, 1990; Cai, Hudson & Keasey, 2004). Market volatility is higher during the trading phase compared to overnight (French & Roll, 1986). The choice of reference price matters: for example, Amihud & Mendelson (1987) find that open-to-open volatility is higher than close-to-close. Cushing and Madhavan (2000) find that closing prices are affected by transitory imbalances, and closing returns tend to reverse overnight. These studies call for better understanding of the noise content of opening and closing prices.

4.2.4 Price discovery in sequential trading: challenges and methodological choices

Starting from Blume & Stambaugh (1983), multiple studies show than noise biases returns and therefore affects inference about price discovery. The challenge then is to estimate price discovery in sequential trading phases in a manner that accounts for noise. Existing literature offers several approaches to sequential price discovery: (i) calculating weighted price contributions (without accounting for noise) as in Barclay & Warner (1993), (ii) calculating the probability of Informed trading

(PIN) following Easley, Kiefer, & O'Hara (1996), (iii) accounting for signal-to-noise ratio in trade prices using "unbiasedness regression" of Biais, Hillion, & Spatt (1999).

Weighted price contributions (WPC) assume that greater price movements indicate more information. Barclay & Warner (1993) use WPC to evaluate which trades move prices around tender offers, finding that it is largely medium-size trades. In discussing their results, Barclay & Warner (1993) acknowledge that this methodology does not account for temporary price change components, such as a bid-ask bounce, temporary price pressures and correlated trade sequences. Following the same methodology, Bacidore & Lipson (2001) study the effects of opening and closing procedures of NYSE and NASDAQ on execution costs and price discovery. In their sequential trading analysis, the weighted price contribution of the last 15 minutes of the trading day is the ratio of price change during these 15 min to the price change during the 24-hour close-to-close period. Others using a similar approach include Cao, Ghysels & Hatheway (2000) and Huang (2000).

Barclay & Hendershott (2003) employ the structural model of Easley, Kiefer, & O'Hara (1996) to establish where the informed traders prefer to trade. They find that the amount of informed trading is significantly higher during the preopen period, compared to post-close. To answer the question how efficient prices are in these two periods, they estimate the "unbiasedness regressions" of Biais, Hillion, & Spatt (1999). For example, to assess how informative the indicative preopen price is, they regress close-to-close return on close-to-indicative-price return. Low beta coefficient on close-to-indicative-price return indicates low signal to noise ratio in the indicative price. Similarly, if root mean squared errors (RMSE) are similar to the variance of close-to-close returns, then the indicative price does not bring much extra information. Crucially, the approach in Biais, Hillion, & Spatt (1999) assumes that closing prices do not contain noise. The model in this chapter relaxes that assumption and allows for dynamic structure of price innovations in each trading period.

4.3 Model for separating information and noise in sequential markets

4.3.1 Structural model of price formation

Consider a 24-hour period t that consists of n phases. Observed prices and returns are illustrated in Figure 4.1. The observed log-price in any given phase i (where $1 \leqslant i \leqslant n$) consists of the efficient price and a pricing error (the noise component):[3]

$$\begin{aligned} p_{i,t} &= m_{i,t} + s_{i,t} \\ &= m_{i-1,t} + w_{i,t} + s_{i,t} \end{aligned} \tag{4.1}$$

where $p_{i,t}$ is the observed price in phase i of the 24-hour period t, $m_{i,t}$ is the efficient price, $s_{i,t} \sim N(0, \sigma^2_{s_i})$ is the pricing error, $m_{i-1,t}$ is the efficient price in phase $i-1$ on day t,[4] and $w_{i,t} \sim N(0, \sigma^2_{w_i})$ is the efficient price innovation in phase i. In this structural model, each phase i can have a different variance of efficient price innovations ($\sigma^2_{w_i}$). For simplicity I assume the pricing errors in different phases are independent.

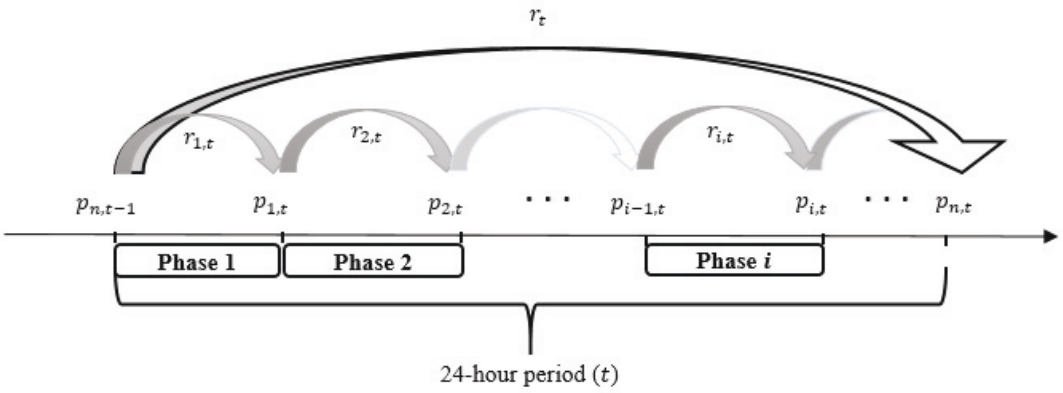


Figure 4.1: Observed prices and returns in sequential phases

This figure illustrates the notation used in the structural model. A 24-hour period t consists of n phases. $r_{i,t} = p_{i,t} - p_{i-1,t}$ is the observed log-return in phase i of 24-hour period t, $p_{i,t}$ is the observed log-price. The 24-hour return is $r_t = \sum_{i=1}^{n} r_{i,t}$.

[3] Going forward, all prices are in natural logarithms. Return in each phase is the difference between two log-prices.

[4] When $i = 1$, $i - 1$ refers to phase n of the previous day $t - 1$.

The observed log-return in phase i is:

$$\begin{aligned} r_{i,t} &= p_{i,t} - p_{i-1,t} \\ &= w_{i,t} + s_{i,t} - s_{i-1,t} \end{aligned} \tag{4.2}$$

Given this simple structural model of price formation, the challenge now is to derive an approach to empirically estimate $\sigma^2_{w_i} = Var(w_{i,t})$, the information impounded into prices by phase i, and $\sigma^2_{s_i} = Var(s_{i,t})$, the noise in phase i's prices. From estimates of $Var(w_{i,t})$ and $Var(s_{i,t})$, I can construct information shares, noise shares, and information-to-noise ratios. Therefore, I now derive expressions for $Var(w_{i,t})$ and $Var(s_{i,t})$ in terms of quantities that can be empirically observed or estimated, including the variance of observed returns, $Var(r_{i,t})$, and temporary/permanent return components that can be estimated from empirical temporary/permanent return decomposition.

The derivation starts by recognizing that the return $r_{i,t}$ consists of permanent and temporary components. The permanent component contains not only new information impounded by phase i, $w_{i,t}$, but also the correction of the previous phase's pricing error, $s_{i,t}$. The temporary component is just the pricing error in phase i. From (4.2):

$$r_{i,t} = \underbrace{w_{i,t} - s_{i-1,t}}_{\text{Permanent return}} + \underbrace{s_{i,t}}_{\text{Temporary return}} \tag{4.3}$$

Further, the permanent return can be broken down into an expected component and an unexpected (shock/surprise) component. From (4.3):

$$\begin{aligned} r_{i,t} = & \underbrace{w_{i,t} - [s_{i-1,t} - \mathbb{E}[s_{i-1,t}]]}_{\text{Unexpected permanent return } UEPR_{i,t}} \\ & - \underbrace{\mathbb{E}[s_{i-1,t}]}_{\text{Expected permanent return } EPR_{i,t}} \\ & + \underbrace{s_{i,t}}_{\text{Temporary return (unexpected)}} \end{aligned} \tag{4.4}$$

The expected permanent return component, $EPR_{i,t}$, is the reversal of the previous phase's pricing error that can be inferred from having observed all information up to time t. Intuitively, a large positive return in the previous trading phase implies that the pricing error in the previous phase was likely positive, leading to an expectation of a negative return in the present phase, as the previous pricing

error is corrected. Therefore, $s_{i-1,t}$ is the only term that leads to some ability to predict returns (that carries over from the past into the present return): $EPR_{i,t} = \mathbb{E}[r_{i,t}] = \mathbb{E}[s_{i-1,t}]$. I estimate $EPR_{i,t}$ using one step ahead forecasts from the structural VAR (described in more detail in the following subsection).

Applying the variance operator to (4.4):[5]

$$Var(r_{i,t}) = Var(UEPR_{i,t}) + Var(EPR_{i,t}) + Var(s_{i,t}) \tag{4.5}$$

In (4.5), I know variances of $r_{i,t}$ and $EPR_{i,t}$, but do not know $UEPR_{i,t}$. The unexpected permanent return component, $UEPR_{i,t}$, contains the efficient price innovation $w_{i,t}$ and the unexpected part of the noise innovation from the previous phase.

The return $r_{i,t}$ I observe directly. $EPR_{i,t}$ I estimate using one step ahead forecasts from the structural VAR (described in more detail in the following subsection). The remaining term that I need before being able to express $Var(s_{i,t})$ is $Var(UEPR_{i,t})$. I discuss $UEPR_{i,t}$ below.

Let $\varepsilon_{i,t}$ denote the unexpected return (i.e., shock / innovation in total return in phase i). It consists of the unexpected *permanent* return $UEPR_{i,t}$ and the unexpected *temporary* return $s_{i,t}$:

$$\varepsilon_{i,t} = UEPR_{i,t} + s_{i,t} \tag{4.6}$$

Let θ_i denote how much of the *total* return innovation gets translated into the *permanent* return innovation,[6] such that:

$$UEPR_{i,t} = \theta_i \varepsilon_{i,t} + \epsilon_{i,t} \tag{4.7}$$

where $\epsilon_{i,t}$ are mean zero error terms uncorrelated with $\varepsilon_{i,t}$, like in regression. Substituting in the expression for $\varepsilon_{i,t}$ from (4.6) into (4.7):

$$UEPR_{i,t} = \theta_i(UEPR_{i,t} + s_{i,t}) + \epsilon_{i,t} \tag{4.8}$$

[5]In Equation (4.5), because $s_{i,t}$ and $w_{i,t}$ are *unexpected*, their covariance with all other return components is zero. The two remaining covariance terms cancel out under assumption that $Cov(s_{i-1,t}, EPR_{i,t}) = Cov(\mathbb{E}[s_{i-1,t}], EPR_{i,t})$. Intuitively, because I do not know the actual noise noise term, my best estimate is that the correlation between the prediction ($\mathbb{E}[s_{i-1,t}]$) and the actual noise ($s_{i,t}$) is 1.

[6]The reason for defining this parameter will become apparent when considering the empirical estimation.

Using (4.8) to express the error term $\epsilon_{i,t}$:

$$\epsilon_{i,t} = UEPR_{i,t}(1-\theta_i) - \theta_i s_{i,t} \tag{4.9}$$

Applying the variance operator to (4.7) and substituting in the expression for $\epsilon_{i,t}$ from (4.9), the variance of $UEPR_{i,t}$ is:

$$Var(UEPR_{i,t}) = \theta_i^2 Var(\varepsilon_{i,t}) + Var\big(UEPR_{i,t}\,(1-\theta_i) - \theta_i s_{i,t}\big) \tag{4.10}$$

From (4.10), the variance of *unexpected permanent* return can be expressed using variances of the *total unexpected* return ($\varepsilon_{i,t}$) and the *temporary unexpected* return ($s_{i,t}$):

$$Var\left(UEPR_{i,t}\right) = \frac{\theta_i^2}{\left[1-(1-\theta_i)^2\right]} Var\left(\varepsilon_{i,t}\right) + \frac{\theta_i^2}{\left[1-(1-\theta_i)^2\right]} Var(s_{i,t}) \tag{4.11}$$

Substituting $Var\left(UEPR_{i,t}\right)$ into (4.5), I have:

$$\begin{aligned} Var\left(r_{i,t}\right) = {} & \frac{\theta_i^2}{\left[1-(1-\theta_i)^2\right]} Var\left(\varepsilon_{i,t}\right) + \frac{\theta_i^2}{\left[1-(1-\theta_i)^2\right]} Var(s_{i,t}) \\ & + Var(EPR_{i,t}) + Var(s_{i,t}) \end{aligned} \tag{4.12}$$

Rearranging (4.12), the variance of temporary returns, or pricing errors (i.e., noise), in phase i is given by:

$$Var\left(s_{i,t}\right) = \left(1-\frac{\theta_i}{2}\right) Var\left(r_{i,t}\right) - \frac{\theta_i}{2} Var\left(\varepsilon_{i,t}\right) - \left(1-\frac{\theta_i}{2}\right) Var(EPR_{i,t}) \tag{4.13}$$

The terms on the right-hand side of (4.13) can all be empirically estimated using the permanent/temporary return decomposition in the spirit of Hasbrouck (1993). Using a vector autoregression (VAR) specified in the following subsection, one can obtain innovation variances (i.e., $Var(\varepsilon_{i,t})$), permanent price impacts of return innovations (i.e., θ_i), and variances of the expected permanent return (i.e., $Var(EPR_{i,t})$).

With an estimate of the noise variance, it is straightforward to express the information variance from (4.2) as:

$$Var\left(w_{i,t}\right) = Var\left(r_{i,t}\right) - Var\left(s_{i-1,t}\right) - Var(s_{i,t}) \tag{4.14}$$

Substituting in (4.13) for noise variance, (4.14) becomes:

$$\begin{aligned} Var\left(w_{i,t}\right) = Var\left(r_{i,t}\right) &- \left(1 - \frac{\theta_{i-1}}{2}\right) Var\left(r_{i-1,t}\right) - \frac{\theta_{i-1}}{2} Var\left(\varepsilon_{i-1,t}\right) \\ &- \left(1 - \frac{\theta_{i-1}}{2}\right) Var(EPR_{i-1,t}) - \left(1 - \frac{\theta_i}{2}\right) Var\left(r_{i,t}\right) \\ &- \frac{\theta_i}{2} Var\left(\varepsilon_{i,t}\right) - \left(1 - \frac{\theta_i}{2}\right) Var(EPR_{i,t}) \end{aligned} \tag{4.15}$$

Equation (4.15) provides an expression for the variance of efficient price changes in terms of quantities that can all be empirically estimated.

The final step is to take the estimates of information variance ($Var(w_{i,t})$) and noise variance ($Var(s_{i,t})$) in (4.15) and (4.13), and construct information shares, noise shares, and information-to-noise ratios. The information share of phase i, being the proportion of total efficient price variation impounded into prices during phase i, is given by:

$$IS_i = \frac{Var\left(w_{i,t}\right)}{\sum_{i=1}^{n} Var\left(w_{i,t}\right)} \tag{4.16}$$

with $Var(w_{i,t})$ given by (4.15).

The noise share of phase i, being the noise in phase i, normalized by the sum of noise in all phases, is given by:

$$NS_i = \frac{Var\left(s_{i,t}\right)}{\sum_{i=1}^{n} Var\left(s_{i,t}\right)} \tag{4.17}$$

with $Var(s_{i,t})$ given by (4.13).

For each trading phase, I also compute an information-to-noise ratio (IN):

$$IN_i = \frac{Var\left(w_{i,t}\right)}{Var\left(w_{i,t}\right) + Var\left(s_{i,t}\right)} \tag{4.18}$$

4.3.2 Empirical estimation of the model

The empirical estimates needed for IS_i, NS_i, and IN_i above are the variance of observed returns $Var(r_{i,t})$ and three parameters estimated from a permanent-temporary decomposition of returns: $Var(\varepsilon_{i,t})$, θ_i, and $Var(EPR_{i,t})$. In this subsection I show the empirical approach to estimating these quantities. For simplicity, I demonstrate the specific case of three phases, $n = 3$, although the approach easily generalizes to any number of phases.

Similar to Hasbrouck (1993), I model each phase's time-series of returns as a structural vector autoregression (VAR):

$$\begin{aligned} r_{1,t} &= \alpha_0 + \sum_{l=1}^{5} \alpha_{1_l} r_{1,t-l} + \sum_{l=1}^{5} \alpha_{2_l} r_{2,t-l} + \sum_{l=1}^{5} \alpha_{3_l} r_{3,t-l} + \varepsilon_{1,t} \\ r_{2,t} &= \beta_0 + \sum_{l=0}^{5} \beta_{1_l} r_{1,t-l} + \sum_{l=1}^{5} \beta_{2_l} r_{2,t-l} + \sum_{l=1}^{5} \beta_{3_l} r_{3,t-l} + \varepsilon_{2,t} \\ r_{3,t} &= \gamma_0 + \sum_{l=0}^{5} \gamma_{1_l} r_{1,t-l} + \sum_{l=0}^{5} \gamma_{2_l} r_{2,t-l} + \sum_{l=1}^{5} \gamma_{3_l} r_{3,t-l} + \varepsilon_{3,t} \end{aligned} \tag{4.19}$$

The contemporaneous causality assumptions in this VAR are strictly satisfied due to the *sequential* nature of the three phases. That is, the returns from previous phases of the trading day can affect returns of subsequent phases of that same day, but not vice versa. For example, phase 2 returns on a given day incorporate information from phase 1, but not from phase 3. For illustration, I allow for five lags of returns to capture the time series dynamics.

The model above is a structural VAR, because it contains contemporaneous terms in the equations for $r_{2,t}$ and $r_{3,t}$. For estimation purposes, I transform the structural VAR into its reduced form:

$$\begin{aligned} r_{1,t} &= \delta_0 + \sum_{l=1}^{5} \delta_{1_l} r_{1,t-l} + \sum_{l=1}^{5} \delta_{2_l} r_{2,t-l} + \sum_{l=1}^{5} \delta_{3_l} r_{3,t-l} + e_{1,t} \\ r_{2,t} &= \kappa_0 + \sum_{l=1}^{5} \kappa_{1_l} r_{1,t-l} + \sum_{l=1}^{5} \kappa_{2_l} r_{2,t-l} + \sum_{l=1}^{5} \kappa_{3_l} r_{3,t-l} + e_{2,t} \\ r_{3,t} &= \rho_0 + \sum_{l=1}^{5} \rho_{1_l} r_{1,t-l} + \sum_{l=1}^{5} \rho_{2_l} r_{2,t-l} + \sum_{l=1}^{5} \rho_{3_l} r_{3,t-l} + e_{3,t} \end{aligned} \tag{4.20}$$

The difference between the structural and the reduced form VAR representations is that the former explicitly models contemporaneous causality between variables, hence the structural innovations $(\varepsilon_{1,t}, \varepsilon_{2,t}, \varepsilon_{3,t})$ are contemporaneously uncorrelated. In the reduced form, however, all right-hand side variables are lagged, and therefore the contemporaneous relations between variables (e.g., that phase 2 returns on a given day are related to phase 1 returns that day) are captured through contemporaneous correlations of the reduced form error terms.

The correspondence between the reduced form residuals $e_{1,t}, e_{2,t}, e_{3,t}$ and the structural innovations $\varepsilon_{1,t}, \varepsilon_{2,t}, \varepsilon_{3,t}$ (which can be found by substituting out the contemporaneous terms in the structural VAR to arrive at the reduced form VAR) is as follows:

$$\begin{aligned} e_{1,t} &= \varepsilon_{1,t} \\ e_{2,t} &= \varepsilon_{2,t} + b_1 \varepsilon_{1,t} \\ e_{3,t} &= \varepsilon_{3,t} + b_2 \varepsilon_{1,t} + b_3 \varepsilon_{2,t} \end{aligned} \tag{4.21}$$

The parameters b_1, b_2, and b_3 can be estimated by regressing $e_{2,t}$ on $e_{1,t}$, and regressing $e_{3,t}$ on $e_{1,t}$ and $e_{2,t}$:

$$\begin{aligned} e_{2,t} &= c_1 e_{1,t} + \varepsilon_{2,t} \\ e_{3,t} &= c_2 e_{1,t} + c_3 e_{2,t} + \varepsilon_{3,t} \end{aligned} \tag{4.22}$$

to obtain coefficients c_1, c_2, and c_3 and then recovering b_1, b_2, and b_3 from those estimated coefficients:

$$\begin{aligned} b_1 &= c_1 \\ b_2 &= c_2 + c_1 c_3 \\ b_3 &= c_3 \end{aligned} \tag{4.23}$$

By applying the variance operator to (4.21), recognizing that the structural errors are uncorrelated by construction, and then rearranging, I recover the structural model variances from the reduced form variances:

$$\begin{aligned} Var\left(\varepsilon_{1,t}\right) &= Var(e_{1,t}) \\ Var\left(\varepsilon_{2,t}\right) &= Var\left(e_{2,t}\right) - b_1^2 Var(e_{1,t}) \\ Var\left(\varepsilon_{3,t}\right) &= Var\left(e_{3,t}\right) + (b_1^2 b_3^2 - b_2^2) Var\left(e_{1,t}\right) - b_3^2 Var\left(e_{2,t}\right) \end{aligned} \tag{4.24}$$

In addition to these variances, another parameter that must be estimated to construct the information and noise shares is θ_i. Recall from (4.7) that θ_i is the long-term (i.e., "permanent") effect on prices from a unit shock to $\varepsilon_{i,t}$. Intuitively, θ_i measures how much of the price innovation, $\varepsilon_{i,t}$, is permanent. This parameter comes naturally from impulse response functions, which trace the dynamic response of prices to a shock. The point at which prices stabilize following a shock provides the estimated permanent effect of the shock.[7]

I estimate θ_i for each trading phase as the cumulative response of returns to a unit shock in the corresponding $\varepsilon_{i,t}$. The b_1, b_2, b_3 parameters allow me to translate any structural shock into an equivalent reduced form shock, so that I can use the reduced form VAR to arrive at structural impulse response functions. For example, a structural shock $(\varepsilon_{1,t}, \varepsilon_{2,t}, \varepsilon_{3,t}) = (1, 0, 0)$ corresponds to the equivalent reduced form shock of $(e_{1,t}, e_{2,t}, e_{3,t}) = (1, b_1, b_2 + b_3)$. Because the sum of returns in the three phases gives the daily return on day t, the permanent effect of a shock to $\varepsilon_{i,t}$ is given by the sum of the cumulative responses of all three phase returns.

The remaining input to compute variance of information in (4.15) and variance of noise in (4.13) is $Var(EPR_{i,t})$. $EPR_{i,t}$ is just the expected value of return in each phase, conditional on returns in prior phases. Hence, I estimate the time series of $EPR_{i,t}$ as fitted values from (4.19), and then compute their variances.

As a final step, I compute variance of information in (4.15) and variance of noise in (4.13) from observed returns $r_{i,t}$ and three parameters estimated from the empirical VAR model: θ_i, $Var(EPR_{i,t})$, $Var(\varepsilon_{i,t})$.

4.4 Monte Carlo simulation evidence

To verify that the proposed empirical methodology can reliably recover the structural model parameters, I simulate log-prices using the Monte Carlo procedure. I then compare the estimated IS, NS, IN ratios to their theoretical true values. I perform this exercise for the model with $n = 3$ phases in the 24-hour period, and for the model with $n = 2$ phases, and show that the results are consistent across the two versions.[8]

[7] An equivalent way of inferring the permanent effects of shocks is by inverting the VAR to an infinite order VMA and using the VMA coefficients to infer the permanent effect of a shock.

[8] In theory, the model recovers IS, NS, IN ratios for any number of phases $n \geq 2$. However, greater values of n result in greater number of VAR parameters to be estimated, and therefore

4.4.1 Model estimation with n=3 phases

The simulations rely on the structural model described in Section 4.3.1, with $n = 3$ phases in the 24-hour period. Observed log-return in phase 1 ($r_{1,t}$) reflects information in phase 1 ($w_{1,t}$), and noise terms ($s_{3,t-1}$ and $s_{1,t}$) that correspond to observed log-prices ($p_{3,t-1}$ and $p_{1,t}$), from which $r_{1,t}$ is calculated. Observed returns in the three phases are: $r_{1,t} = w_{1,t} + s_{1,t} - s_{3,t-1}$, $r_{2,t} = w_{2,t} + s_{2,t} - s_{1,t}$, $r_{3,t} = w_{3,t} + s_{3,t} - s_{2,t}$. I fix the information and noise variances for phases 1 and 2, and vary them for phase 3. Specifically, $w_{1,t} \sim N(0, 0.01)$, $w_{2,t} \sim N(0, 0.04)$, $s_{1,t} \sim N(0, 0.01)$, $s_{2,t} \sim N(0, 0.04)$, where $r_{1,t}$, $r_{2,t}$, $r_{3,t}$ are observed log-returns in phase 1, 2 and 3, $w_{1,t}$, $w_{2,t}$, $w_{3,t}$ are efficient log-returns, $s_{1,t}$, $s_{2,t}$, $s_{3,t}$ are pricing errors. This setting allows me to compare the estimated ratios IS_3, NS_3, IN_3 to the true ratios IS_3, NS_3, IN_3 computed from the Monte Carlo parameters.

Table 4.1: Estimates of information shares from the model with $n = 3$ phases

This table reports the mean estimates of information shares in phase 3 (IS_3) from the VAR model with three phases ($n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phases 1 and 2 are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. Information share in phase 3 is $IS_3 = \frac{Var(w_{3,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(w_{3,t})}$. The detailed description of the simulated model is in Section 4.3.1. The description of Monte Carlo procedure is in Section 4.4.

Noise variance, $\sigma^2_{s_3}$	Information variance, $\sigma^2_{w_3}$				
	0.10^2	0.20^2	0.30^2	0.40^2	0.50^2
0.20^2	0.1781 (0.1667)	0.4400 (0.4444)	0.6319 (0.6429)	0.7483 (0.7619)	0.8187 (0.8333)
0.25^2	0.1804 (0.1667)	0.4388 (0.4444)	0.6301 (0.6429)	0.7471 (0.7619)	0.8175 (0.8333)
0.30^2	0.1830 (0.1667)	0.4376 (0.4444)	0.6282 (0.6429)	0.7456 (0.6429)	0.8166 (0.8333)
0.35^2	0.1862 (0.1667)	0.4365 (0.4444)	0.6263 (0.6429)	0.7439 (0.7619)	0.8156 (0.8333)
0.40^2	0.1900 (0.1667)	0.4354 (0.4444)	0.6242 (0.6429)	0.7420 (0.7619)	0.8144 (0.8333)

I zoom in on phase 3 information shares in Table 4.1. This table reports IS_3 for different values of information and noise variances in this phase. I vary

lower degrees of freedom. This may result in lower precision of IS, NS, IN, when sample size becomes too small relative to n.

phase 3 information variance from 0.01 to 0.25, and phase 3 noise variance from 0.04 to 0.16. The true values of information shares are reported in parentheses below the estimated values. I follow Equation (4.16) to compute $IS_3 = \frac{Var(w_{3,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(w_{3,t})}$.

Table 4.2: Estimates of noise shares from the model with $n = 3$ phases

This table reports the mean estimates of noise shares in phase 3 (NS_3) from the VAR model with three phases ($n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true noise shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phases 1 and 2 are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. Noise share in phase 3 is $NS_3 = \frac{Var(s_{3,t})}{Var(s_{1,t})+Var(s_{2,t})+Var(s_{3,t})}$. The detailed description of the simulated model is in Section 4.3.1. The description of Monte Carlo procedure is in Section 4.4.

Noise variance, $\sigma^2_{s_3}$	Information variance, $\sigma^2_{w_3}$				
	0.10^2	0.20^2	0.30^2	0.40^2	0.50^2
0.20^2	0.4489 (0.4444)	0.4462 (0.4444)	0.4388 (0.4444)	0.4231 (0.4444)	0.4022 (0.4444)
0.25^2	0.5613 (0.5556)	0.5604 (0.5556)	0.5567 (0.5556)	0.5473 (0.5556)	0.5305 (0.5556)
0.30^2	0.6487 (0.6429)	0.6488 (0.6429)	0.6469 (0.6429)	0.6412 (0.6429)	0.6307 (0.6429)
0.35^2	0.7156 (0.7101)	0.7162 (0.7101)	0.7152 (0.7101)	0.7113 (0.7101)	0.7043 (0.7101)
0.40^2	0.7668 (0.7619)	0.7675 (0.7619)	0.7670 (0.7619)	0.7641 (0.7619)	0.7589 (0.7619)

Consider the Table 4.1 values of IS_3 for different combinations of $\sigma^2_{w_3}$ and $\sigma^2_{s_3}$, holding other parameters constant. As I would expect, information shares increase with the variance of information (going from left to right), but not with the variance of noise (they stay relatively unchanged from top to bottom of each column). Qualitatively, this confirms that the model recovers the simulated variation in information and noise.

Consider how the estimated values in Table 4.1 compare to the true values of IS_3 (reported in parentheses). The results suggest that the model recovers unbiased information shares, as mean estimates from the simulation are in line with the true values of IS_3. I confirm that the IS estimates for other phases (i.e., IS_1, IS_2, in addition to IS_3) are also in line with their true values. These results are in Appendix 4.1 (Table 4.7).

Table 4.3: Estimates of information-to-noise ratios from the model with $n = 3$ phases

This table reports the mean estimates of information-to-noise ratios in phase 3 (IN_3) from the VAR model with three phases ($n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information-to-noise ratios are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phases 1 and 2 are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. Information-to-noise ratio in phase 3 is $IN_3 = \frac{Var(w_{3,t})}{Var(w_{3,t})+Var(s_{3,t})}$. The detailed description of the simulated model is in Section 4.3.1. The description of Monte Carlo procedure is in Section 4.4.

Noise variance, $\sigma^2_{s_3}$	Information variance, $\sigma^2_{w_3}$				
	0.10^2	0.20^2	0.30^2	0.40^2	0.50^2
0.20^2	0.2352 (0.2000)	0.5230 (0.5000)	0.7105 (0.6923)	0.8157 (0.8000)	0.8749 (0.8621)
0.25^2	0.1704 (0.1379)	0.4131 (0.3902)	0.6087 (0.5902)	0.7354 (0.7191)	0.8149 (0.8000)
0.30^2	0.1297 (0.1000)	0.3300 (0.3077)	0.5186 (0.5000)	0.6565 (0.5000)	0.7505 (0.7353)
0.35^2	0.1029 (0.0755)	0.2678 (0.2462)	0.4421 (0.4235)	0.5830 (0.5664)	0.6865 (0.6711)
0.40^2	0.0846 (0.0588)	0.2210 (0.2000)	0.3785 (0.3600)	0.5167 (0.5000)	0.6253 (0.6098)

In addition to information shares, the proposed method also allows me to discern noise terms in each phase. I call the relative contribution of noise in a given phase "noise share". The model estimates for phase 3 noise shares are in Table 4.2. As described before for IS_3, I vary the phase 3 information variance and noise variance, and estimate NS_3 for each combination of these parameters. I follow Equation (4.17) to compute $NS_3 = \frac{Var(s_{3,t})}{Var(s_{1,t})+Var(s_{2,t})+Var(s_{3,t})}$.

Qualitatively, the results in Table 4.2 show that the model captures variation in noise. Noise shares increase from top to bottom, as I dial up the variance of noise. At the same time, noise shares are relatively unchanged from left to right, as higher information variances do not affect noise shares. Quantitatively, the mean estimates are close to the true noise shares reported in parentheses. In Appendix 4.1 (Table 4.8), I confirm that this is the case not only for phase three, but also for other phases.

In addition to information shares and noise shares, a third way to gauge the

price informativeness in a given phase is to estimate information-to-noise ratios. I scale information variances in each phase by the sum of information and noise variance, and call the resulting measure "information-to-noise ratio" (IN). Table 4.3 reports IN_3 recovered from the simulated data. I follow Equation (4.18) to compute $IN_3 = \frac{Var(w_{3,t})}{Var(w_{3,t})+Var(s_{3,t})}$.

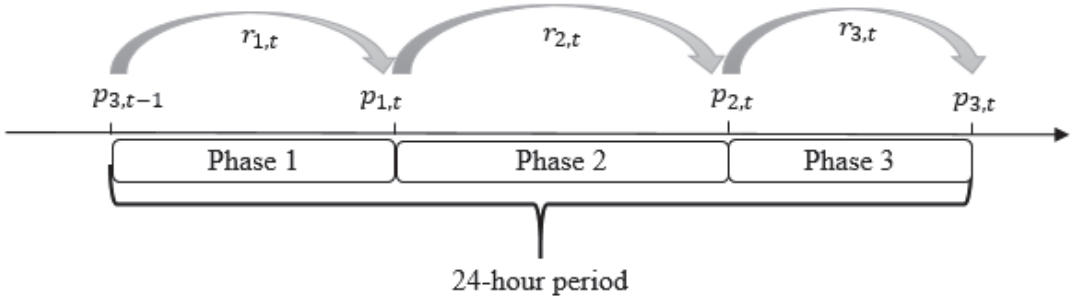
Results in Table 4.3 confirm that information-to-noise ratios are broadly in line with the theoretical values computed from the simulated data. The estimates are somewhat less precise than those for information shares and noise shares; this is consistent with greater parameter uncertainty when both information variance estimates and noise variance estimates are combined in a single measure (as opposed to having only information variances in, e.g., IS ratios). Table 4.9 in Appendix 4.1 confirms this result for all three phases.

4.4.2 Model estimation with n=2 phases

For certain applications, it might be of interest to estimate the model with $n = 2$ phases. For example, I do so in the next section to infer intraday patterns in information and noise. But do the estimated IS, NS, IN align between the versions of the empirical VAR model with $n = 2$ and $n = 3$? To test this, I estimate the VAR model using two phases instead of three. Phases 1 and 2 are joined into one (which I denote $1 \cup 2$), and phase 3 stays the same. I rely on the same simulated price series as before, but compute observed returns differently. Observed return in phase $(1 \cup 2)$ is $r_{1\cup 2,t} = p_{2,t} - p_{3,t-1}$, and return in phase 3 is the same as before: $r_{3,t} = p_{3,t} - p_{2,t}$. Figure 4.2 illustrates the correspondence between observed returns in three phases and two phases.

Because phase 3 is the same across the two estimation approaches, the information incorporated in that phase should account for the same proportion of overall 24-hour variance, regardless which estimation approach I use. Therefore, the model with $n = 2$ should recover IS_3 that is in line with IS_3 estimated from the model with $n = 3$. I show that it is indeed the case: the results in Table 4.4 (using the model with two phases) are well aligned with the results in Table 4.1 (using the model with three phases). As I compare IS_3 estimates to their true values, the model with $n = 2$ recovers IS_3 somewhat more accurately compared to the model with $n = 3$. The latter is to be expected, as standard errors are smaller when

Panel A. Observed returns with three sequential phases



Panel B. Observed returns with two sequential phases

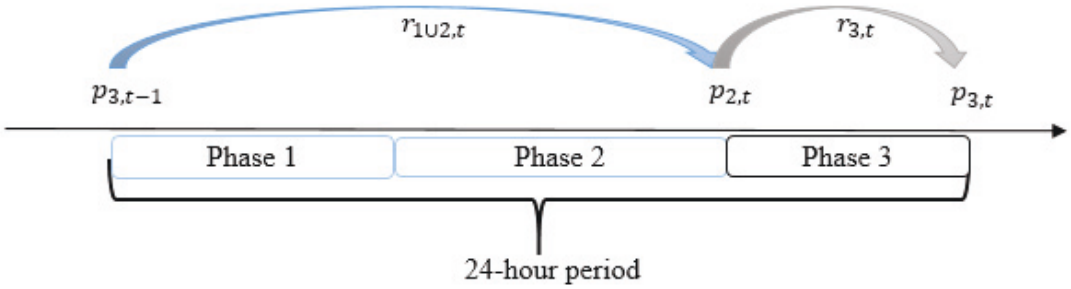


Figure 4.2: Correspondence between observed returns with three vs two sequential phases

This figure illustrates the notation with $n = 3$ and $n = 2$ sequential phases. I simulate the observed log-prices $p_{1,t}, p_{2,t}, p_{3,t}$ as described in Section 4.3.1.When I estimate the model with $n = 3$ phases in the 24-hour period t, the observed log-returns are: $r_{1,t} = p_{1,t} - p_{3,t-1}$, $r_{2,t} = p_{2,t} - p_{1,t}$, $r_{3,t} = p_{3,t} - p_{2,t}$. When I estimate the model with $n = 2$ phases, the observed log-returns are: $r_{1\cup2,t} = p_{2,t} - p_{3,t-1}$, $r_{3,t} = p_{3,t} - p_{2,t}$.

fewer parameters are estimated (as is the case in the model with $n = 2$ compared to the model with $n = 3$).

As I combine phases 1 and 2 into a single phase (denoted as phase $1 \cup 2$), the amount of information incorporated during this phase is equivalent to the sum of information in phases 1 and 2. Therefore, the efficient return innovation in phase $(1 \cup 2)$ is $w_{1\cup2,t} = w_{1,t} + w_{2,t}$. Hence, the recovered information share in phase $(1 \cup 2)$ should be close to the theoretical value $IS_{1\cup2} = \frac{Var(w_{1,t})+Var(w_{2,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(w_{3,t})}$. The results in Appendix 4.1 (Table 4.10) confirm that this is indeed the case.

I have shown that the model recovers information shares that are consistent between the different number of phases. Therefore, I can apply the model to settings

Table 4.4: Estimates of information shares from the model with $n = 2$ phases

This table reports the mean estimates of information shares in phase 3 (IS_3) from the VAR model with two phases: $(1 \cup 2)$ and 3 (phase $(1 \cup 2)$ corresponds to joint phases 1 and 2 in the model with $n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phase $(1 \cup 2)$ are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. Information share in phase 3 is $IS_3 = \frac{Var(w_{3,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(w_{3,t})}$. The detailed description of the simulated model is in Section 4.3.1. The description of Monte Carlo procedure is in Section 4.4.

Noise variance, $\sigma^2_{s_3}$	Information variance, $\sigma^2_{w_3}$				
	0.10^2	0.20^2	0.30^2	0.40^2	0.50^2
0.20^2	0.1756 (0.1667)	0.4440 (0.4444)	0.6402 (0.6429)	0.7591 (0.7619)	0.8309 (0.8333)
0.25^2	0.1784 (0.1667)	0.4441 (0.4444)	0.6397 (0.6429)	0.7584 (0.7619)	0.8296 (0.8333)
0.30^2	0.1812 (0.1667)	0.4443 (0.4444)	0.6393 (0.6429)	0.7578 (0.7619)	0.8289 (0.8333)
0.35^2	0.1843 (0.1667)	0.4445 (0.4444)	0.6389 (0.6429)	0.7572 (0.7619)	0.8282 (0.8333)
0.40^2	0.1876 (0.1667)	0.4448 (0.4444)	0.6387 (0.6429)	0.7566 (0.7619)	0.8275 (0.8333)

with phases of unequal duration. One scenario of such application is when markets open sequentially, with unequal duration of trading sessions. Another scenario is when different trading mechanisms (such as opening auctions, continuous trading, and closing auctions) take place sequentially, but operate for unequal time periods.

I also estimate noise shares in the model with $n = 2$ phases. Because noise terms enter the observed returns calculation at a point in time rather than adding up over a period of time (as information terms do), I do not expect noise shares to align between the models with $n = 2$ and $n = 3$ phases. I only check whether the noise shares are in line with their theoretical values: $NS_{1\cup2} = \frac{Var(s_{2,t})}{Var(s_{2,t})+Var(s_{3,t})}$, $NS_3 = \frac{Var(s_{3,t})}{Var(s_{2,t})+Var(s_{3,t})}$. Table 4.5 confirms that the recovered NS_3 are close to their theoretical values. Appendix 4.1 (Table 4.11) confirms this is observed for the phase $(1 \cup 2)$ as well.

Finally, I estimate information-to-noise ratios in the model with $n = 2$ phases.

Table 4.5: Estimates of noise shares from the model with $n = 2$ phases

This table reports the mean estimates of noise shares in phase 3 (IS_3) from the VAR model with two phases: $(1 \cup 2)$ and 3 (phase $(1 \cup 2)$ corresponds to joint phases 1 and 2 in the model with $n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true noise shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phase $(1 \cup 2)$ are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. Noise share in phase 3 is $NS_3 = \frac{Var(s_{3,t})}{Var(s_{2,t})+Var(s_{3,t})}$. The detailed description of the simulated model is in Section 4.3.1. The description of Monte Carlo procedure is in Section 4.4.

Noise variance, $\sigma^2_{s_3}$	Information variance, $\sigma^2_{w_3}$				
	0.10^2	0.20^2	0.30^2	0.40^2	0.50^2
0.20^2	0.5049 (0.4444)	0.5031 (0.4444)	0.4978 (0.4444)	0.4854 (0.4444)	0.4675 (0.4444)
0.25^2	0.6152 (0.5556)	0.6151 (0.5556)	0.6137 (0.5556)	0.6089 (0.5556)	0.5973 (0.5556)
0.30^2	0.6975 (0.6429)	0.6981 (0.6429)	0.6984 (0.6429)	0.6973 (0.6429)	0.6931 (0.6429)
0.35^2	0.7584 (0.7101)	0.7594 (0.7101)	0.7604 (0.7101)	0.7608 (0.7101)	0.7597 (0.7101)
0.40^2	0.8040 (0.7619)	0.8050 (0.7619)	0.8063 (0.7619)	0.8074 (0.7619)	0.8075 (0.7619)

Given that information variance and noise variance attributed to phase 3 is the same across $n = 2$ and $n = 3$, I expect the IN_3 to align between the two approaches. Table 4.6 confirms that it is indeed the case. The 2-phase model recovers IN_3 that is close to the theoretical value $IN_3 = \frac{Var(w_{3,t})}{Var(w_{3,t})+Var(s_{3,t})}$. Appendix 4.1 (Table 4.12) results confirm that it also recovers information-to-noise ratio in the phase $(1 \cup 2)$: $IN_{1\cup2} = \frac{Var(w_{1,t})+Var(w_{2,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(s_{1,t})+Var(s_{2,t})}$.

4.5 Illustrative application to intraday patterns in price discovery

To illustrate the empirical application of the price discovery model, I estimate information shares, noise shares, and information-to-noise ratios for intraday prices of 233 US stocks. The sample spans from 2002 to 2018, a period that is sufficiently

Table 4.6: Estimates of information-to-noise ratios from the model with $n = 2$ phases

This table reports the mean estimates of information-to-noise ratios in phase 3 (IN_3) from the VAR model with two phases: $(1 \cup 2)$ and 3 (phase $(1 \cup 2)$ corresponds to joint phases 1 and 2 in the model with $n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information-to-noise ratios are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phase $(1 \cup 2)$ are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. Information-to-noise ratio in phase 3 is $IN_3 = \frac{Var(w_{3,t})}{Var(w_{3,t})+Var(s_{3,t})}$. The detailed description of the simulated model is in Section 4.3.1. The description of Monte Carlo procedure is in Section 4.4.

Noise variance, $\sigma^2_{s_3}$	Information variance, $\sigma^2_{w_3}$				
	0.10^2	0.20^2	0.30^2	0.40^2	0.50^2
0.20^2	0.2263 (0.2000)	0.5178 (0.5000)	0.7067 (0.6923)	0.8123 (0.8000)	0.8719 (0.8621)
0.25^2	0.1629 (0.1379)	0.4083 (0.3902)	0.6049 (0.5902)	0.7320 (0.7191)	0.8117 (0.8000)
0.30^2	0.1229 (0.1000)	0.3255 (0.3077)	0.5150 (0.5000)	0.6532 (0.6400)	0.7473 (0.7353)
0.35^2	0.0965 (0.0755)	0.2633 (0.2462)	0.4386 (0.4235)	0.5798 (0.5664)	0.6833 (0.6711)
0.40^2	0.0783 (0.0588)	0.2165 (0.2000)	0.3750 (0.3600)	0.5135 (0.5000)	0.6221 (0.6098)

long, yet representative of modern market structure post-decimalization. [9] I then examine the average IS, NS, IN ratios across these stocks. This exercise allows me to gauge market-wide patterns in information and noise throughout the trading day. I compare these patterns with the predictions from microstructure theory, and with the prior empirical studies of intraday variances.

In the estimation, I use the method from Section 4.3.1, and apply the empirical VAR described in Section 4.3.2. I estimate multiple VAR models with $n = 2$

[9]The stock sampling approach is the same as in Section 2.4 of Chapter 2 to ensure that the sample is representative of modern markets. That is, I select 241 US stocks at random, using stratification by size. For the purpose of VAR estimation, I apply two filters to ensure sufficient data: (i) a stock must have at least 100 daily price observations in a given year, and (ii) have no gaps in consecutrive trading days that are longer than one week. After filters, I have 233 stocks in year 2018. The number of stocks varies slightly from year to year due to the two filters mentioned above. The number of stocks is within the range from 191 (in year 2002) to 233 (in year 2018). The official open and close, and intraday trading prices are from Thomson Reuters Tick History database.

phases. For example, the first VAR uses close-to-open[10] log-return and open-to-9:40 AM log-return on day t: $r_{1,t} = p_{open,t} - p_{close,t-1}$, $r_{2,t} = p_{9:40,t} - p_{open,t}$. The second VAR and all subsequent VARs use 10-minute sliding window of returns: $r_{1,t} = p_{9:40,t} - p_{9:50,t-1}, r_{2,t} = p_{9:50,t} - p_{9:40,t}$, and so on. In the last VAR, $r_{1,t} = p_{15:50,t} - p_{close,t-1}, r_{2,t} = p_{close,t} - p_{15:50,t}$. Each VAR (except for the first one) uses returns from two phases of 24-hour perios: (i) phase 1 is a period of 24 hours less 10 minutes, and (ii) phase 2 is a period of 10 minutes. The log-returns in each phase rely on trade prices from the US consolidated data feed.

Because I am interested in intraday patterns, I split the trading day into 10-minute buckets, and select the price of the last trade within each bucket. For example, the 10:30 AM price is the last trade price available within the 10:20 AM – 10:30 AM bucket. If no trades occurred within the period 10:20 AM – 10:30 AM, I take the last available price from the 10:10 AM – 10:20 AM bucket. For any missing values of trade prices, I fill forward prices within a stock-day, but not overnight. I run each VAR model per stock-year, hence the estimates of IS, NS, IN are at stock-year frequency. I plot IS, NS, IN averaged per stock-year, and these estimates of IS, NS, IN correspond to 10-minute window. For example an average IS of 0.04 at 9:50 AM means the average information share of the period 9:40 AM (t) – 9:50 AM (t) is 0.04. The IS of the period 9:50 AM ($t-1$) – 9:40 AM (t) is therefore approximately (1-0.04)=0.96 on average (it is approximate due to being the average across stocks; it is exact at stock-year level).

I start by estimating the VAR for year 2018, and plot the resulting intraday patterns in Figure 4.3. The IS, NS, IN estimates are at stock-year-frequency, averaged across around 233 stocks in year 2018. Intuitively, I expect the information shares to be relatively high at the start of the trading day, as overnight news is incorporated into prices. However, opening prices and early trade prices of the day are also likely to be noisy, as early in the trading day there has been still relatively little time to process the overnight news. Hence, I expect high noise shares at the start of the day. The plots in Figure 4.3 confirm this intuition, and also show that net price informativeness (information-to-noise ratios) follows an L-shaped curve throughout the trading day. This patterns suggests that as the trading session progresses, relatively little new information is incorporated in each subsequent 10-minute period. A slight increase in information-to-noise at the end

[10]The US market opens at 9:30 AM, and closes at 4:00 PM. In the data, the 9:30 AM price is the official opening price, and the 4:00 PM price is the official closing price.

of the trading day corresponds to the substantial trading in the closing auctions of NYSE and NASDAQ.

I also examine the IS, NS, IN ratios in years 2002–2018. Note that the mid-day patterns (fairly flat IS, NS, IN) are parsimonious is year 2018, so to reduce the computational burden in the multi-year analysis, I split the trading day into five periods (rather into than into 37 periods of 10 minutes). This means I estimate five VAR models with $n = 2$, and the general estimation procedure is the same as described before. The results are plotted in Figure 4.4.

The intraday IS patterns in those years resemble the L-shaped curve that I obtain for year 2018. IN and NS patterns are close to the U-shape, also similar to year 2018. There is more variability in ratios that involve noise terms (NS, IN), as standard errors are higher for noise variances compared to information variances.

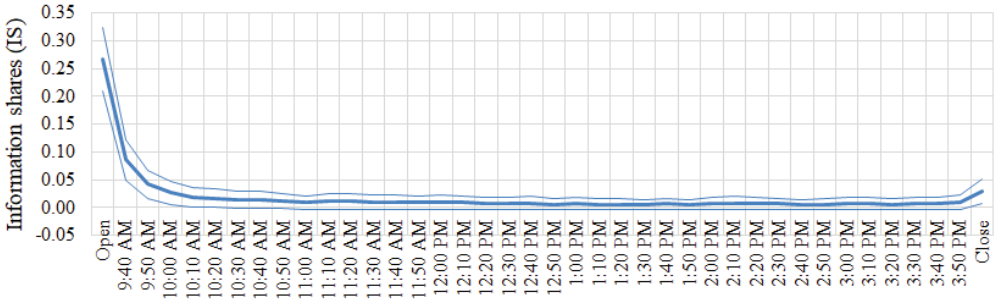
The results corroborate the earlier studies on volatility, liquidity, and trading volume throughout the day. However, I offer a more nuanced picture that one does not get from intraday patterns of spreads or volatility. For example, existing literature suggests that volatility and spreads tend to follow a U-shaped curve throughout the trading day (McInish & Wood, 1990; Wood, Mcinish, & Ord, 1985; Harris, 1986). I confirm this stylized fact in the analysis in Figure 4.3 and Figure 4.4.

The IS pattern follows an L-shaped curve, consistent with increased spreads, volumes, and volatility early in the trading day (as in McInish & Wood, 1990; Lockwood & Linn, 1990; Cai, Hudson & Keasey, 2004). Because information asymmetry at the start of the day is high, it is not surprising to observe relatively high price informativeness, albeit in the presence of relatively high noise.

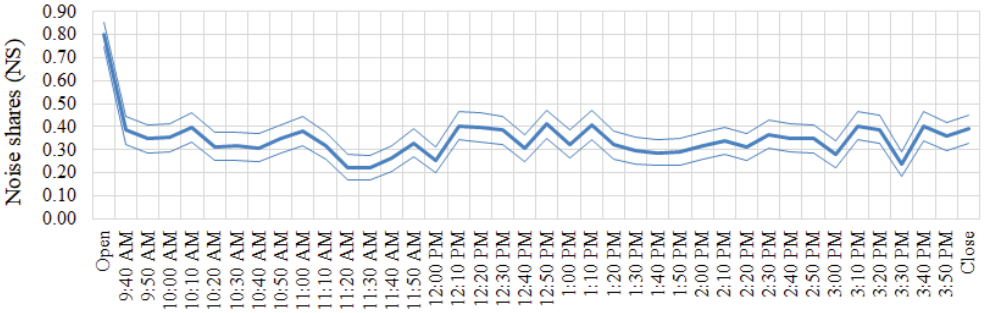
The IN results are consistent with Hasbrouck (1991a), who finds that trades tend to be more informative at the beginning of trading. Figure 4.3 confirms that information-to-noise ratios tend to be high at the market open, but decrease gradually thereafter. Following the open, every subsequent 10-minute trading phase incorporates relatively less new information into prices.

The NS results suggest that noise shares are elevated at the start and at the end of the trading day, consistent with Hasbrouck (1993). The results suggest that there is only a slight increase in relative price informativeness towards the close, consistent with Amihud & Mendelson (1987). The latter show that open-to-open

Panel A. Information shares



Panel B. Noise shares



Panel C. Information-to-noise ratios

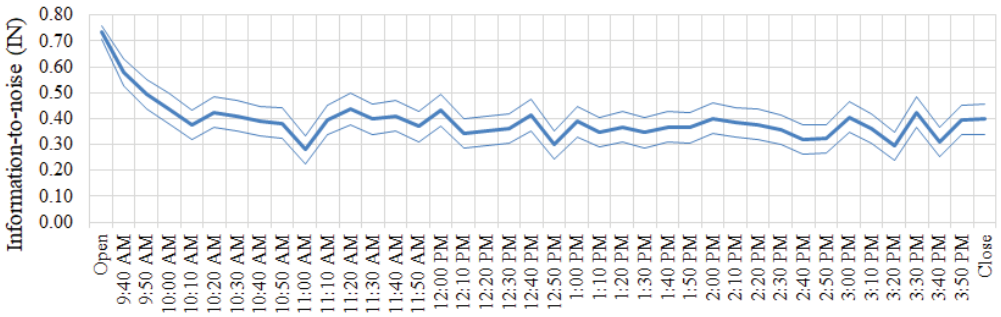
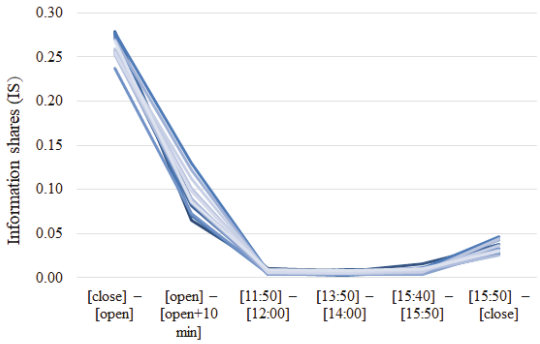


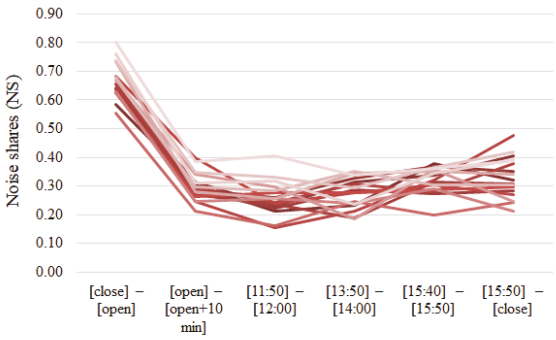
Figure 4.3: Intraday patterns in the US market, 2018

This figure plots mean information shares (IS), noise shares (NS), and information-to-noise ratios (IN = Info / (Info + Noise)) of the intraday prices in the US market. The thick line represents mean estimates, the thin lines represent 95 % confidence bounds. Prices are sampled every 10 minutes from the consolidated trading feed for the year 2018. Averages are computed across 233 randomly selected stocks (stratified by market cap). The estimates are from the 2-equation VAR model with a sliding 10-min window. For example, the IS estimate at 10:30 AM corresponds to information share of the period 10:20 AM - 10:30 AM, relative to the rest of the 24-hour period (10:30 AM on day $t-1$ till 10:20 AM on day t).

Panel A. Information shares



Panel B. Noise shares



Panel C. Information-to-noise ratios

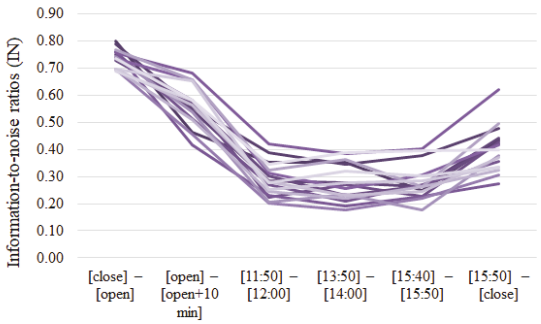


Figure 4.4: Intraday patterns in the US market, 2002 – 2018

This figure plots mean information shares (IS), noise shares (NS), and information-to-noise ratios (IN = Info / (Info + Noise)). of the intraday prices in the US market. Each line represents a different year from 2002 to 2018, with darker lines corresponding to earlier years. Estimates are averaged across 233 stocks per year. The estimates are from the 2-equation VAR model with a sliding 10-min window. For example, the IS estimate at 10:30 AM corresponds to information share of the period 10:20 AM – 10:30 AM, relative to the rest of the 24-hour period (10:30 AM on day $t-1$ till 10:20 AM on day t).

volatility is higher than close-to-close. Because the total amount of information between open-to-open and close-to-close should be the same, their findings imply that the close is relatively noisier than open, consistent with Figure 4.3. Similarly, Cushing & Madhavan (2000) find systematic return reversals following pre-close imbalance publications,[11] which is also in line with Figure 4.3 showing an uptick in noise towards the market close.

Overall, the empirical application of the proposed methodology suggests that the model recovers the expected patterns in price informativeness and noise. The advantage of the proposed methodology lies in separating information variances from noise variances. Based on evidence from the Monte Carlo simulations and from the illustrative application to intraday patterns, the model offers a useful tool for variance decomposition in sequential trading phases.

4.6 Conclusions and future research

This chapter develops a novel methodology to decompose sequential returns into information and noise components. The decomposition provides "information shares", "noise shares", and "information-to-noise ratios" for each trading phase. The method can be viewed as the sequential markets analogue of traditional price discovery methods like Hasbrouck (1995) that are instead applicable to parallel markets. Additionally, the proposed approach is more explicit in separating noise from information in prices.

Using Monte Carlo simulations, I confirm that the proposed methodology recovers true information / noise shares when pricing errors in each phase are independent. I illustrate an empirical application of the method by quantifying the intraday patterns of information and noise in stock prices. The results show that information-to-noise ratios follow an L-shaped pattern during the trading day, with relatively high information-to-noise ratios at market open, and a smaller increase at the market close.

Potential future applications of this new approach to measuring price discovery include sequential market settings such as the opening mechanism vs continuous trading vs the closing mechanism vs overnight trading. Another example is studying securities that trade almost continuously, but in different time zones (e.g., many

[11] Order imbalances are published every 5 seconds between 15:50 and 16:00 on NYSE.

futures markets, forex markets), and quantifying how much each time zone contributes to price discovery (information) and noise. Finally, ultra high-frequency applications of this methodology could be used to examine the effects of trading algorithms clustering their order submissions and revisions intrasecond.

Appendix 4.1. Additional results from Monte Carlo simulations

This Appendix reports additional results from Monte Carlo simulations that are referenced in the main text.

Table 4.7: Estimates of information shares from the model with $n = 3$ phases

This table reports the mean estimates of information shares in phases 1, 2, and 3 (IS_1, IS_2, IS_3) from the VAR model with three phases ($n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phases 1 and 2 are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. The simulated returns are: $r_{i,t} = w_{i,t} + s_{i,t} - s_{i-1,t}$, where $r_{i,t}$ is the observed return in phase i of the 24-hour period t, $w_{i,t} \sim N(0, \sigma^2_{w_i})$ is the efficient price innovation, $s_{i,t} \sim N(0, \sigma^2_{s_i})$ is the pricing error. I estimate variances of information and noise ($Var\,(w_{i,t}) = \sigma^2_{w_i}$ and $Var\,(s_{i,t}) = \sigma^2_{s_i}$) following the formulas in the main text. Information share in phase i is $IS_i = \frac{Var(w_{i,t})}{\sum_{i=1}^{n} Var(w_{i,t})}$, where the number of phases is $n = 3$.

Results from VAR model with n=3 phases in the 24-hour period															
Noise variance, $\sigma^2_{s_1}, \sigma^2_{s_2}, \sigma^2_{s_3}$	Information variances, $\sigma^2_{w_1}, \sigma^2_{w_2}, \sigma^2_{w_3}$														
	0.10^2	0.20^2	0.10^2	0.10^2	0.20^2	0.20^2	0.10^2	0.20^2	0.30^2	0.10^2	0.20^2	0.40^2	0.10^2	0.20^2	0.50^2
$0.10^2, 0.20^2, 0.20^2$	0.1715 (0.1667)	0.6504 (0.6667)	0.1781 (0.1667)	0.1206 (0.1111)	0.4393 (0.4444)	0.4400 (0.4444)	0.0846 (0.0714)	0.2835 (0.2857)	0.6319 (0.6429)	0.0632 (0.0476)	0.1885 (0.1905)	0.7483 (0.7619)	0.0501 (0.0333)	0.1312 (0.1333)	0.8187 (0.8333)
$0.10^2, 0.20^2, 0.25^2$	0.1734 (0.1667)	0.6463 (0.6667)	0.1804 (0.1667)	0.1234 (0.1111)	0.4378 (0.4444)	0.4388 (0.4444)	0.0871 (0.0714)	0.2828 (0.2857)	0.6301 (0.6429)	0.0651 (0.0476)	0.1879 (0.1905)	0.7471 (0.7619)	0.0522 (0.0333)	0.1303 (0.1333)	0.8175 (0.8333)
$0.10^2, 0.20^2, 0.30^2$	0.1759 (0.1667)	0.6411 (0.6667)	0.1830 (0.1667)	0.1265 (0.1111)	0.4359 (0.4444)	0.4376 (0.4444)	0.0899 (0.0714)	0.2819 (0.2857)	0.6282 (0.6429)	0.0671 (0.0476)	0.1873 (0.1905)	0.7456 (0.7619)	0.0538 (0.0333)	0.1296 (0.1333)	0.8166 (0.8333)
$0.10^2, 0.20^2, 0.35^2$	0.1790 (0.1667)	0.6347 (0.6667)	0.1862 (0.1667)	0.1300 (0.1111)	0.4335 (0.4444)	0.4365 (0.4444)	0.0929 (0.0714)	0.2808 (0.2857)	0.6263 (0.6429)	0.0695 (0.0476)	0.1866 (0.1905)	0.7439 (0.7619)	0.0555 (0.0333)	0.1289 (0.1333)	0.8156 (0.8333)
$0.10^2, 0.20^2, 0.40^2$	0.1827 (0.1667)	0.6273 (0.6667)	0.1900 (0.1667)	0.1338 (0.1111)	0.4308 (0.4444)	0.4354 (0.4444)	0.0964 (0.0714)	0.2795 (0.2857)	0.6242 (0.6429)	0.0721 (0.0476)	0.1858 (0.1905)	0.7420 (0.7619)	0.0573 (0.0333)	0.1283 (0.1333)	0.8144 (0.8333)

Table 4.8: Estimates of noise shares from the model with $n = 3$ phases

This table reports the mean estimates of noise shares in phases 1, 2, and 3 (NS_1, NS_2, NS_3) from the VAR model with three phases ($n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true noise shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phases 1 and 2 are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. The simulated returns are: $r_{i,t} = w_{i,t} + s_{i,t} - s_{i-1,t}$, where $r_{i,t}$ is the observed return in phase i of the 24-hour period t, $w_{i,t} \sim N(0, \sigma^2_{w_i})$ is the efficient price innovation, $s_{i,t} \sim N(0, \sigma^2_{s_i})$ is the pricing error. I estimate variances of information and noise ($Var\,(w_{i,t}) = \sigma^2_{w_i}$ and $Var\,(s_{i,t}) = \sigma^2_{s_i}$) following the formulas in the main text. Noise share in phase i is $NS_i = \frac{Var(s_{i,t})}{\sum_{i=1}^{n} Var(s_{i,t})}$, where the number of phases is $n = 3$.

Results from VAR model with n=3 phases in the 24-hour period															
Noise variance, $\sigma^2_{s_1}, \sigma^2_{s_2}, \sigma^2_{s_3}$	Information variances, $\sigma^2_{w_1}, \sigma^2_{w_2}, \sigma^2_{w_3}$														
	0.10^2	0.20^2	0.10^2	0.10^2	0.20^2	0.20^2	0.10^2	0.20^2	0.30^2	0.10^2	0.20^2	0.40^2	0.10^2	0.20^2	0.50^2
$0.10^2, 0.20^2, 0.20^2$	0.1074 (0.1111)	0.4436 (0.4444)	0.4489 (0.4444)	0.1064 (0.1111)	0.4473 (0.4444)	0.4462 (0.4444)	0.1075 (0.1111)	0.4537 (0.4444)	0.4388 (0.4444)	0.1125 (0.1111)	0.4644 (0.4444)	0.4231 (0.4444)	0.1203 (0.1111)	0.4765 (0.4444)	0.4022 (0.4444)
$0.10^2, 0.20^2, 0.25^2$	0.0852 (0.0889)	0.3536 (0.3556)	0.5613 (0.5556)	0.0837 (0.0889)	0.3559 (0.3556)	0.5604 (0.5556)	0.0841 (0.0889)	0.3593 (0.3556)	0.5567 (0.5556)	0.0881 (0.0889)	0.3646 (0.3556)	0.5473 (0.5556)	0.0950 (0.0889)	0.3745 (0.3556)	0.5305 (0.5556)
$0.10^2, 0.20^2, 0.30^2$	0.0680 (0.0714)	0.2833 (0.2857)	0.6487 (0.6429)	0.0664 (0.0714)	0.2847 (0.2857)	0.6488 (0.6429)	0.0665 (0.0714)	0.2866 (0.2857)	0.6469 (0.6429)	0.0696 (0.0714)	0.2892 (0.2857)	0.6412 (0.6429)	0.0752 (0.0714)	0.2940 (0.2857)	0.6307 (0.6429)
$0.10^2, 0.20^2, 0.35^2$	0.0550 (0.0580)	0.2294 (0.2319)	0.7156 (0.7101)	0.0535 (0.0580)	0.2304 (0.2319)	0.7162 (0.7101)	0.0533 (0.0580)	0.2315 (0.2319)	0.7152 (0.7101)	0.0560 (0.0580)	0.2328 (0.2319)	0.7113 (0.7101)	0.0607 (0.0580)	0.2350 (0.2319)	0.7043 (0.7101)
$0.10^2, 0.20^2, 0.40^2$	0.0450 (0.0476)	0.1882 (0.1905)	0.7668 (0.7619)	0.0436 (0.0476)	0.1888 (0.1905)	0.7675 (0.7619)	0.0435 (0.0476)	0.1895 (0.1905)	0.7670 (0.7619)	0.0457 (0.0476)	0.1902 (0.1905)	0.7641 (0.7619)	0.0499 (0.0476)	0.1912 (0.1905)	0.7589 (0.7619)

Table 4.9: Estimates of information-to-noise ratios from the model with $n = 3$ phases

This table reports the mean estimates of information-to-noise ratios in phases 1, 2, and 3 (IN_1, IN_2, IN_3) from the VAR model with three phases ($n = 3$). The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information-to-noise ratios are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phases 1 and 2 are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. The simulated returns are: $r_{i,t} = w_{i,t} + s_{i,t} - s_{i-1,t}$, where $r_{i,t}$ is the observed return in phase i of the 24-hour period t, $w_{i,t} \sim N(0, \sigma^2_{w_i})$ is the efficient price innovation, $s_{i,t} \sim N(0, \sigma^2_{s_i})$ is the pricing error. I estimate variances of information and noise ($Var(w_{i,t}) = \sigma^2_{w_i}$ and $Var(s_{i,t}) = \sigma^2_{s_i}$) following the formulas in the main text. Information-to-noise ratio in phase i is $IN_i = \frac{Var(w_{i,t})}{Var(w_{i,t})+Var(s_{i,t})}$.

Results from VAR model with n=3 phases in the 24-hour period															
Noise variance, $\sigma^2_{s_1}, \sigma^2_{s_2}, \sigma^2_{s_3}$	Information variances, $\sigma^2_{w_1}, \sigma^2_{w_2}, \sigma^2_{w_3}$														
	0.10^2	0.20^2	0.10^2	0.10^2	0.20^2	0.20^2	0.10^2	0.20^2	0.30^2	0.10^2	0.20^2	0.40^2	0.10^2	0.20^2	0.50^2
$0.10^2, 0.20^2, 0.20^2$	0.5335 (0.5000)	0.5207 (0.5000)	0.2352 (0.2000)	0.5362 (0.5000)	0.5206 (0.5000)	0.5230 (0.5000)	0.5383 (0.5000)	0.5193 (0.5000)	0.7105 (0.6923)	0.5298 (0.5000)	0.5161 (0.5000)	0.8157 (0.8000)	0.5207 (0.5000)	0.5117 (0.5000)	0.8749 (0.8621)
$0.10^2, 0.20^2, 0.25^2$	0.5358 (0.5000)	0.5211 (0.5000)	0.1704 (0.1379)	0.5430 (0.5000)	0.5213 (0.5000)	0.4131 (0.3902)	0.5483 (0.5000)	0.5197 (0.5000)	0.6087 (0.5902)	0.5388 (0.5000)	0.5156 (0.5000)	0.7354 (0.7191)	0.5220 (0.5000)	0.5100 (0.5000)	0.8149 (0.8000)
$0.10^2, 0.20^2, 0.30^2$	0.5384 (0.5000)	0.5214 (0.5000)	0.1297 (0.1000)	0.5462 (0.5000)	0.5217 (0.5000)	0.3300 (0.3077)	0.5572 (0.5000)	0.5199 (0.5000)	0.5186 (0.5000)	0.5395 (0.5000)	0.5151 (0.5000)	0.6565 (0.6400)	0.5238 (0.5000)	0.5083 (0.5000)	0.7505 (0.7353)
$0.10^2, 0.20^2, 0.35^2$	0.5412 (0.5000)	0.5215 (0.5000)	0.1029 (0.0755)	0.5509 (0.5000)	0.5220 (0.5000)	0.2678 (0.2462)	0.5586 (0.5000)	0.5201 (0.5000)	0.4421 (0.4235)	0.5448 (0.5000)	0.5147 (0.5000)	0.5830 (0.5664)	0.5305 (0.5000)	0.5066 (0.5000)	0.6865 (0.6711)
$0.10^2, 0.20^2, 0.40^2$	0.5445 (0.5000)	0.5216 (0.5000)	0.0846 (0.0588)	0.5540 (0.5000)	0.5222 (0.5000)	0.2210 (0.2000)	0.5640 (0.5000)	0.5203 (0.5000)	0.3785 (0.3600)	0.5524 (0.5000)	0.5143 (0.5000)	0.5167 (0.5000)	0.5353 (0.5000)	0.5052 (0.5000)	0.6253 (0.6098)

Table 4.10: Estimates of information shares from the model with $n = 2$ phases

This table reports the mean estimates of information shares in phases $(1 \cup 2)$, and 3 $(IS_{1\cup2}, IS_3)$ from the VAR model with two phases $(n = 2)$. Phase $(1 \cup 2)$ corresponds to joint phases 1 and 2 in the model with $n = 3$. The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phase $(1 \cup 2)$ are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. The simulated returns are: $r_{i,t} = w_{i,t} + s_{i,t} - s_{i-1,t}$, where $r_{i,t}$ is the observed return in phase i of the 24-hour period t, $w_{i,t} \sim N(0, \sigma^2_{w_i})$ is the efficient price innovation, $s_{i,t} \sim N(0, \sigma^2_{s_i})$ is the pricing error. I estimate variances of information and noise ($Var(w_{i,t}) = \sigma^2_{w_i}$ and $Var(s_{i,t}) = \sigma^2_{s_i}$) following the formulas in the main text. Information share in phase $(1 \cup 2)$ is $IS_{1\cup2} = \frac{Var(w_{1,t})+Var(w_{2,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(w_{3,t})}$. Information share in phase 3 is $IS_3 = \frac{Var(w_{3,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(w_{3,t})}$.

Results from VAR model with n=2 phases in the 24-hour period										
Noise variance, $\sigma^2_{s_1}, \sigma^2_{s_2}, \sigma^2_{s_3}$	Information variances, $[\sigma^2_{w_1} + \sigma^2_{w_2}]$, $\sigma^2_{w_3}$									
	$[0.10^2 + 0.20^2]$	0.10^2	$[0.10^2 + 0.20^2]$	0.20^2	$[0.10^2 + 0.20^2]$	0.30^2	$[0.10^2 + 0.20^2]$	0.40^2	$[0.10^2 + 0.20^2]$	0.50^2
$[0.10^2 + 0.20^2], 0.20^2$	0.8244 (0.8333)	0.1756 (0.1667)	0.5560 (0.5556)	0.4440 (0.4444)	0.3598 (0.3571)	0.6402 (0.6429)	0.2409 (0.2381)	0.7591 (0.7619)	0.1691 (0.1667)	0.8309 (0.8333)
$[0.10^2 + 0.20^2], 0.25^2$	0.8216 (0.8333)	0.1784 (0.1667)	0.5559 (0.5556)	0.4441 (0.4444)	0.3603 (0.3571)	0.6397 (0.6429)	0.2416 (0.2381)	0.7584 (0.7619)	0.1704 (0.1667)	0.8296 (0.8333)
$[0.10^2 + 0.20^2], 0.30^2$	0.8188 (0.8333)	0.1812 (0.1667)	0.5557 (0.5556)	0.4443 (0.4444)	0.3607 (0.3571)	0.6393 (0.6429)	0.2422 (0.2381)	0.7578 (0.7619)	0.1711 (0.1667)	0.8289 (0.8333)
$[0.10^2 + 0.20^2], 0.35^2$	0.8157 (0.8333)	0.1843 (0.1667)	0.5555 (0.5556)	0.4445 (0.4444)	0.3611 (0.3571)	0.6389 (0.6429)	0.2428 (0.2381)	0.7572 (0.7619)	0.1718 (0.1667)	0.8282 (0.8333)
$[0.10^2 + 0.20^2], 0.40^2$	0.8124 (0.8333)	0.1876 (0.1667)	0.5552 (0.5556)	0.4448 (0.4444)	0.3613 (0.3571)	0.6387 (0.6429)	0.2434 (0.2381)	0.7566 (0.7619)	0.1725 (0.1667)	0.8275 (0.8333)

Table 4.11: Estimates of noise shares from the model with $n = 2$ phases

This table reports the mean estimates of noise shares in phases $(1 \cup 2)$, and 3 $(NS_{1\cup2}, NS_3)$ from the VAR model with two phases $(n = 2)$. Phase $(1 \cup 2)$ corresponds to joint phases 1 and 2 in the model with $n = 3$. The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true noise shares are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phase $(1 \cup 2)$ are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. The simulated returns are: $r_{i,t} = w_{i,t} + s_{i,t} - s_{i-1,t}$, where $r_{i,t}$ is the observed return in phase i of the 24-hour period t, $w_{i,t} \sim N(0, \sigma^2_{w_i})$ is the efficient price innovation, $s_{i,t} \sim N(0, \sigma^2_{s_i})$ is the pricing error. I estimate variances of information and noise ($Var(w_{i,t}) = \sigma^2_{w_i}$ and $Var(s_{i,t}) = \sigma^2_{s_i}$) following the formulas in the main text. Noise share in phase $(1\cup2)$ is $NS_{1\cup2} = \frac{Var(s_{2,t})}{Var(s_{2,t})+Var(s_{3,t})}$. Noise share in phase 3 is $NS_3 = \frac{Var(s_{3,t})}{Var(s_{2,t})+Var(s_{3,t})}$.

Results from VAR model with n=2 phases in the 24-hour period										
Noise variance, $\sigma^2_{s_1}, \sigma^2_{s_2}, \sigma^2_{s_3}$	Information variances, $[\sigma^2_{w_1} + \sigma^2_{w_2}]$, $\sigma^2_{w_3}$									
	$[0.10^2 + 0.20^2]$	0.10^2	$[0.10^2 + 0.20^2]$	0.20^2	$[0.10^2 + 0.20^2]$	0.30^2	$[0.10^2 + 0.20^2]$	0.40^2	$[0.10^2 + 0.20^2]$	0.50^2
$[0.10^2 + 0.20^2], 0.20^2$	0.4951 (0.5556)	0.5049 (0.4444)	0.4969 (0.5556)	0.5031 (0.4444)	0.5022 (0.5556)	0.4978 (0.4444)	0.5146 (0.5556)	0.4854 (0.4444)	0.5315 (0.5556)	0.4675 (0.4444)
$[0.10^2 + 0.20^2], 0.25^2$	0.3848 (0.4444)	0.6152 (0.5556)	0.3849 (0.4444)	0.6151 (0.5556)	0.3863 (0.4444)	0.6137 (0.5556)	0.3911 (0.4444)	0.6089 (0.5556)	0.4027 (0.4444)	0.5973 (0.5556)
$[0.10^2 + 0.20^2], 0.30^2$	0.3025 (0.3571)	0.6975 (0.6429)	0.3019 (0.3571)	0.6981 (0.6429)	0.3016 (0.3571)	0.6984 (0.6429)	0.3027 (0.3571)	0.6973 (0.6429)	0.3069 (0.3571)	0.6931 (0.6429)
$[0.10^2 + 0.20^2], 0.35^2$	0.2416 (0.2899)	0.7584 (0.7101)	0.2406 (0.2899)	0.7594 (0.7101)	0.2396 (0.2899)	0.7604 (0.7101)	0.2392 (0.2899)	0.7608 (0.7101)	0.2403 (0.2899)	0.7597 (0.7101)
$[0.10^2 + 0.20^2], 0.40^2$	0.1960 (0.2381)	0.8040 (0.7619)	0.1950 (0.2381)	0.8050 (0.7619)	0.1937 (0.2381)	0.8063 (0.7619)	0.1926 (0.2381)	0.8074 (0.7619)	0.1925 (0.2381)	0.8075 (0.7619)

Table 4.12: Estimates of information-to-noise ratios from the model with $n = 2$ phases

This table reports the mean estimates of information-to-noise ratios in phases $(1 \cup 2)$, and 3 $(IN_{1\cup2}, IN_3)$ from the VAR model with two phases $(n = 2)$. Phase $(1 \cup 2)$ corresponds to joint phases 1 and 2 in the model with $n = 3$. The data are from Monte Carlo simulations (1,000 samples with 1,000 observations per sample). The true information-to-noise ratios are in parentheses below the estimated values. The horizontal dimension corresponds to different values of $\sigma^2_{w_3}$ (variance of information in phase 3), the vertical dimension corresponds to different values of $\sigma^2_{s_3}$ (variance of noise in phase 3). The fixed parameters for phase $(1 \cup 2)$ are: $\sigma_{w_1} = 0.1$, $\sigma_{s_1} = 0.1$, $\sigma_{w_2} = 0.2$, $\sigma_{s_2} = 0.2$. The simulated returns are: $r_{i,t} = w_{i,t} + s_{i,t} - s_{i-1,t}$, where $r_{i,t}$ is the observed return in phase i of the 24-hour period t, $w_{i,t} \sim N(0, \sigma^2_{w_i})$ is the efficient price innovation, $s_{i,t} \sim N(0, \sigma^2_{s_i})$ is the pricing error. I estimate variances of information and noise ($Var(w_{i,t}) = \sigma^2_{w_i}$ and $Var(s_{i,t}) = \sigma^2_{s_i}$) following the formulas in the main text. Information-to-noise ratio in phase $(1 \cup 2)$ is $IN_{1\cup2} = \frac{Var(w_{1,t})+Var(w_{2,t})}{Var(w_{1,t})+Var(w_{2,t})+Var(s_{1,t})+Var(s_{2,t})}$. Information-to-noise ratio in phase 3 is $IN_3 = \frac{Var(w_{3,t})}{Var(w_{3,t})+Var(s_{3,t})}$.

Results from VAR model with n=2 phases in the 24-hour period										
Noise variance, $\sigma^2_{s_1}, \sigma^2_{s_2}, \sigma^2_{s_3}$	Information variances, $[\sigma^2_{w_1} + \sigma^2_{w_2}]$, $\sigma^2_{w_3}$									
	$[0.10^2 + 0.20^2]$	0.10^2	$[0.10^2 + 0.20^2]$	0.20^2	$[0.10^2 + 0.20^2]$	0.30^2	$[0.10^2 + 0.20^2]$	0.40^2	$[0.10^2 + 0.20^2]$	0.50^2
$[0.10^2 + 0.20^2], 0.20^2$	0.5733 (0.5000)	0.2263 (0.2000)	0.5753 (0.5000)	0.5178 (0.5000)	0.5760 (0.5000)	0.7067 (0.6923)	0.5737 (0.5000)	0.8123 (0.8000)	0.5668 (0.5000)	0.8719 (0.8621)
$[0.10^2 + 0.20^2], 0.25^2$	0.5747 (0.5000)	0.1629 (0.1379)	0.5771 (0.5000)	0.4083 (0.3902)	0.5784 (0.5000)	0.6049 (0.5902)	0.5765 (0.5000)	0.7320 (0.7191)	0.5708 (0.5000)	0.8117 (0.8000)
$[0.10^2 + 0.20^2], 0.30^2$	0.5761 (0.5000)	0.1229 (0.1000)	0.5787 (0.5000)	0.3255 (0.3077)	0.5804 (0.5000)	0.5150 (0.5000)	0.5789 (0.5000)	0.6532 (0.6400)	0.5739 (0.5000)	0.7473 (0.7353)
$[0.10^2 + 0.20^2], 0.35^2$	0.5773 (0.5000)	0.0965 (0.0755)	0.5801 (0.5000)	0.2633 (0.2462)	0.5821 (0.5000)	0.4386 (0.4235)	0.5811 (0.5000)	0.5798 (0.5664)	0.5763 (0.5000)	0.6833 (0.6711)
$[0.10^2 + 0.20^2], 0.40^2$	0.5786 (0.5000)	0.0783 (0.0588)	0.5812 (0.5000)	0.2165 (0.2000)	0.5833 (0.5000)	0.3750 (0.3600)	0.5829 (0.5000)	0.5135 (0.5000)	0.5786 (0.5000)	0.6221 (0.6098)

Chapter 5

The rise in trading on close: Drivers and effects on price formation

Consistently buy an S&P 500 low-cost index fund. I think it's the thing that makes the most sense practically all of the time.

Warren Buffett.

5.1 Introduction

The last five minutes of trading have become the busiest time of the day for many stock markets. In Europe's major equity markets, as much as one-quarter to one-half of the entire day's trading is now executed at the market close. This is twice the level of a mere three years ago, with the US following similar trends of increasing trading on close.[1] Around the world, the share of trading at the close continues to grow year on year. Given that bursts of intense trading activity are associated with increased volatility and large price movements, the concentration of trading in a small window around the close can potentially harm the quality of closing prices. The more trading that occurs in this short window of time, the higher is the risk of large order imbalances that can create temporary price swings and distort closing prices. Given the importance of closing prices as benchmarks

[1]For example, a recent report from the French market regulator (AMF) estimates that in 2019, closing auctions account for as much as 44% of daily volume in Spain, 41% in France, 40% in the UK, 36% in Germany, 32% in Netherlands, and 24% in Italy (see AMF, 2019). The same report estimates that in the US trading at the close represents around 12%–14% of daily volume, and that this share is rapidly growing.

in many contracts, the tendency towards increasing volumes on close could potentially harm the market's role in providing accurate prices.[2] As the French market regulator AMF (2019) put it: "The associated risks are a deterioration in price formation and liquidity during trading sessions, not to mention the operational vulnerabilities at the end of the day, given the volumes concentrated in the closing auction."

This chapter investigates the drivers and potential harm caused by the trend towards trading on close. What is driving the strong appetite to trade on close? Is the tendency harming closing prices by making them more prone to temporary price pressures? Or is the concentration of trading making closing prices more informative? How is the continuous trading session impacted by trading activity shifting away? Has the rest of the day become "meaningless" for price discovery, as suggested by some market participants?[3]

To answer these questions, and ensure the results are not specific to a particular type of closing mechanism, I empirically examine two different markets representing the two main closing mechanisms. As a representative of markets that have an "on-close" trading facility (meaning order types that can be placed during the day to be matched at the closing price), I use the US markets. And as a representative of the single-price batch auction closing mechanism (in which continuous trading is halted for a period of time before a batch auction that sets the closing price), I use the Australian equity market.[4]

I find that index investing is by far the most important driver of trading on close. The empirical tests rely on discontinuities in the activity of index funds around index inclusion thresholds. For example, in Australia, index funds invest almost ten times more in a stock just within the S&P ASX 300 index than in a stock just outside of this size-based index. Because index funds seek to minimize their tracking error against the index based on closing prices, they are incentivized to trade on close. Furthermore, creation and redemption of ETF units typically

[2]For example, Reuters (August 16, 2019): "The growing popularity of passive and index-tracking funds and tougher regulations are driving the shift [to trading on close], raising concerns about big price swings and possible disruption to price discovery".

[3]Reuters (August 16, 2019): "The rest of the day is meaningless and you can't see flow" (quote attributed to Andrea Vismara, Chief Executive Officer at Italian boutique investment bank Equita).

[4]Going forward, "trading on close" for Australian markets refers to dollar volume executed in ASX closing call auction, as a percentage of total ASX daily dollar volume. For the US markets, "trading on close" refers to dollar volume executed between 3:50 pm and 4:05 pm, as a percentage of total daily dollar volume.

occurs after the daily net asset value of the ETF is established based on closing prices. The growth in index investing, including ETFs, therefore leads to increased trading on close.

I also find that changes to market structure contribute to the rise of trading on close. I find some evidence of increased trading on close in stocks with more HFT activity after the introduction of fast data feeds that increased the speed advantage of HFTs. This finding is in line with the hypothesis that market participants seek auction trading facilities to avoid being picked off by faster traders during the continuous trading session. I do not find evidence that dark trading and block trading restrictions have a causal impact on trading on close.

To examine how the concentration of trading at the close is impacting price discovery and noise in markets, I use a novel variance decomposition method in each phase of the trading day (open, continuous trading session, and close), which separates the information and noise components of returns. The decomposition (described in detail in Chapter 4) allows me to measure the "information shares" and "noise shares" of each phase and examine how they change with the shifts in trading volume. The results are consistent across the US and Australian markets: closing volumes have increased over the past decade, but the information shares of the closing price have decreased. These results suggest that the trading that migrates to the close is relatively uninformed, consistent with much of it being attributed to index investing.

In fact, the price discovery analysis reveals that the closing phase of the market has never really contributed much to price discovery, beyond the price formation that happens during the continuous trading phase in the lead up to the close. I estimate that the average information share of the closing phase is 2.89% in the US and 0.85% in Australia. These estimates are for the year 2018, averaged across 200 representative stocks in each market.[5] In contrast, the major contributor to price discovery is the continuous trading phase. This phase accounts for an average of 76.07% of the price discovery in the US, and 60.85% in Australia. Despite trading activity shifting to the close, the continuous trading phase continues to play a dominant role for price discovery. The remainder of price discovery occurs in the opening auctions, which impound a substantial amount of overnight information.

[5]The Australian sample covers S&P ASX 200 constituents. After filters, the number of stocks is in the range of 205–214 names in each year during the period 2002–2018. The US sample covers 241 stocks stratified by market cap, as described in Section 2.4 of Chapter 2. After filters, the number of stocks is in the range of 185–230 names in each year during the period 2002–2018.

The price discovery method developed in Chapter 4 of this thesis also allows me to quantify the noise in prices in each phase. This analysis addresses the concerns that increasing volumes at the close can lead to large temporary dislocations in closing prices, or in other words, a substantial increase in the noise in closing prices. I find little support for such concerns. The results suggest that so far the closing mechanisms in both markets are able to digest the increasing trading volume with little or no negative side effects on price quality. The estimates show that the close accounts for about 23.46% of the total noise in the three phases of the day in the US and 18.19% in Australia.[6] Thus, the "noise-to-signal" ratio of the close is considerably higher than for the other phases of trading, consistent with the notion that little or no new information is impounded at the close that was not already reflected in pre-close prices. Interestingly, this "noise-to-signal" ratio of the close shows no signs of increasing through time as trading migrates to the close.

I implement a number of robustness tests to validate the price discovery results. Simple proxies of price informativeness (e.g., the proportion of zero returns in each phase, the realized variances, and the extent of overnight reversals) all point to the same conclusion: that the closing phase brings little or no new information that was not already in prices from continuous trading. This finding is somewhat surprising, given the prediction by Admati & Pfleiderer (1988) that prices are more informative in periods of concentrated liquidity, such as the closing auction.

This chapter's findings imply that closing mechanisms play an important role in pooling liquidity, but not in information production or price discovery. Therefore, regulators should consider the importance of continuous trading session for price discovery when they assess proposals of shorter trading hours in some markets (e.g., Europe). The findings also suggest that different closing mechanisms (on-close facilities vs closing call auctions) produce rather similar outcomes with respect to closing price informativeness.

This chapter proceeds as follows. Section 5.2 reviews relevant literature and develops hypotheses, Section 5.3 describes the institutional setting, and Section 5.4 presents the data. Section 5.5 examines the drivers of trading on close, and Section

[6]Noise estimates should be interpreted as "point in time" quantities. The noise share of the closing phase is therefore the proportion of noise variance of the closing price, relative to the total of noise variances of the opening price, pre-close price, and closing price.

5.6 presents the price discovery analysis of trading on close. Section 5.7 concludes the chapter.

5.2 Literature review and hypotheses

The early analysis of how equilibrium prices are formed in response to information dates back to Walras's (1874) field observations of the tâtonnement process on Paris Bourse. At the heart of this process lies gradual revelation of the aggregate demand and supply schedules as traders express their willingness to transact at certain prices. These prices then converge to a single equilibrium price through the auction process. Theoretical studies of opening and closing auctions show that the concentration of liquidity at a single point in time can be informationally efficient. Admati & Pfleiderer (1988) and Pagano (1989) argue that liquidity can become concentrated in single periods endogenously, via the network effect of "liquidity begetting liquidity". Vayanos (2007) shows that a single auction serves as a commitment mechanism to trade in larger size than through slicing orders thinly during the continuous session. To the extent that trading is costly, the commitment to a single trade is welfare-enhancing. For example, Copeland & Galai (1983) model how the time limit on batch auctions puts a cap on adverse selection costs to traders using limit orders.

Broadly, two types of closing mechanisms exist: single-price closing call auctions (like those of the Australian Stock Exchange, London Stock Exchange and Deutsche Börse) and on-close facilities (like those of the Toronto Stock Exchange, New York Stock Exchange and NASDAQ). Smaller, less liquid markets calculate the closing price from the last minutes of continuous trading. However, as markets evolve, they tend to adopt either a closing call auction or an on-close facility to form the closing price. Cordi, Foley, & Putnins (2015) find that adopting a closing mechanism (as opposed to relying on last trade prices of the day) is beneficial to market quality, with greatest benefits arising from randomized closing times, volatility extensions, no-order-altering policy in the pre-close period, and non-displayed indicative closing prices.

A number of empirical studies examine the effects of closing mechanisms on liquidity and price discovery: Pagano & Schwartz (2003) study Euronext Paris, Aitken,

Comerton-Forde & Frino (2005) — Australian Stock Exchange, Battig & Chelley-Steeley (2010) — London Stock Exchange, Pagano, Peng, & Schwartz (2013) — NASDAQ US. These papers show that closing mechanisms typically improve market quality and reduce the potential for market manipulation, especially when closing time randomization and volatility extensions are implemented as part of the closing mechanism (Comerton-Forde & Rydge, 2006; Kandel, Rindi, & Bosetti, 2012; Comerton-Forde & Putnins, 2013; Cordi, Foley, & Putnins, 2015). Given that multiple layers of safeguards have been implemented in the closing mechanisms in advanced markets like the ASX or NASDAQ, instances of closing price manipulation have become increasingly rare (Cordi, Foley, & Putnins, 2015).

Even in the absence of closing price manipulation, noise in the closing prices might make these prices less informative. The focus in this chapter is on information vs noise content in closing prices. The chapter examines the period 2002–2018, when a greater share of daily volume has shifted to the close of the market.

5.2.1 Literature on closing auctions

Trading on close has been in the spotlight of major industry publications in the past year. The Financial Times wrote in August 2018: "An increased concentration of volumes from 3:30 to 4 pm is causing concern." In September 2018, The Trade published an analysis piece by ITG, titled "Liquidity is for closers". However, no academic papers have yet investigated what drives this shift in trading volumes. Even more importantly, no formal analysis has considered whether closing prices have become more informative as a result. I address these questions in the present chapter.

This study links several strands of literature, namely studies on passive investing, intraday volume patterns, and market closing mechanisms. The inquiry into closing volume builds on Appel et al.'s (2016) regression discontinuity design to infer the effects of passive investing on closing volumes. It also speaks to the literature on the effects of passive ownership (e.g., Chinco & Fos, 2019; Ben-David, Franzoni, & Moussawi, 2018). Finally, it relies on the studies of closing mechanism design, including Cordi, Foley, & Putnins (2015), Comerton-Forde & Putnins (2013), and Pagano & Schwartz (2003).

The price discovery methodology in this study estimates information shares and noise shares in sequential trading, as described in Chapter 4. Unlike previous studies, this method does not assume that all prices are equally noisy, and instead derives noise components explicitly. Related work on intraday and overnight variances includes French & Roll (1986) and Cushing & Madhavan (2000). This study is different in that it does not treat return variances as representing efficient price innovations only. Instead, price innovations are isolated after accounting for transitory shocks that revert to zero (i.e., noise).

5.2.2 Informativeness of auction prices and continuous trading prices

Medrano & Vives (2001) model the way information is incorporated into auction prices. They show theoretically that information flow accelerates towards the auction match time. Empirical evidence from Paris Bourse supports this view (Biais, Hilion, & Spatt, 1999). The results in Cao, Ghyles & Hatheway (2000), and experimental results in Schnitzlein (1996) and Biais & Pouget (2000) suggest that the crucial ingredient for improved auction price efficiency is the pre-auction order accumulation phase. For an in-depth literature review on the informational efficiency of auction prices, see Biais, Glosten, & Spatt (2005).

Previous studies show that price informativeness varies between the opening auction, intraday phase and closing auction. Amihud & Mendelson (1987) find NYSE opening prices to be noisier than closing prices. According to Amihud & Mendelson (1991) and Amihud, Mendelson, & Murgia (1990), this result reflects the accumulation of information overnight rather than inefficiency of the auction mechanism itself. The closing auction, on the other hand, occurs after an extended period of continuous trading, so it is fair to compare the closing price to the pre-close price in the continuous session. Using weighted price contributions (a methodology that does not separate information variances from noise), Bacidore & Lipson (2001) find the official close on NYSE to be more informative than either the last daily transactions or last daily quotes.

Auctions are not equally resilient in all stocks and market conditions. Ellul, Shin, & Tonks (2005) find that when traders can choose between the auction and off-exchange dealership system, the auction suffers from high failure rate, especially

in volatile markets, and in small-sized and medium stocks. Similarly, Theissen & Westheide (2017) emphasize the role of designated market makers (DMMs) in intraday call auctions. They show that designated market makers provide the greatest benefits to the market in illiquid stocks and during volatile periods.

5.2.3 Trading on close and passive investing

The shift from active to passive investing has raised concerns among regulators (Anadu et al., 2019), particularly with regard to potential risks to financial stability. Multiple academic studies have emerged to assess how passive investing affects market volatility, individual stocks' volatility, cross-stocks correlations, liquidity risk etc. Ben-David, Franzoni, & Moussawi (2018) find that ETF ownership increases individual stock volatility. Evidence from Chinco & Fos (2019) suggests that this increase in volatility is likely to be related to noise rather than information. Chinco & Fos (2019) show that rules-based trading by index funds generates liquidity shocks, which are economically large and statistically unpredictable. These liquidity shocks can be reflected in higher volumes on close and temporary price pressures.

Passive investing and the rise of ETFs are frequently mentioned in conjunction with increased trading on close (Financial Times, 2018). Both passive mutual funds and ETFs use closing prices to calculate their net asset values. This may generate increased trading on close when ETFs or passive mutual funds: (i) face net in- or out-flows on a given day, prompting them to buy or sell their holdings in the closing auction, or (ii) rebalance their portfolios simulataneously around the time of index rebalancing events (to minimize the tracking error relative to the index). Additionally, certain leveraged or smart beta ETFs might rebalance their holdings daily to reflect a pre-specified leverage or other target ratio.

Some markets and stocks have experienced greater increases in trading on close than others. For example, Tabb Forum (2019) reports especially fast recent growth in trading on close in Europe. The recent report by AMF (2019), the French financial markets regulator, identifies the rapid growth in passive investment as one of the key drivers in substantial concentration of volume at the end of the day (41% in CAC-40 stocks). Far from being a local phenomenon, however, the shift of trading volumes to the close, and the shift of investments to passive funds, have affected the global equity markets. The Economist (2019) special briefing on

financial markets claims that "stock markets are now run by computers, algorithms, and passive managers". To establish the causal link between passive investing and trading on close, one can exploit discontinuity in passive ownership around certain index thresholds. For example, Appel et al. (2016) and Heath et al. (2020) examine the effect of passive ownership on corporate governance by instrumenting passive ownership with discontinuity around Russell 1000 / Russell 2000 index inclusion. This chapter employs a similar approach to study the effect of passive ownership on closing auction trading. The first hypothesis is as follows:

Hypothesis 1. Trading on close is positively related to the extent of passive investing.

5.2.4 Trading on close and dark / block trading

Recent analysis from Tabb Forum (2019) suggests that MiFID II limitations on dark trading in Europe have prompted investors to make greater use of closing auctions. Johann, Putnins, Sagade, & Westheide (2019) analyze post-MiFID effects across stocks and find a substitution effect between different forms of quasi-dark trading. When limited in their ability to trade in dark pools, European investors shift order flow to block trading venues, periodic auctions and internalizing dealers. Gomber, Sagade, Theissen, Weber, & Westheide (2016) investigate the determinants of traders' order routing decisions between different trading mechanisms (continuous lit trading, auctions, dark pools, internalization platforms and OTC markets) in Europe. Relying on the premise that market participants simultaneously choose a trade size and a trading venue, they find that market shares of lit markets are relatively higher in conditions of high information asymmetry (e.g., around earning announcements), while market shares of OTC markets are relatively higher when the proportion of uninformed trading increases (e.g., around post-dividend dates). Another argument is that investors route to venues with lowest execution costs. Boehmer, Jennings, & Wei (2006) analyze the effects of the SEC requirement for publishing monthly execution quality reports and show that market participants care about the execution quality on different trading venues and adjust their routing decisions when new information about execution quality is published.

The microstructure theory recognizes that informed traders choose between different types of markets by weighting trading costs against execution probability. For

example, Pagano & Roell (1996) model the effect of dark trading on liquidity by considering four stylized types of markets: a transparent auction, batch auction, continuous auction, and a dealership. Their model predicts that liquidity providers offer narrower spreads in fully transparent markets, as pre-trade information about the order flow reduces adverse selection costs imposed by insiders. More recent theories (e.g., Ye, 2011; Zhu, 2014; Buti, Rindi & Werner, 2017) show that traders willing to forgo immediacy are more likely to trade in dark pools, allowing them to save on execution costs. Menkveld, Yueshen & Zhu (2017) show that investors rate trading venues along two dimensions: liquidity costs and immediacy. Venues with high execution probability (high immediacy) and low transactions costs (high liquidity) are at the top of the pecking order, and therefore these venues have lower market share in high adverse selection and high volatility conditions, as investors move down the pecking order in times of increased immediacy needs.

Overall, earlier studies suggest that different trading mechanisms (e.g., dark pools, upstairs block markets, auction facilities and lit markets) can be substitutes depending on their design features, such as minimum tick size, immediacy and anonymity. Therefore, I hypothesize that market participants' ability to execute dark and block trades influences their propensity to trade in the closing auction.

Hypothesis 2. Trading on close is negatively related to the ability to trade in the dark.

Hypothesis 3. Trading on close is negatively related to the ability to trade in blocks.

5.2.5 Trading on close and HFT

Another reason why traders might resort to closing auctions is to minimize the risk of being picked off by high-frequency traders. For example, Budish, Cramton, & Shim (2015) argue that continuous trading in the presence of fast traders (HFTs) disadvantages slower traders, hence the latter should respond by moving to call auctions. O'Hara (2014) argues that in fragmented markets, speed is crucial for effective trade execution, and O'Hara & Ye (2011) show that fragmented stocks have lower execution costs, but smaller trade sizes. O'Hara (2015) points out that fragmentation has given rise to electronic execution strategies and enabled HFTs to pick off slower traders by detecting their order routing patterns. Korajczyk &

Murphy (2018) and van Kervel & Menkveld (2019) trace the large institutional orders and find that early in the lifetime of an order, HFTs provide liquidity, but later switch and start trading in the same direction as the order. Goldstein, Kwan, & Philip (2018) study HFT activity on ASX and find that HFTs tend to provide liquidity on the "thick" size of the limit order book and demand liquidity on the "thin" size, thus exacerbating order imbalances. In these conditions, the closing auction, which offers a large pool of liquidity without the risk of being picked off by faster traders, provides an attractive alternative for institutional order flow.

Several theory models show that HFTs have incentives to prey on large liquidity-motivated orders (Brunnermeier Pedersen, 2005), as well as on large informed orders (Li, 2019; Yang & Zhu, 2019; Boulatov, Bernhardt, & Larionov, 2016). Li (2019) and Yang & Zhu (2019) argue that the optimal response of informed traders is to trade less aggressively by delaying order submissions or slicing orders more finely. Degryse, Tombeur, & Wuyts (2018) show that algorithmic trading is negatively related to dark trading, in line with order hiding behavior of informed traders. They also document substitution effects between hidden orders on lit exchanges and dark trading on off-exchange venues.

Therefore, increased trading speeds, fragmentation and algorithmic trading go hand in hand with lower trade sizes and higher signaling risk faced by large institutional traders. I hypothesize that in response to increased HFT activity, market participants resort to trading more in the closing auction.

Hypothesis 4. Trading on close increases with the extent of HFT activity.

5.2.6 Trading on close and price discovery

Given that the proportion of trading on close has increased substantially, it is of interest how this dynamic has affected information and noise components of the closing price. On one hand, the theoretical argument in Admati & Pfleiderer (1988) is that informed traders have an incentive to "pool" with the uninformed, and hence trade substantially in the closing auction. On the other hand, empirical analysis in Cushing & Madhavan (2000) suggests that closing prices reverse overnight, suggesting low information content on close.

The composition of market participants who trade in the closing auction matters for price discovery. For example, greater amount of trading by passive funds (including ETFs) can result in price inefficiencies, such as increased volatility, higher correlations, temporary price pressures and liquidity co-movements, as shown in Chinco & Fos (2019), Ben-David, Franzoni & Moussawi (2018), Hamm (2014), Israeli, Lee, & Sridharan (2017), Da & Shive (2017). This chapter proposes a potential channel through which ETF holdings can introduce noise in the stock prices: via price dislocations on close. Hence, I hypothesize that greater closing volumes are related to noisier closing prices:

Hypothesis 5. Greater proportion of dollar volumes executed on close is associated with higher noise share of the closing price.

5.3 Institutional setting

5.3.1 Trading on close in the US and Australian markets

I consider two main closing mechanisms in this study: (i) the single-price closing call auction (like those operated by the ASX, LSE and Deutsche Börse), and (ii) an on-close facility (like those operated by NASDAQ, NYSE and TSX). The key difference between these closing mechanisms is that (i) has a distinct end of day call auction period, while (ii) allows traders to enter on-close orders throughout the day.[7] To ensure that the findings are not driven by one specific closing mechanism, I investigate trading on close in two major markets (the US and Australia), which operate different mechanisms for market close (a closing call auction in Australia, and an on-close facility in the US).

The US on-close facilities are operated by two listing exchanges — NASDAQ and NYSE. The official closing price in a stock is set on an exchange that lists that stock. NYSE accepts market-on-close and limit-on-close orders from 6:30 am till 3:50 pm, and NASDAQ — from 4 am till 3:55 pm. During the period 3:50 pm–4 pm (3:55 pm–4 pm), only imbalance-offsetting orders can be entered on NYSE (NASDAQ), and during that period NYSE (NASDAQ) publishes order imbalance information every 5 seconds (every 1 second). Appendix 5.1 illustrates the timeline

[7]See Cordi, Foley, & Putnins (2015) for a detailed review of different closing mechanisms and their characteristics.

of on-close facilities of NYSE and NASDAQ. Unlike the single-price closing call auction of ASX, on-close facilities of NYSE and NASDAQ operate in parallel with continuous markets. At 4 pm, the closing cross occurs, and continuous trading stops for the day.

In Australia, the ASX closing call auction occurs after the continuous trading session stops at 4 pm. Buy and sell orders are consolidated over the period 4 pm–4:10 pm, and during that time indicative order imbalances and indicative prices are available to all market participants. The closing cross occurs at 4:10 pm, with randomized closing time of up to + 60 seconds. The ASX matching algorithm calculates the volume-maximizing closing price between 4:10 pm and 4:12 pm.[8] Appendix 5.1 illustrates the timeline of the ASX single-price call auction.

5.3.2 The microstructure of Australian markets

I test Hypotheses 1–4 using Australian data, which allows me to exploit several natural experiments related to dark trading, block trading, and HFT activity. One advantage of this setting is the data availability on dark trading and block trading in the Australian market. Another advantage is that Australian markets have undergone a number of regulatory changes (e.g., a change to dark trading and block trading rules) during the sample period, which is helpful as an identification strategy in the regression analysis. Also, compared to the US, Australian markets are less complex, which makes the effects of market structure changes more tractable.

The Australian financial markets regulator, ASIC, implemented a number of market structure changes over the sample period considered in this chapter. In 2011, competition was introduced in exchange markets, and trading fragmented between the ASX and Chi-X Australia as the latter entered the market on October 25, 2011. On April 2, 2012, ASX launched the ultra-low latency ITCH trading protocol, which reportedly increased the speed of obtaining ASX trading information up to seven times. The ITCH protocol is used by institutional brokers who pay a monthly access fee. Information on the identities of ITCH subscribers is not available to the public. Goldstein, Kwan & Philip (2018) discuss the institutional details of HFT trading in Australia. They argue that speed-sensitive traders (HFTs) are able to

[8] If no single volume-maximizing price can be established, then the matching algorithm optimization problem considers the second principle — minimizing the order imbalance. The third principle is market pressure, and the fourth principle is last trade price. See Comerton-Forde & Rydge (2006) for a detailed analysis of the matching algorithms.

trade more aggressively after obtaining a speed advantage in the post-ITCH period. I use the ITCH introduction as an exogenous shock to HFT activity in the 2SLS regression analysis.

Dark trading in the Australian market can occur in ASX-operated dark pool (Centre Point), on Chi-X Australia (using several dark order types)[9], or in 15 broker-operated dark pools.[10] The broker dark pools (crossing systems) are accessible by institutional traders who are clients of a particular broker operating the dark pool. Additionally, there are three operators aggregating orders from different dark pools: ITG, Instinet and Liquidnet. Since 2017, ASIC requires crossing network operators to disclose which other crossing systems the client orders may be routed to, and whose orders may access a given crossing system. Clients can then opt out from interacting with the order flow from a particular crossing system or aggregator, as well as place restrictions on where their orders may be routed. Most of the crossing network transactions are client-to-client trades, with brokers acting as an agent.[11] Dark trading in broker-operated dark pools is available only to institutional traders. Payment for order flow is not allowed in Australia, and retail brokers route to exchanges rather than to broker-operated dark pools. Retail flow can interact with Chi-X dark orders (by executing against mid-, near-, or far- resting orders), but not with ASX Centre Point, as ASX requires brokers to connect to Centre Point to be able to interact with its dark flow.[12]

On March 27, 2013, ASIC introduced a minimum price improvement regulation, which requires that dark trades improve on lit quotes by at least one full tick, or else (if spread is tick-constrained) execute at mid-point of NBBO. Foley & Putnins (2015) show that dark trading decreased from 15.2% to 9.2% of dollar volume in

[9]Chi-X Australia operates an integrated order book, which means that dark order flow can interact with the lit limit order book. Chi-X participants can use "Chi-X Mid" dark order type, or "Far" and "Near" pegged orders to access dark liquidity on Chi-X.

[10]As of April 2019, ASIC-registered crossing systems (dark pools) are: Best Block Event, Citi Match, CLSA, Credit Suisse CrossFinder, DeutscheBank Super X, Goldman Sachs Sigma X, Instinet BLX, ITG POSIT, J.P. Morgan JPM-X, Liquidnet, Macquarie Securities MACB, Macquarie Securities MAQX, Morgan Stanley MS Pool, State One SOSL 1, and UBS PIN.

[11]Internalization of order flow, when brokers take the other side of the transaction (i.e., act as a principal), is not prohibited in Australia. However, ASIC Market Integrity Rules updates from 2017 require brokers to disclose whenever they are trading as a principal, and obtain a client's consent. In practice, this means that electronic crossing networks are not well suited for internalization, and are mostly used to facilitate client-to-client trades. Transactions in which a participant acts as a principal, or as partly a principal and partly an agent are marked with a special reference flag in trade reports.

[12]In theory, retail brokers can connect to Centre Point, but in practice, they do not have an incentive to do so, given that Centre Point charges higher trading fees than either ASX lit or Chi-X.

the two months following the regulation. There was also a clear substitution from two-sided to one-sided dark trading: after price improvement rules came into effect, the share of dark dollar volume executed at mid-quote increased from 46% to 81%. I use the price improvement regulation as an exogenous shock to dark trading in the 2SLS regressions.

The number of crossing systems has decreased gradually after the price improvement regulation, as some of the brokers found it unprofitable to operate one. For example, Foley & Putnins (2016) note the existence of 21 dark pools in Australia in 2013, while in April 2019, there were only 15 dark existing dark pools. The 2017 ASIC disclosure requirements (on crossing systems' routing of clients' order flow) led to further reduction in the number of dark pools. Overall, since the price improvement regulation, the composition of dark trading has changed: dark volumes shifted from broker-operated dark pools to exchange-operated dark pools, and from two-sided dark trading to midpoint dark trading. The proportion of dark trading has been increasing gradually, and now accounts for about 12.5% of total value traded (compared to about 9.2% immediately following the regulation).

With respect to block trading, ASIC divides equities into three tiers: the minimum block value for Tier 1 securities is $1 million, for Tier 2 – $500 thousand, and for Tier 3 – $200 thousand. The block trading tiers were introduced on May 27, 2013, and prior to this date the minimum block value for all securities was $1 million. ASIC publishes the additions and deletions from each tier on a quarterly basis, giving market participants two weeks' notice prior to the effective date. Securities are assigned to tiers based on average value traded, with more widely traded securities facing higher block trading thresholds. I use block tier assignment as an instrument for ability to trade in blocks in the 2SLS regressions.

5.4 Data

The Australian analysis of drivers of trading on close covers the sample of 881,667 stock-day observations during the period 2012–2018. This period coincides with the growth in trading on close, and has passive investing data available from ETF Global (the ETF Global data, which is needed to construct the passive ownership variable, starts in May 2012). I sample stocks that belong to the ASX All Ordinaries Index in any given year. There are 618 stocks on average on a

given day in the sample, covering a broad range of market capitalizations. The data are from Thomson Reuters Tick History, CRSP, ETF Global and SIRCA. The variable definitions are provided in Table 5.1.

Table 5.1: Variable definitions

Variable name	Variable definition
Australian data	
$PercentOnClose$	Fraction of stock's daily dollar volume on ASX executed in the ASX closing auction.
$PercentPassive$	Fraction of stock's market capitalization held by passive funds. The list of passive funds is in Table 5.2.
$Passive$	Dollar value of stock's market capitalization held by passive funds.
$MktCap$	Total market capitalization of a stock.
$Frag$	Fraction of daily lit dollar volume on Chi-X Australia, relative to total daily lit dollar volume on ASX and Chi-X combined.
$Depth$	Combined time-weighted depth at best bid and offer, in dollar terms.
$RelSpread$	Average daily relative spread, calculated as the difference between ask and bid, scaled by midpoint.
$TickToPrice$	Tick size divided by the closing price.
$HighLowVolat$	Daily stock volatility measure, computed as high minus low price, divided by the average of daily high and low prices.
$TradeSize$	Daily dollar volume on ASX divided by the number of trades.
$OTTR$	Number of ASX order entry, amend and cancel messages at the best quotes, divided by the ASX number of trades. Winsorized at 1% level. In ITCH regressions, OTTR is the number of total order entries, cancellations and amendments (at all levels of the order book), divided by the number of trades.
D_{ITCH}	A dummy variable taking the value of 1 after April 1, 2012, and 0 otherwise. Captures the trading speed upgrades following the introduction of the ITCH protocol by ASX.
D_{reg}	Price improvement regulation dummy. Takes the value of 1 after May 27, 2013, and 0 otherwise. The regulation requires that dark trades provide a full one tick price improvement, or half a tick for stocks with spread at one tick.
$D_{S\&P300}$	A dummy variable taking the value of 1, if a security is in S&P ASX 300 index on a given day, and 0 otherwise.
D_{Tier2}	A dummy variable taking the value of 1, if a security belongs to the block trading Tier 2 list after May 27, 2013, and 0, if the date is before May 27, 2013 or a security does not belong to Tier 2 list.
D_{Tier3}	A dummy variable taking the value of 1, if a security belongs to the block trading Tier 3 list after May 27, 2013, and 0, if the date is before May 27, 2013 or a security does not belong to Tier 3 list.
$PercentBlock$	Fraction of stock's ASX dollar volume executed in block trading.
$PercentCtPoint$	Fraction of stock's ASX dollar volume executed on Centre Point (ASX dark pool).
$PercentDark$	Fraction of stock's ASX dollar volume executed in dark pools.
$PercentTickConstr$	Fraction of time a stock is tick-constrained on a given trading day.
US data	
$PercentOnClose$	Share of dollar volume executed between 3:45 pm and 4:05 pm.
$PercenPassive$	Fraction of stock's market capitalization held by US-listed ETFs.
$PercentDark$	Fraction of stock's dollar volume trade-reported in FINRA TRF.
$OTTR$	Number of order entry, amend and cancel messages at the best quotes, divided by number of trades. Winsorized at 1% level.
$TradeSize$	Dollar volume between 10:00 am and 4 pm divided by number of trades over the same period.
$Frag$	Number of markets with non-zero dollar volume for a given stock.
$RelSpread$	Bid-ask spread divided by midpoint. Computed as daily average from intraday hourly data between 10:00 am and 4 pm, using the consolidated data feed.
$HighLowVolat$	Daily stock volatility measure, computed as high minus low price, divided by the average of daily high and low prices, using the consolidated data feed.

The US analysis of trading on close relies on 997,409 stock-day observations for the period 2000–2017. I do not run 2SLS regressions for the US market, and only use descriptive time series and cross-sectional results to compare against the Australian market. The sample starts in year 2000, given the US passive ownership variable is proxied by ETF ownership (obtained from Thomson Reuters linked with CRSP,

rather than from ETF Global). The sample ends in 2017 due to data limitations on passive ownership by ETFs.

The sampling approach for the US stocks seeks to ensure the composition of securities that are representative of today's market.[13] In Australia, this composition remains relatively unchanged between 2012 and 2018, and is well captured by the All Ordinaries index. In the US, the composition of stocks changed rather significantly, becoming more skewed towards larger and older companies. To remove the effects of composition changes (in terms of size distribution), I use stratified random sampling approach. This allows me to focus on how market structure (rather than the composition of stocks in the market) affects trading on close.

The percent passive variable from Table 5.1 deserves a separate explanation. For US stocks, the percent of passive ownership is proxied by the share of stock's market cap held by ETFs. Similarly to Easley, Michayluk, O'Hara, and Putnins (2020) and Ben-David et al. (2018), I use the data from CRSP to identify the ETFs, and from Thomson Reuters Fund Holdings database to obtain their quarterly holdings. I also adjust the data for several ETF share classes. The % passive estimates are based on US$ 1.48 trillion combined holdings of 413 ETFs as of the end of 2017.

Passive investing data for Australian stocks accounts for holdings by both listed (i.e., ETFs) and unlisted passive funds. I back out the passive holdings in ASX-listed stocks from (i) assets under management (AUM) of Australian passive funds (including ETFs) with exposure to Australian equities, (ii) AUM of US-listed ETFs with exposure to Australian equities (and their unlisted counterparts), and (iii) the weightings of each stock in the respective indices that these passive funds track. The data on US ETFs (ETF and unlisted fund AUM) are from ETF Global, the data on Australian passive funds is from SIRCA (assets under management) and Thomson Reuters Tick History (stock weights).[14] The list of passive funds is provided in Table 5.2.

The differences in market structure between the US and Australian markets also result in somewhat different approaches to measuring dark trading in these two markets. The dark trading variable for US markets is FINRA TRF volume, scaled by total daily volume. The FINRA TRF volume captures the sum of dark volume

[13]The US sample consists of 241 randomly selected US stocks, stratified by market capitalization. See the data section in Chapter 2 for a detailed description of the sampling approach.

[14]I reconstruct the weights from float-adjusted market capitalization, and identify the index composition by applying Thomson Reuters RIC chain expansion on index rebalance dates.

Table 5.2: List of passive funds with significant exposure to AU stocks

This table reports the list of passive funds used to calculate *PercentPassive* variable for the Australian market. The list contains Australian and US index funds with the most significant holdings in ASX-listed stocks. Funds are ranked from highest to lowest, based on AU$ AUM invested in Australian stocks. RIC identifiers are for listed versions of these index funds. AUM are the sum of holdings in listed and unlisted versions of these funds, as of July 2019.

RIC	Index	AUM in ASX-listed stocks, AU$ mln.
Australian funds holding ASX-listed stocks		
VAS	S&P/ASX300 (Vanguard)	15,788.10
-	S&P/ASX300 (iShares, unlisted only)	3,438.12
STW	S&P/ASX 200 (SPDR)	3,856.92
IOZ	S&P/ASX 200 (iShares)	1,289.85
A200	S&P/ASX 200 (Beta Shares)	715.24
US funds holding ASX-listed stocks		
EFA	MSCI EAFE NR USD	4,683.31
VEA	FTSE Developed ex North America NR USD	3,736.26
IEFA	MSCI EAFE IMI NR USD	3,419.95
SCHF	FTSE Dv Ex US NR USD	1,398.86
VEU	FTSE AW Ex US TR USD	994.02
FV	Dorsey Wright Focus Five USD	583.51
EFAV	MSCI EAFE Minimum Volatility NR USD	493.64
SCZ	MSCI EAFE Small Cap NR USD	474.80
VXUS	FTSE Global All Cap ex US TR USD	461.63
DBEF	MSCI EAFE 100% Hedged NR USD	407.79
IXUS	MSCI ACWI Ex USA IMI NR USD	406.53
EFV	MSCI EAFE Value NR USD	392.06
GUNR	Morningstar Gbl Upstream Ntl Res TR USD	366.14
IDV	DJ EPAC Select Dividend TR USD	320.08
GDXJ	MV Global Junior Gold Miners NR USD	309.34

from three active TRFs (trade reporting facilities): FINRA / NASDAQ TRF Carteret, FINRA / NASDAQ TRF Chicago and FINRA / NYSE TRF. The data are from Thomson Reuters Tick History Elektron Time Series.

The Centre Point variable for Australian markets is dollar volume from Centre Point (ASX-operated dark pool), scaled by total daily dollar volume. The dark trading variable includes volume from all dark pools (as in Foley & Putnins, 2016), and is only available for the time period around price improvement regulation.

5.5 Drivers of trading on close

5.5.1 Descriptive analysis of trading on close

A median US (Australian) stock in the sample has market cap of US$ 584 million (US $287 million), a relative spread of 22 bps (57 bps), and trade size of US$ 5,199 (US $1,488). The split of percent dollar volume on close by market cap quintiles reveals, as expected, a positive cross-sectional relation between these variables. The split by passive holdings, similarly, shows that stocks with more passive holdings trade more on close. Descriptive statistics are provided in Table 5.3.

Table 5.3: Descriptive statistics

This table reports descriptive statistics for Australian (Panel A) and US (Panel B) markets. All variables are at stock-day level. Variable definitions are in Table 5.1. The split by *PercentPassive* quintiles is conditional on non-zero *PercentPassive*. Panel A covers the the sample of 630 ASX All Ordinaries stocks during the period of May 2012–December 2018. Panel B covers 241 US stocks during the period January 2010–December 2017.

Panel A. Australian data

	Mean	StDev	25th pctl	50th pctl	75th pctl
PercentOnClose	0.0955	0.1203	0.0031	0.0633	0.1408
PercentPassive	0.0093	0.0388	0.0000	0.0038	0.0106
PercentCtPoint	0.0524	0.0855	0.0000	0.0200	0.0728
OTTR	2.9109	1.9533	1.9889	2.5412	3.2561
Frag	0.1562	0.1670	0.0000	0.1119	0.2636
MktCap, AU$ mln	3,063.53	10,865.38	162.32	428.79	1,686.25
RelSpread	0.0143	0.0290	0.0027	0.0057	0.0164
HighLowVolat	0.0314	0.0302	0.0146	0.0241	0.0395
TradeSize, AU$	5,463.04	316,359.27	1,208.02	2,215.49	4,121.39
PercentOnClose cross-sectional means by quintile					
	Q1	Q2	Q3	Q4	Q5
By MktCap quintile	0.0369	0.0475	0.0638	0.1015	0.1583
By PercentPassive quintile	0.1249	0.1367	0.1372	0.1445	0.1638

Panel B. US data

	Mean	StDev	25th pctl	50th pctl	75th pctl
PercentOnClose	0.1723	0.8511	0.0639	0.1245	0.2112
PercentPassive	0.0188	0.0247	0.0000	0.0079	0.0301
PercentCtPoint	0.2877	0.1780	0.0000	0.2448	0.3711
OTTR	17.7146	35.4015	4.7419	7.9259	14.3976
Frag	5.4102	3.0346	3.0000	5.0000	8.0000
MktCap, US$ mln	5,972.4658	2,4847.5139	130.3511	584.8128	2479.4557
RelSpread	0.0073	0.0176	0.0009	0.0022	0.0064
HighLowVolat	0.0406	0.0399	0.0181	0.0296	0.0495
TradeSize, US$	11,552.40	28,731.33	2,289.60	5,199.64	11,673.63
PercentOnClose cross-sectional means by quintile					
	Q1	Q2	Q3	Q4	Q5
By MktCap quintile	0.1234	0.2013	0.1910	0.1795	0.1659
By PercentPassive quintile	0.1299	0.1350	0.1507	0.1959	0.2516

I investigate multivariate relations between trading on close and explanatory variables in the time series regressions. As part of exploratory time series analysis, I regress first differences of percent on close on first differences of explanatory variables: percent dark trading, order-to-trade ratios, fragmentation, relative spread and volatility (see Appendix 5.2 for results). For an average stock, the percent on close tends to increase compared to the day before, when average passive holdings increase, volume in the ASX dark pool goes down, fragmentation increases, relative spread on ASX narrows, and market volatility increases.

These time series relations are broadly in line with the hypotheses: trading on close increases with passive ownership (due to passive funds that trade on close) and decreases with the proportion of dark trading (indicative of a substitution effect). The positive relation between trading on close and fragmentation corroborates the hypothesis that market participants trade more on close when the risk of being picked off by HFTs increases.

In the US data, I do not find significant results for dark trading and volatility, while the relations with fragmentation and relative spread are similar to the Australian markets. Note that the market structure of dark trading is different between Australia and the US: in Australia, investors are restricted by the price improvement regulation (they must use the same tick schedule as in lit markets), while in the US, dark trading is is not restricted trading (apart from several groups of securities during the Tick Size Pilot).

I also investigate which stock characteristics matter for trading on close, by using cross-sectional regressions of trading on close on explanatory variables (see Appendix 5.2 for results). I find that stocks that trade more on close tend to have higher passive holdings, trade more on Centre Point, are more fragmented, and have wider intraday spreads. The cross-sectional results for passive holdings, fragmentation, and spreads are consistent between Australian and US markets.

In the cross-section, passive holdings are strongly correlated with trading on close, after controlling for market cap. Stocks with high dark volume tend to also have high closing volumes, an effect consistent with the substitution between dark and block trading within a particular stock. Finally, more fragmented stocks trade more on close, consistent with "liquidity pooling" argument: traders in those stocks seek to minimize the risk of being picked off by resorting to large liquidity pools that do not give an advantage to faster traders.

5.5.2 2SLS analysis of trading on close

I test the causal relations between trading in the closing auction and four hypothesized drivers: passive investing, dark trading, block trading, and HFT activity. I use the Australian markets for this analysis, because regulatory changes in Australia allow for causal inference in a relatively tractable market. I use two stage least squares (2SLS) regressions to test for causal effects.

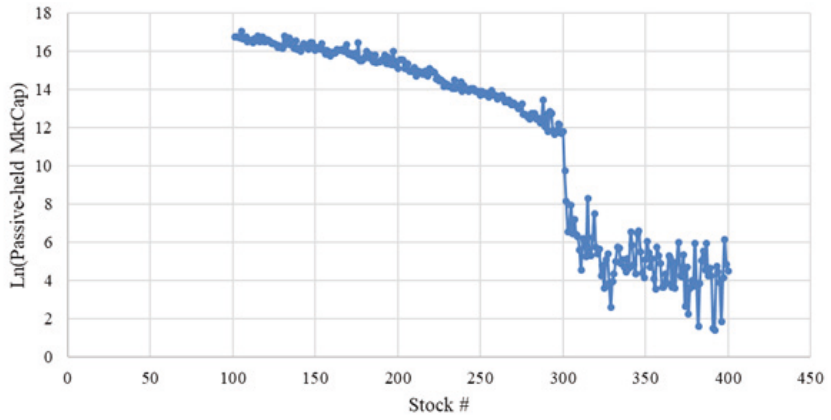
5.5.2.1 Passive investing

Testing Hypothesis 1, I investigate whether the degree of passive ownership of Australian stocks causally affects the extent of trading in the ASX closing call auction. First, I extract the variation in passive ownership that is not related to trading on close (either directly or through other variables, such as market cap). Second, I ask whether this exogenous variation in passive ownership can explain the variation in trading on close.

The identification strategy is similar to Appel, Gormley, & Keim (2016). There might be common factors, such as market capitalization, that affect both trading on close and ownership by passive funds. Therefore, I choose an instrumental variable (index membership in S&P ASX 300) that does not affect market participants' decision to trade in the closing auction, other than through the passive ownership channel. I compare the passive dollars invested in the Australian stocks ranked 201–300 (*in* S&P ASX 300) to those stocks number 301–400 (*outside* S&P ASX 300), controlling for market capitalization, stock- and time-fixed effects.

S&P ASX 300 is the most widely followed benchmark among Australian passive funds (see Table 5.2): AU $19.2 billion assets under management in Australian passive funds track S&P ASX300. In comparison, AU $5.86 billion tracks S&P ASX 200. Therefore, the chosen instrument (a dummy variable for the S&P ASX 300 index membership) is strongly related to passive ownership. At the same time, there is no reverse causality from trading on close to passive ownership, as S&P index composition follows a predefined methodology based on market cap weights adjusted for free float. Furthermore, S&P ASX 300 portfolios are not subject to index arbitrage activity to the same extent as S&P ASX 200, hence trading in stocks number 201–300 is driven by passive funds rather than by index arbitrageurs.

Panel A. Passive ownership by stock rank in S&P ASX 300



Panel B. Percent dollar volume on close, split by % passive ownership

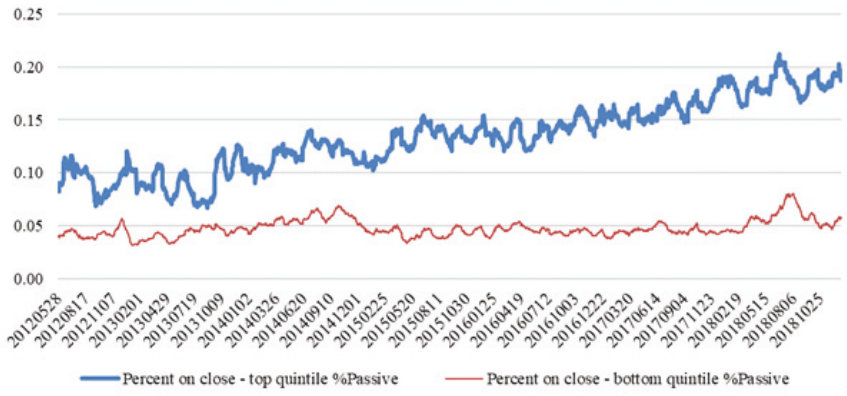


Figure 5.1: Trading on close and passive ownership (AU)

Panel A plots *LogPassive* against the stock rank (by market cap) in S&P ASX300 index. Panel B plots the time series of *PercentOnClose* split by *Passive*. The sample covers ASX All Ordinaries stocks for the period January 1, 2012–December 31, 2018.

I control for time and stock fixed effects in both stages of the 2SLS framework and include market capitalization controls with squared and cubic terms. The baseline regression specification is as follows:

Stage 1:

$$\begin{aligned} LogPassive_{it} = b_0 + b_1 D_{S\&P300_{it}} + b_2 LogMktCap_{it} + b_3[LogMktCap_{it}]^2 + \\ + b_4[LogMktCap_{it}]^3 + \mu_i + \tau_t + \epsilon_{it} \end{aligned} \tag{5.1}$$

Stage 2:

$$\begin{aligned} PercentOnClose_{it} = {} & k_0 + k_1 \widehat{LogPassive}_{it} \\ & + k_2 LogMktCap_{it} + k_3 [LogMktCap_{it}]^2 + k_4 [LogMktCap_{it}]^3 \\ & + k_5 CtPoint_{it} + k_6 Frag_{it} + k_7 OTTR_{it} + k_8 HighLowVolat_{it} \\ & + k_9 RelSpread_{it} + \mu_i + \tau_t + e_{it} \end{aligned} \tag{5.2}$$

Where the unit of observation is stock-day, and i and t are stock and day subscripts respectively. $LogPassive_{it}$ is the natural logarithm of passive holdings (in AU$) in stock i on day t, $D_{S\&P300_{it}}$ is a dummy variable that takes the value of one, if stock i is in S&P ASX 300 index on day t, and 0 otherwise, $LogMktCap_{it}$ is the natural logarithm of market capitalization (in AU$), $\widehat{LogPassive}_{it}$ is the fitted value of passive ownership from the first stage, $PercentCtPoint_{it}$ is the proportion of dollar volume executed in the ASX dark pool Centre Point, $Frag_{it}$ is the measure of fragmentation, computed as the percent dollar volume on Chi-X Australia (a second stock exchange operating alongside ASX), $OTTR_{it}$ is the order-to-trade ratio on ASX, $HighLowVolat_{it}$ is the volatility measure computed as daily high-low range scaled by high-low average. For detailed definitions of all variables, see Table 5.1.

The first stage regression results (see model (2) in Table 5.4) suggest that after controlling for market capitalization, stocks 201–300 (*in* S&P ASX 300) have ten times higher passive holdings than stocks 301–400 (*outside* S&P ASX 300). This result is graphically illustrated in Figure 5.1: there is a sharp discontinuity in passive holdings around stock number 300, hence the S&P ASX 300 dummy is a strong instrument for passive ownership. The results are similar when using 50 stocks (rather than 100) around stock 300 (see model (1) in Table 5.4) .

The second stage regression results (see Table 5.5 below and Table 5.15 in Appendix 5.3) show that percent dollar volume on close is indeed higher for stocks with higher passive ownership. The effect is statistically significant. To evaluate the economic magnitude, consider an average stock (mean passive holdings of 0.93% of market cap): if its passive holdings increase to 1.86%, keeping market cap constant, its fraction of dollar volume on close increases by 0.17% (according to model (2) in Table 5.5). As illustrated in Figure 5.1, the top 20% stocks in terms of passive ownership experienced a visible run-up in trading on close, while

Table 5.4: Stage 1 – 2SLS regression results for passive investing (AU)

This table reports stage 1 results from 2SLS. The dependent variable is $LogPassive_{it}$. The instrumental variable is dummy for inclusion in S&P ASX 300 index ($D_{S\&P300}$). All variables are at stock-day level. Variable definitions are in Table 5.1. The sample period is May 2012 – December 2018. In model (1), the sample stocks are within 50-stocks band on each side of stock number 300 by market capitalization. In model (2), they are within 100-stocks band. T-statistics are reported in parentheses.***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Stage 1 regression results		
	$LogPassive$ (1)	$LogPassive$ (2)
$D_{S\&P300}$	9.2654***	9.3807***
	(10.6934)	(21.7450)
$LogMktCap$	57.0069	10.2796
	(1.2432)	(0.2920)
$LogMktCap^2$	-3.1113	-0.8392
	(-1.3702)	(-0.4842)
$LogMktCap^3$	0.0572	0.0208
	(1.5388)	(0.7345)
Adjusted R^2	82%	73%
Clustered std. errors	Stock & Date	Stock & Date
Fixed effects	Stock & Date	STock & Date

the bottom 20% stocks had almost no change in trading on close over the period 2012–2018.

5.5.2.2 Block and dark trading

Testing hypotheses 2 and 3, I investigate the effect of regulatory changes in block and dark trading on the percent of trading on close. Starting from March 27, 2013, ASIC lowered the hurdles for block executions, while simultaneously increasing the hurdles for dark trading. After the change, traders face lower block trading thresholds: the minimum block size decreased from AU$ 1 million to AU$ 200 thousand for Tier 3 securities, and to AU$ 500 thousand for Tier 2 securities. In terms of dark trading below the block size, from March 27, 2013, the price improvement regulation requires traders to execute dark trades with at least one full tick price improvement in bid or offer, or else trade at midpoint. Because the regulations of block and dark trading were introduced simultaneously, I exploit cross-sectional heterogeneity in the effects in how this regulation affected (i) the extent of dark trading, and (ii) the extent of block trading. I then run 2SLS regressions with two exogenous variables, instrumenting (i) and (ii) separately.

Table 5.5: Stage 2 – 2SLS regression results for passive investing (AU)

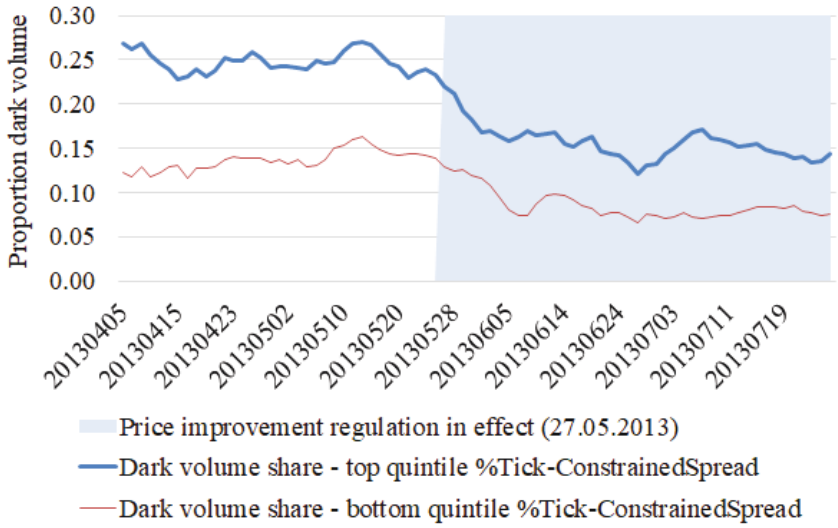
This table reports stage 2 results from 2SLS. The dependent variable is *PercentOnClose*. The fitted value of *LogPassive* is from model (1) in Table 5.4. All variables are at stock-day level. See Table 5.1 for detailed variable definitions. The sample stocks are within 50-stocks band on each side of stock number 300 by market capitalization. The sample period is May 2012–December 2018. T-statistics are reported in parentheses.***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	PercOnClose (1)	*PercOnClose* (2)	*PercOnClose* (3)	*PercOnClose* (4)	*PercOnClose* (5)
LogPassiveFitted	0.0016***	0.0017***	0.0017***	0.0014***	0.0017***
	(3.3402)	(3.4071)	(3.4684)	(2.9690)	(3.4782)
LogMktCap	0.8651**	1.0911**	0.9511***	0.8795**	0.7047
	(2.0583)	(2.5195)	(2.1731)	(2.1664)	(1.6203)
$LogMktCap^2$	−0.0466**	−0.0576***	−0.0510**	−0.0482**	−0.0389*
	(-2.2508)	(-2.7038)	(-2.3735)	(-2.4091)	(-1.8194)
$LogMktCap^3$	0.0008**	0.0010***	0.0009***	0.0009***	0.0007***
	(2.4575)	(2.9025)	(2.5856)	(2.6671)	(2.0318)
PercentCtPoint	−0.0306***			−0.0353***	−0.0327***
	(-6.5807)			(-7.8092)	(-7.2307)
Frag	0.1300***	0.1244***	0.1231***	0.1234***	0.1295***
	(11.6547)	(11.4738)	(11.4280)	(11.0538)	(11.6856)
OTTR	0.0021***	0.0021***	0.0021***	0.0021***	0.0019***
	(6.1380)	(6.2803)	(6.1611)	(6.2566)	(5.8763)
$TickToPrice \times D_{reg}$		0.3404	0.3984		
		(1.1362)	(1.3342)		
HighLowVolat			−0.1685***	−0.1562***	−0.1755***
			(-7.4308)	(-7.2216)	(-7.6226)
LogTradeSize				−0.0064***	
				(-7.1699)	
RelSpread					0.1329*
					(1.7517)
Adjusted R^2	12%	12%	12%	13%	12%
Clustered std. errors	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date
Fixed effects	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date

The instrument for dark trading is the interaction term between tick-to-price ratio and a dummy for the period post price improvement regulation. I use replicate the result in Foley & Putnins (2016), who find that the price improvement regulation affects stocks differently, depending on whether those stocks' spreads are constrained to one tick. In stage one, I confirm their finding that after price improvement regulation, dark trading decreased more in stocks with constrained spreads. In the second stage, I use the fitted value of dark trading, which captures the decrease in dark trading that is unrelated to trading on close.

Because price improvement regulation was introduced simultaneously with lowering block trading restrictions, I estimate the first stage for block trading within the same system of equations. The instruments for block trading are block tier dummies, interacted with the post regulation period dummy. I include the same set of instruments and control variables in the first stage regressions for both dark and block trading. In the second stage, therefore, the fitted values for dark and

Panel A. Proportion of dark trading, split by % tick-constrained



Panel B. Proportion of trading on close, split by % dark trading

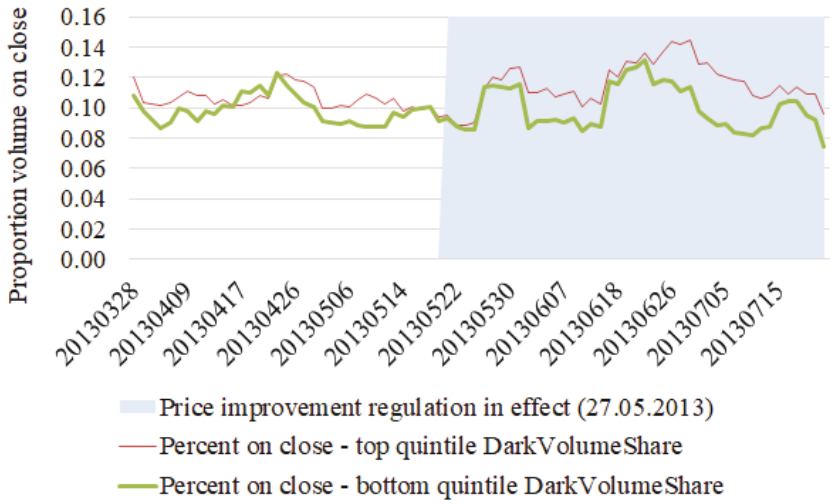


Figure 5.2: Trading on close and dark trading

Panel A plots the time series of *PercentDark* split by *PercTickConstrained*. Panel B plots plots the time series of *PercentOnClose* split by *PercentDark*. The time series cover the period March 1, 2012—July 31, 2013 (two months before and after the price improvement regulation 27.05.2013). The time series are weekly moving averages. The sample securities are S&P ASX 200 stocks.

block trading are exogenous with respect to trading on close. I control for time and stock fixed effects in both stages of the 2SLS framework and include market capitalization and stock volatility controls. The sample period covers a two-month window before and after March 27, 2013. The baseline regression specification is as follows:

Stage 1:

$$\begin{aligned} PercentDvolDark_{it} = \gamma_0 &+ \gamma_1 D_{Reg_t} PercentTickConstr_{it} + \\ &+ \gamma_2 D_{Tier2_{it}} + \gamma_3 D_{Tier3_{it}} + \gamma_4 LogMktCap_{it} + \\ &+ \gamma_5 HighLowVolat_{it} + \tau_t + \mu_i + \epsilon_{it} \end{aligned} \tag{5.3}$$

$$\begin{aligned} PercentDvolBlock_{it} = \beta_0 &+ \beta_1 D_{Reg_t} PercentTickConstr_{it} + \\ &+ \beta_2 D_{Tier2_{it}} + \beta_3 D_{Tier3_{it}} + \beta_4 LogMktCap_{it} + \\ &+ \beta_5 HighLowVolat_{it} + \tau_t + \mu_i + \varepsilon_{it} \end{aligned} \tag{5.4}$$

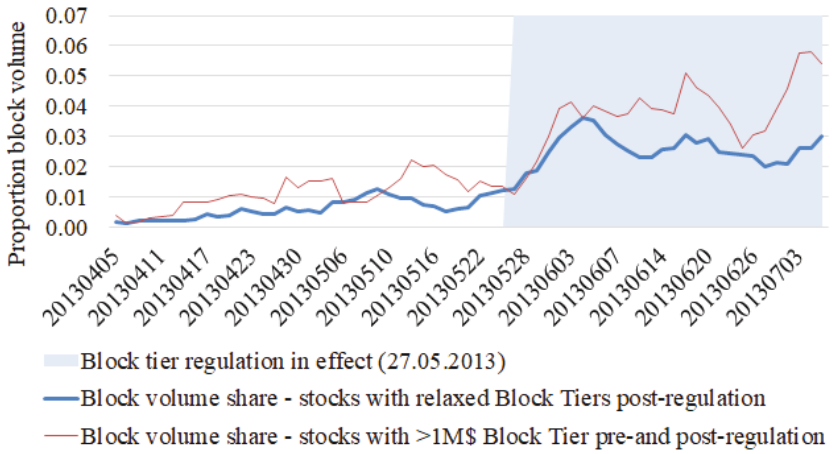
Stage 2:

$$\begin{aligned} PercentOnClose_{it} = \theta_0 &+ \theta_1 \widehat{PercentDvolDark}_{it} + \theta_2 \widehat{PercentDvolBlock}_{it} + \\ &+ \theta_3 LogMktCap_{it} + \theta_4 HighLowVolat_{it} + \theta_5 OTTR_{it} + \\ &+ \theta_6 LogPassive_{it} + \theta_7 LogDepth_{it} + \\ &+ \theta_8 Frag_{it} + \theta_9 RelSpread_{it} + \tau_t + \mu_i + \sigma_{it} \end{aligned} \tag{5.5}$$

where $PercentDvolDark_{it}$ is percent dollar volume executed in dark pools in stock i on day t, $PercentDvolBlock_{it}$ is percent dollar volume executed in blocks, D_{reg_t} is a dummy variable for the period post price improvement regulation, $PercentTickConstr_{it}$ is the percent of time in a trading day during which the stock's spread is tick-constrained, $D_{Tier2_{it}}$ is a dummy variable that takes the value of 1 after price improvement regulation, if a stock belongs to block trading Tier 2, and 0 otherwise, $LogDepth_{it}$ is the natural logarithm of depth (in AU$) at best quotes.

The first stage regression results (see Table 5.6) suggest that there is a substitution effect from dark to block trading among tick-constrained stocks. After the price improvement regulation, dark trading decreased by 4.68% for tick-constrained stocks, with an additional 3.47% decrease in Tier 2 stocks (block trading threshold lowered from AU$ 1 million to AU$ 500 thousand), and an additional 2.70% decrease in Tier 3 stocks (block trading threshold lowered from AU$ 1 million to AU$ 200 thousand). At the same time, block trading increased by 1.58% for tick-constrained stocks. The results are graphically illustrated in Figures 5.2 and 5.3.

Panel A. Proportion of block trading, split by Block Tiers



Panel B. Proportion of trading on close, split by % block trading

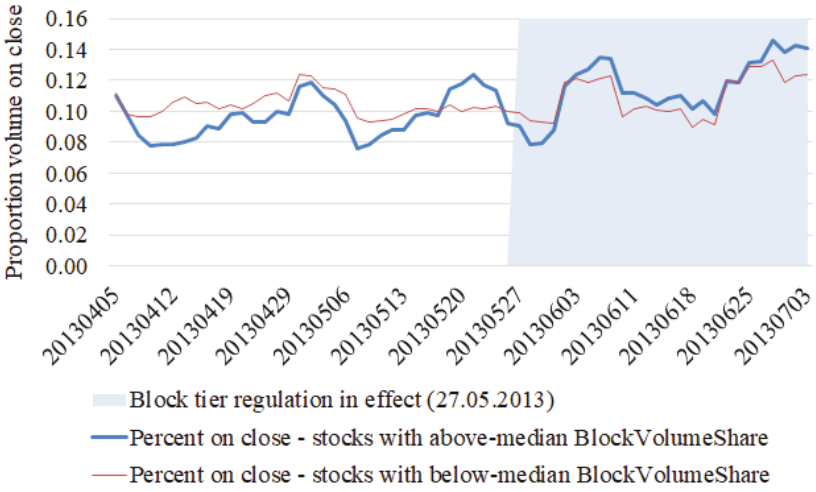


Figure 5.3: Trading on close and block trading

Panel A plots the time series of *PercentBlock* split by Block Tiers. The horizontal axis plots dates in YYYYMMDD format. Tier 2 and Tier 3 correspond to lowered block trading thresholds, from $1M to $500K and from $1M to $200K respectively, on 27.05.2013). Panel B plots the time series of *PercentOnClose* split by *PercentBlock*, before vs after the Block Tiers regulation (27.05.2013). The time series are weekly moving averages. The sample covers S&P ASX 200 stocks for the period March 1, 2012 — July 31, 2013.

In the second stage, I do not find evidence that trading on close is strongly related to exogenous changes in block or dark trading. I only confirm the evidence from descriptive analysis, suggesting that trading on close is strongly related to intraday liquidity, volatility, and the degree of market fragmentation. Trading on close on ASX is higher on less liquid stock-days on ASX (i.e., wider spreads and lower depth at best quotes), and for more fragmented stocks.

Table 5.6: Stage 1 – 2SLS regression results for dark and block trading (AU)

This table reports stage 1 results from 2SLS with two endogenous variables and three instruments. In regression (1), the dependent variable is $PercentDark$. In regression (2), the dependent variable is $PercentBlock$. The instrumental variables are: $PercTickConstr \times D_{reg}$, D_{Tier_2}, D_{Tier_3}. All variables are at stock-day level. See table 5.1 for variable definitions. The sample period is March 2013–July 2013. The sample stocks are S&P ASX 200 constituents. T-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Stage 1 regression results		
	$PercentDark$ (1)	$PercentBlock$ (2)
$PercentTickConstr \times D_{reg}$	-0.0468^{***}	0.0158^{***}
	(-3.3250)	(2.0656)
D_{Tier2}	-0.0347^{***}	0.0023
	(-3.7877)	(0.2533)
D_{Tier3}	-0.0270^{***}	-0.0063
	(-3.3349)	(-0.7570)
$LogMktCap^3$	-0.0151	0.0245^{***}
	(-1.1561)	(3.8536)
$HighLowVolat$	-0.3969^{***}	0.0126
	(-6.8690)	(0.5175)
Adjusted R^2	1.05%	0.23%
Clustered std. errors	Stock & Date	Stock & Date
Fixed effects	Stock & Date	Stock & Date

Given that the ASX closing auction is fairly transparent (information on order submissions, volumes and indicative prices is disseminated continuously), it is perhaps not surprising that there is no direct causal link between dark trading / block trading (which have no pre-trade transparency) and auction trading (which is pre-trade transparent in the ASX closing auction). One interpretation of these results is that the reduction in dark trading post price improvement regulation was partly offset by an increase in block trading, without an effect on auction trading.

5.5.2.3 High-frequency trading

Testing Hypothesis 4, I analyze the response of auction trading to an exogenous shock to HFT activity. I use the exchange connectivity speed upgrades (ITCH protocol introduction by ASX) as an instrument for HFT activity. Previous studies (e.g., Goldstein, Kwan, & Philip, 2018) document an increase in HFT activity after the ITCH protocol and co-location services were launched by ASX in February 2012. This event is unlikely to be driven by trading on close, while being strongly related to the degree of HFT activity. To allow for cross-sectional variation in the instrumental variable, I interact the ITCH dummy with tick-to-price ratio, relying on Yao & Ye (2018) argument HFTs compete on speed in tick-constrained stocks.

Table 5.7: Stage 2 – 2SLS regression results for dark and block trading (AU)

This table reports stage 2 results from 2SLS. The dependent variable is percent dollar volume on close (*PercentOnClose*). The instrumented variables are *PercentDarkFitted* from model (1) in Table 5.6 and *PercentBlockFitted* from model (2) in Table 5.6. All variables are at stock-day level. See Table 5.1 for variable definitions. The sample stocks are S&P ASX200 constituents. T-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Stage 2 regression results. (Both Percent Dark and Percent Block instrumented)					
	PercOnClose (1)	*PercOnClose* (2)	*PercOnClose* (3)	*PercOnClose* (4)	*PercOnClose* (5)
PercentBlockFitted	0.2772 (0.6859)	0.4831 (1.1985)	0.7466* (1.9165)	0.7491* (1.9275)	0.6419* (1.6545)
PercentDarkFitted	-0.0181 (-0.1343)	-0.0451 (-0.3376)	0.1038 (0.7832)	0.1126 (0.8483)	0.1940 (1.5818)
LogMktCap	-0.0060 (-0.5417)	-0.0137 (-1.2313)	-0.0126 (-1.2395)	-0.0046 (-0.4533)	0.0031 (0.2966)
HighLowVolat	−0.3645*** (-4.5700)	−0.4020*** (-4.9415)	−0.2923*** (-3.8714)	−0.2964*** (-3.9722)	−0.2025*** (-3.1259)
OTTR	0.0036*** (5.2183)	0.0035*** (5.2252)	0.0020*** (3.2179)	0.0020*** (3.1727)	
LogPassive	0.0002 (0.0463)	0.0000 (0.0033)	-0.0003 (-0.0609)	-0.0006 (-0.1350)	-0.0012 (-0.2365)
LogDepth		−0.0152*** (-8.8989)	−0.0162*** (-10.3628)	−0.0163*** (-10.4271)	−0.0085*** (-3.6066)
Frag			0.3464*** (15.9761)	0.3438*** (16.1774)	0.3056*** (12.1722)
RelSpread				1.2486** (2.3677)	1.4351*** (2.7238)
LogTradeSize					−0.0160*** (-4.0858)
Adjusted R^2	1.96%	3.05%	9.32%	9.38%	10.35%
Clustered std. errors	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date
Fixed effects	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date

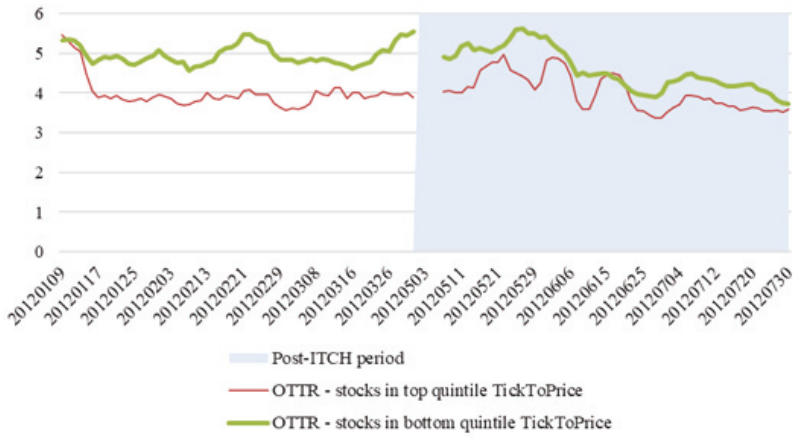
Further, Yao & Ye (2018) point out that after speed upgrades, speed competition between HFTs is intensified further in more fragmented stocks that are also tick-constrained. Therefore, a second version of the first-stage regression for HFT activity allows for a three-way interaction term between ITCH dummy, tick-to-price, and the fragmentation variable.

Following Goldstein, Kwan, & Philip (2018), I include two months prior to April 1, 2012 (the date of ITCH launch), and two months after May 1, 2012. The month of April is excluded to allow for adjustments in market participants' order routing behaviour. The baseline regression specification is as follows:

Stage 1 (version 1):

$$\begin{aligned} OTTR_{it} = {} & c_0 + c_1 D_{ITCH_t} TickToPrice_{it} + \\ & + c_3 Frag_{it} + c_4 LogMktCap_{it} + c_5 HighLowVolat_{it} + \\ & + \tau_t + \mu_i + \epsilon_{it} \end{aligned} \tag{5.6}$$

Panel A. Order-to-trade ratio, split by tick-to-price



Panel B. Proportion of trading on close, split by order-to-trade ratio

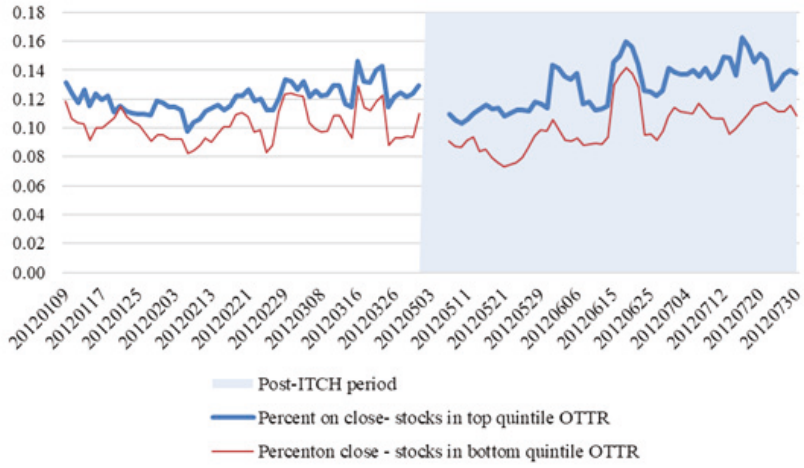


Figure 5.4: Trading on close and HFT

Panel A plots the time series of $OTTR$ split by $TickToPrice$. Panel B plots the time series of $PercentOnClose$ split by $OTTR$. The time series are weekly moving averages, before vs after the rollout of ITCH on ASX on April 1, 2012. The sample period is February 2012–June 2012, with the month of April excluded. The sample stocks are S&P ASX 200 constituents.

Stage 1 (version 2):

$$\begin{aligned} OTTR_{it} = c_0 + c_1 D_{ITCH_t} TickToPrice_{it} + c_2 Frag_{it} D_{ITCH_t} TickToPrice_{it} + \\ c_3 Frag_{it} + c_4 LogMktCap_{it} + c_5 HighLowVolat_{it} + \\ + \tau_t + \mu_i + \varepsilon_{it} \end{aligned} \tag{5.7}$$

Table 5.8: Stage 1 – 2SLS regression results for HFT (AU)

This table reports stage 1 results from 2SLS. The dependent variable is $OTTR$. The instrument in regression (1) is $(D_{ITCH} \times TickToPrice)$. The instruments in regression (2) are: $(D_{ITCH} \times Frag)$ and $(D_{ITCH} \times Frag \times TickToPrice)$. All variables are at stock-day level. Variable definitions are in Table 5.1. The sample period is February 2012 – June 2012, with the month of April excluded. The sample stocks are S&P ASX200 constituents. T-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Stage 1 regression results		
	OTTR (1)	*OTTR (2)*
$D_{ITCH} \times TickToPrice$	25.2475* (1.7376)	52.9740*** (3.5360)
$D_{ITCH} \times Frag \times TickToPrice$	893.5558*** (4.6927)	
$LogMktCap$	-0.0787 (-0.4231)	-0.0459 (-0.2450)
$HighLowVolat$	−7.9250*** (-8.3662)	−7.9005*** (-8.4499)
$Frag$		3.5670*** (3.5359)
Adjusted R^2	1.06%	0.97%
Clustered std. errors	Stock & Date	Stock & Date
Fixed effects	Stock & Date	Stock & Date

Stage 2:

$$\begin{aligned} PercentOnClose_{it} = g_0 + g_1\widehat{OTTR}_{it} + g_2 LogMktCap_{it} + g_3 HighLowVolat_{it} \\ + g_4 Frag_{it} + g_5 PercCtPoint_{it} + g_6 RelSpread_{it} \\ + \tau_t + \mu_i + e_{it} \end{aligned} \tag{5.8}$$

The first stage regression results (see Table 5.8) suggest that both proposed instruments are relevant. As in Ye & Yao (2018), ASX order-to-trade ratios are lower for high tick-to-price stocks than in low tick-to-price ones (see Figure 5.4). However, after the ITCH introduction, high tick-to-price stocks experience a greater jump in OTTR on ASX, compared to low tick-to-price stocks. This jump after ITCH upgrades reflects greater speed competition for lucrative market making in tick-constrained stocks. In version 2 of stage 1, this speed competition effect is amplified for more fragmented stocks. The more fragmented the stock, the greater the speed competition for liquidity provision in the stock.

The exogenous shock to HFT activity has only weak effect on trading on close: the percent value traded in the closing auction increases by 1% for each unit increase in OTTR (see Table 5.9). However, this becomes insignificant, if I control for stock

Table 5.9: Stage 2 – 2SLS regression results for HFT (AU)

This table reports stage 2 results from 2SLS. The dependent variable is *PercentOnClose*. In regressions (1)-(3), the fitted value of $OTTR$ is from model (1) in Table 5.8. In regressions (4)-(5), the fitted value of $OTTR$ is from model (2) in Table 5.8. All variables are at stock-day level. Variable definitions are in Table 5.1. The sample period is February 2012–June 2012, with the month of April excluded. The sample stocks are S&P ASX 200 constituents. T-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Stage 2 regression results.					
	PercOnClose (1)	*PercOnClose* (2)	*PercOnClose* (3)	*PercOnClose* (4)	*PercOnClose* (5)
OTTRFitted	0.0129**	0.0048	-0.0093	0.0132	-0.0043
	(2.1754)	(0.8328)	(-1.3848)	(1.3303)	(-0.3662)
LogMktCap	−0.0089*	0.0016	0.0010	-0.0083	0.0013
	(-1.7847)	(0.2725)	(0.1729)	(-1.5712)	(0.2343)
HighLowVolat	−0.5128***	−0.5934***	−0.7015***	−0.5091***	0.1840***
	(-6.4155)	(-8.2557)	(-9.1667)	(-4.8304)	(-5.9741)
PercentCtPoint	−0.0489***	−0.0504***	−0.0499***	−0.0489***	−0.6627***
	(-4.1847)	(-4.3623)	(-4.3490)	(-4.1749)	(-4.3050)
RelSpread		2.7063**	3.4974***		−0.0498***
		(2.5569)	(3.4747)		(2.7871)
Frag			0.1976***	0.1203**	3.1991***
			(4.0590)	(2.0444)	(2.8732)
Adjusted R^2	2.41%	2.51%	2.73%	2.59%	2.71%
Clustered std. errors	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date
Fixed effects	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date

liquidity. Hence, I find no strong evidence that increased HFT speed competition leads to greater trading on close. One possibility is that order routing takes longer than two months to reflect market participants' reaction to increased picking-off risk from HFT activity. Additionally, the 2012 levels of trading on close (at about 12% for top-quintile OTTR stocks) were not quite as high as they are today. This might be another reason market participants did not immediately tap into closing liquidity as an alternative to the intraday continuous limit order book.

5.6 Price discovery analysis of trading on close

For price discovery estimation, I use the methodology developed in Chapter 4. I estimate a structural VAR model with $n = 3$ phases, using the daily open, pre-close and closing prices from Thomson Reuters Tick History. The data cover the period 2002–2018. As the pre-close price, I use the last available trade price 10 minutes before the closing auction.

The Australian sample covers S&P ASX 200 constituents, and the US sample covers 241 US stocks stratified by market cap (see Section 2.4 of Chapter 2 for

a more detailed description of the US sample). This sampling approach ensures that the selection of stocks is representative of today's markets. To be able to estimate the VARs per stock-year, I require a stock to have at least 100 stock-day observations in a given year, and have no gaps in consecutive trading days that are longer than a week.

I present the descriptive statistics of the estimation results in Table 5.10. Note that information shares (noise shares) across the three phases do not sum up to exactly one, as they are yearly averages across stocks (they do, however, sum up to one for each stock-year estimation). Note also that IS and NS estimates are always between zero and one.

Table 5.10: Information shares and noise shares results

This table provides summary estimates of information shares (IS) and noise shares (NS) from the structural VAR. The VAR models are estimated by stock-year, using 3,678 (3,722) stock-day observations for Australian (US) markets over the period 2002—2018.

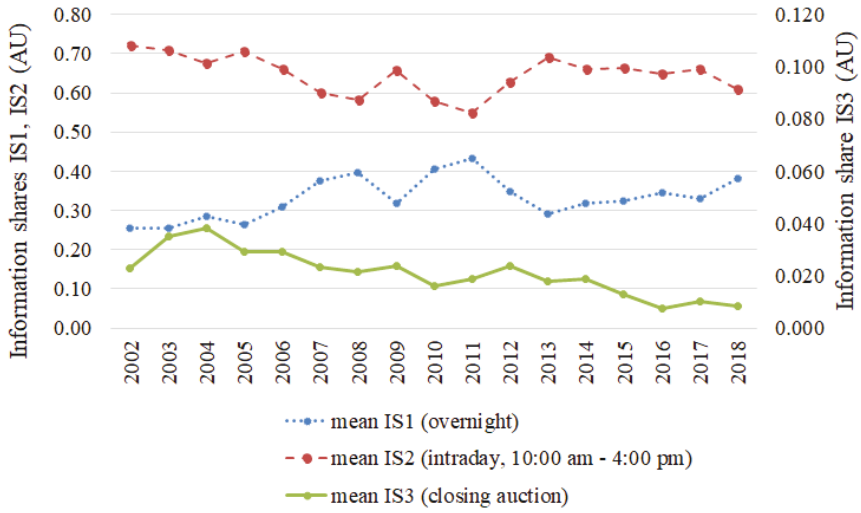
	Mean	StDev	25th pctl	50th pctl	75th pctl
Panel A: Australian market					
IS1 (Overnight, 4:00 pm – 10:00 am)	0.3323	0.1689	0.2225	0.3078	0.4052
IS2 (Intraday, 10:00 am – 4:00 pm)	0.6467	0.1655	0.5728	0.6693	0.7535
IS3 (Close, 4:00 pm – 4:12 pm)	0.0210	0.0456	0.0000	0.0021	0.0281
NS1 (Overnight, 4:00 pm – 10:00 am)	0.2793	0.3595	0.0000	0.0000	0.5751
NS2 (Intraday, 10:00 am – 4:00 pm)	0.2589	0.3425	0.0000	0.0000	0.5095
NS3 (Close, 4:00 pm – 4:12 pm)	0.2647	0.3493	0.0000	0.1022	0.3856
Panel B: US market					
IS1 (Overnight, 4:00 pm – 9:30 am)	0.2534	0.1902	0.1266	0.2183	0.3312
IS2 (Intraday, 9:30 am – 3:50 pm)	0.7163	0.1967	0.6229	0.7617	0.8514
IS3 (Close, 3:50 pm – 4:00 pm)	0.0303	0.0566	0.0000	0.0041	0.0360
NS1 (Overnight, 4:00 pm – 9:30 am)	0.3544	0.3831	0.0000	0.2106	0.7094
NS2 (Intraday, 9:30 am – 3:50 pm)	0.2566	0.3506	0.0000	0.0000	0.5112
NS3 (Close, 3:50 pm – 4:00 pm)	0.2001	0.3085	0.0000	0.0446	0.2526

Table 5.10 results offer an insight into the distribution of information and noise shares in the US and Australian markets. The continuous trading session is responsible for most of the price discovery in both markets. In Australia (the US), the intraday trading phase contributes 64.67% (71.63%) of information on average. This suggests that the continuous trading hours bring in a disproportionately high share of information, relative to their time duration, consistent with French & Roll (1986).

The average information content of the closing phase is rather low in both Australian (2.10%) and US markets (3.03%). However, even more revealing is the time series dynamic of the closing price informativeness (see Figures 5.5 and 5.6). In Australia, the information share of the close decreased dramatically: from 2.27% in 2002 to 0.85% in 2018. In the US, it decreased from 4.09% to 2.89% over the

same period. This is rather surprising, given that both markets saw substantial growth in volumes on close over the same period (the close represents about 25% of volume in Australia, and up to 12% in the US). This result suggests that informed traders did not shift to the close at the same rate as the passive index funds did, contrary to Admati & Pfleiderer's (1988) prediction that the informed "pool" with the uninformed in large liquidity events like closing auctions.

Panel A. Information shares (AU market)



Panel B. Noise shares (AU market)

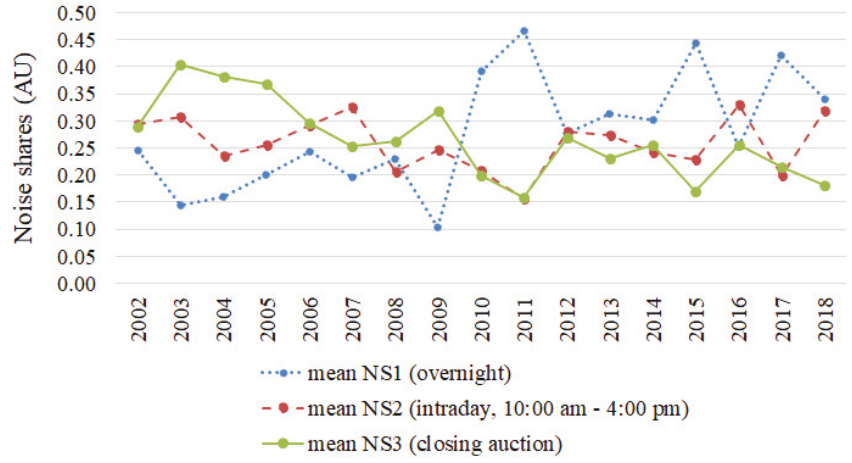


Figure 5.5: Information shares and noise shares for the Australian market

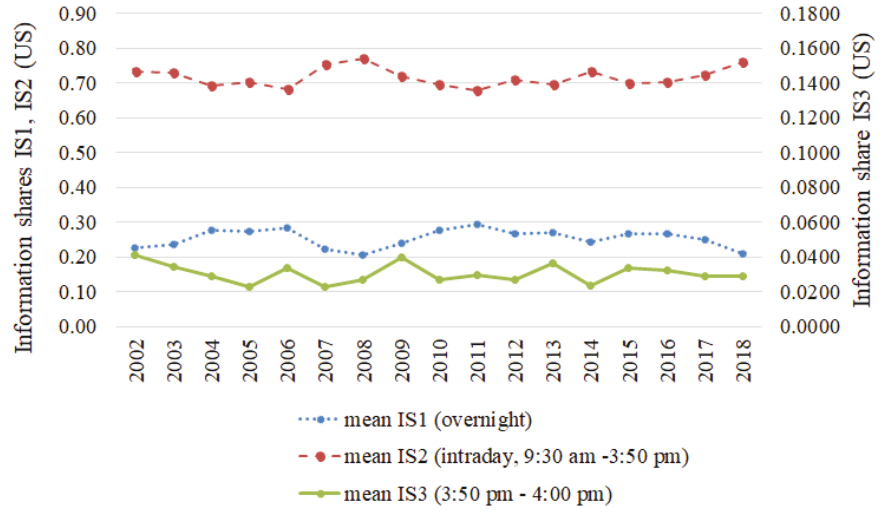
Panel A plots yearly averages of information shares (IS), and Panel B plots yearly averages of noise shares (NS) for the Australian market. The VAR models are estimated by stock-year, using 3,678 stock-day observations for the Australian market over the sample period of 2002—2018.

Given the substantial increase in volumes, are the closing mechanisms producing noisier prices? Noise shares analysis suggests not (or not beyond historical levels). In fact, the noise share of the Australian close has decreased from 28.94% to 18.19% during 2002–2018. In the US, the noise share of close is almost the same in 2018 (23.46%) as in 2002 (23.07%). Interestingly, the US noise share of close has been following an upward trend (increasing from 14.06 % to 23.46%) during the period 2011–2018, which is the period of substantial ETF growth. In 2018, US closing prices are more noisy, on average, than pre-close prices: the close accounts for 23.46% of noise out of the three prices considered, while pre-close — for 19.55%.

The results in Figures 5.5 and 5.6 also highlight the importance of the opening auction in incorporating overnight information into prices. In the US, the 2018 information share of overnight period is 21.04%, on average, while in Australia it is 38.30%. At the same time, the US opening price is more noisy, with the noise share of 41.77% (compared to 34.09% in Australia). These results confirm the intuition that opening prices are noisy when a substantial amount of overnight information is being incorporated into prices during a relatively short period of time. It is to be expected that the Australian market (which opens after the US and European markets have closed) has less noisy opening prices. Given the information available from the US and European markets, Australian traders enjoy extra time to arrive at consensus prices before the trading commences.

Overall, the price discovery results suggest that the increase in trading on close during 2002–2018 did not go hand in hand with greater price discovery in the closing auction in the US and Australia. Annual averages of information shares on close have declined (in Australia) or stayed unchanged (in the US) over the sample period, although trading on close increased in both markets. However, the noise in closing prices also has declined over 2002–2018 (in Australia), or has not exceeded the levels of a decade ago (in the US). This means that so far, the shift of volumes to the close has not made prices much noisier than before. At the same time, the low information share on close (and high information share of the intraday session) means that continuous trading remains an important vehicle for price discovery.

Panel A. Information shares (US market)



Panel B. Noise shares (US market)

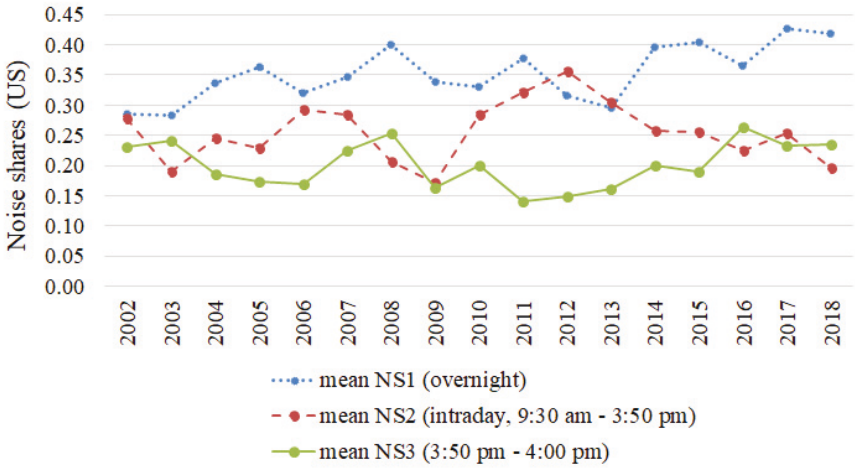


Figure 5.6: Information shares and noise shares for the US market

Panel A plots yearly averages of information shares (IS), and Panel B plots yearly averages of noise shares (NS) for the US market. The VAR models are estimated by stock-year, using 3,722 stock-day observations for the US market over the sample period of 2002—2018.

5.6.1 Validation checks for the price discovery results

I perform several simple checks to confirm the price discovery results (see Figures 5.8–5.11 in Appendix 5.3).[15] First, I shift the pre-close time back by one hour and plot the information shares and noise shares in each phase. The resulting plots

[15]To conserve space, these validation checks cover the later part of the sample, namely the period 2008–2018.

for "placebo" pre-close at 3 pm, 2 pm, 1 pm, and 12 pm are presented in Figure 5.8. As expected, the intraday information share decreases (and closing phase information share increases) as I shift the pre-close from 3 pm to 12 pm, reflecting the scaling up of intraday information variances with time. Noise variances, on the other hand, do not scale up with time, as expected by the properties of white noise.

Second, I plot realized variances in each phase of the trading day (Figure 5.9) and find that the closing auction period has the lowest realized variance of the three phases, followed by overnight period, while intraday trading is associated with the highest realized variance. These results are in line with French & Roll (1986), who document that returns are more volatile during exchange trading hours, compared to overnight hours. Realized variances are also in line with my price discovery results, as they illustrate that prices move more intraday (incorporating new information), and relatively less during the closing phase (reflecting little information being incorporated in this phase).

Third, I plot the proportion of zero-return observations in each phase of the trading day (Figure 5.10). These observations capture instances when e.g., the closing price is the same as pre-close (then, $r_3 = 0$), or pre-close is the same as the opening price (then, $r_2 = 0$). Consistent with the price discovery results, I find that the closing phase of the market produces the highest proportion of instances with zero-return observations, which reflects the fact that there is little informed trading to move prices in this phase.

Fourth, I illustrate the extent of overnight return reversals in Figure 5.11. I run a simple regression $r_{1i,t} = \beta_0 + \beta_1 r_{3i,t-1} + \mu_i + \varepsilon_{i,t}$, where $r_{1i,t}$ is overnight log-return in stock i on day t, $r_{3i,t-1}$ is the closing period log-return in stock i on day $t-1$, and μ_i is stock fixed effect. I obtain significant negative coefficients in the regression results for all years (except for 2015 for AU, and 2013 for US), which suggests that overnight returns "undo" the closing returns on average. This finding is consistent with overnight reversals documented in prior literature (e.g., Cushing & Madhavan, 2000); in the context of my price discovery results, it confirms that the closing phase is not informed, as stocks consistently experience overnight reversals.

5.7 Conclusions

This chapter investigates what drives the shift in trading volumes to the close of the market over the past decade, and whether this shift is accompanied by greater informativeness of the closing price. The results suggest that index investing is by far the most significant factor explaining the increase in dollar volume share on close (consistent with Hypothesis 1). In line with this evidence is the finding that closing price informativeness did not improve as more volume shifted towards the close. I show that two major closing mechanisms (the closing call auction in Australia and the on-close facility in the US) display similar trends: relatively low and decreasing (in Australia) information share of the close.

With respect to other hypothesized drivers of trading on close (block- and dark trading limitations, and the extent of HFT activity), I do not find strong evidence that these factors causally affect trading on close in Australia during my sample period (contrary to Hypotheses 2–4).

I also do not find support for regulators' concerns about closing prices becoming noisier as the closing mechanisms experience an influx of trading volume. Contrary to Hypothesis 5, the noise share of the closing price has decreased over the period 2002–2018 (in Australia), or fluctuated within the historically observed levels (in the US). However, since 2011, the US market has witnessed an upwards trend in the noise share of the close, and this might suggest that trading by ETFs (which increased over the same period) is getting reflected in noisier closing prices.

This chapter contributes several novel insights. Firstly, it shows the market structure implications of changing fund management scene (shift from active to passive). Secondly, it evaluates the price informativeness of the closing phase, which now accounts for a substantial proportion of intraday volumes. Finally, it addresses regulatory concerns about the shift of volumes towards the close of the trading session by showing how the noise share of close changes around such shifts. The findings imply that increased volumes on close can result in noisy prices, while the continuous limit order book remains the key driver of price discovery.

Appendix 5.1. Closing mechanism timeline in Australian and US markets

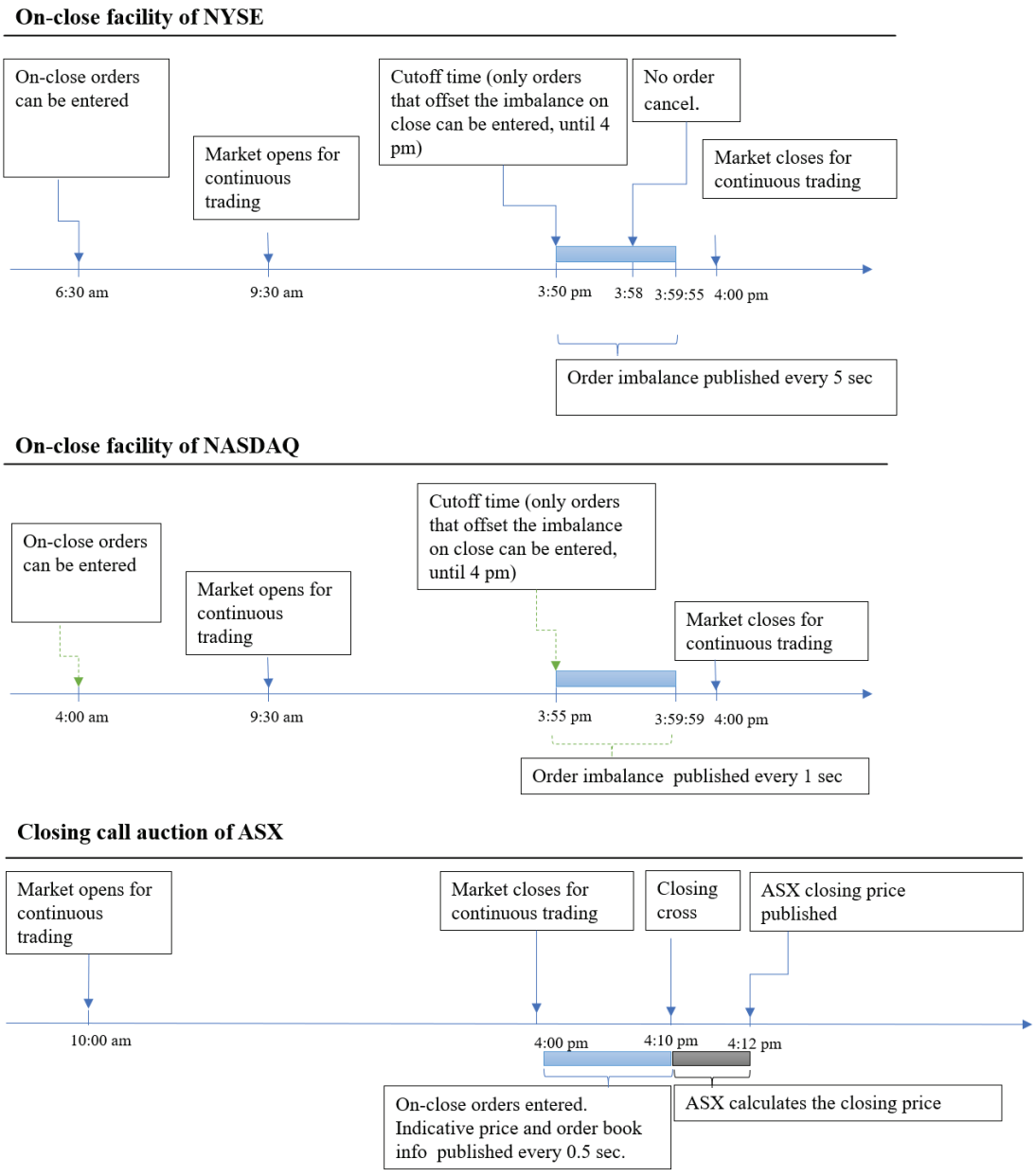


Figure 5.7: Closing mechanisms in AU and US markets

Appendix 5.2. Additional regression results for trading on close

Table 5.11: Time series regression results for trading on close (AU)

This table reports the time series regression results for the Australian market. The dependent variable is $\Delta PercentOnClose$. The independent variables are in column one. Detailed variable definitions are in Table 5.1. Δ indicates first differences between day t and day $t-1$. Each observation is a daily average across stocks in ASX All Ordinaries index. The number of observations is 1,687 and the sample period is May 2012 – December 2018. T-statistics are reported in parentheses. ***,** , and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	$\Delta Percent$ $OnClose(1)$	$\Delta Percent$ $OnClose(2)$	$\Delta Percent$ $OnClose(3)$	$\Delta Percent$ $OnClose(4)$	$\Delta Percent$ $OnClose(5)$	$\Delta Percent$ $OnClose(6)$	$\Delta Percent$ $OnClose(7)$
$\Delta LogPassive$	0.01	0.04	0.04	0.04	0.04	0.04^{*}	0.04^{*}
	(0.55)	(1.53)	(1.52)	(1.63)	(1.60)	(1.68)	(1.67)
$\Delta PercCtPoint$	-1.08^{***}	-1.06^{***}	-0.90^{***}	-1.07^{***}	-1.11^{***}	-1.01^{***}	-0.98^{***}
	(-6.28)	(-6.19)	(-5.60)	(-6.53)	(-6.79)	(-6.67)	(-6.61)
$\Delta OTTR$	-0.02^{***}	-0.02^{***}			-0.02^{***}	-0.02^{***}	-0.01^{***}
	(-3.46)	(-3.58)			(-4.10)	(-3.88)	(-3.44)
$\Delta LogMktCap$		-0.16^{**}	-0.13^{*}	-0.18^{**}	-0.19^{***}	-0.23^{***}	-0.13^{*}
		(-2.18)	(-1.77)	(-2.42)	(-2.64)	(-3.11)	(-1.92)
$\Delta LogTradeSize$			0.04^{***}			0.04^{***}	0.04^{***}
			(4.41)			(4.43)	(4.25)
$\Delta Frag$				0.67^{***}	0.69^{***}	0.69^{***}	0.72^{***}
				(6.84)	(7.02)	(7.16)	(7.16)
$\Delta RelSpread$						-2.60^{***}	-2.51^{***}
						(-3.50)	(-3.53)
$\Delta HighLowVolat$							1.81^{***}
							(4.69)
R^2	6%	6%	9%	11%	12%	16%	18%

Table 5.12: Time series regression results for trading on close (US)

This table reports the time series regression results for the US market. The dependent variable is $\Delta PercentOnClose$. The independent variables are in column one. Detailed variable definitions are in Table 5.1. Δ indicates first differences between day t and day $t-1$. Each observation is a daily average across 241 US stocks. The number of observations is 4,502 and the sample period is January 2000 – December 2017. T-statistics are reported in parentheses. ***,**, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	$\Delta Percent$ $OnClose(1)$	$\Delta Percent$ $OnClose(2)$	$\Delta Percent$ $OnClose(3)$	$\Delta Percent$ $OnClose(4)$	$\Delta Percent$ $OnClose(5)$	$\Delta Percent$ $OnClose(6)$
$\Delta PercentDark$	0.0009	0.0009	0.0008	0.0009	0.0009	0.0009
	(0.9700)	(1.0720)	(0.9282)	(0.9955)	(0.9966)	(0.9679)
$\Delta OTTR$	-0.0047^{***}		-0.0044^{***}	-0.0038^{***}	-0.0038^{***}	-0.0039^{***}
	(-8.8519)		(-8.3413)	(-7.4055)	(-7.4178)	(-7.5569)
$\Delta LogMktCap$	-0.0830	-0.0007	-0.1135	-0.0267	-0.0292	-0.0021
	(-0.9499)	(-0.0081)	(-1.2819)	(-0.3014)	(-0.3746)	(-0.0271)
$\Delta LogTradeSize$		-0.0605^{***}				-0.0415^{***}
		(-5.9708)				(-4.4460)
$\Delta Frag\ (Num\ mkts)$			0.0817^{***}	0.0359^{**}	0.0366^{**}	0.0353^{**}
			(4.7787)	(2.1678)	(2.2810)	(2.2106)
$\Delta RelSpread$				-4.2355^{***}	-4.2397^{***}	-3.9925^{***}
				(-14.5305)	(-13.8784)	(-13.1400)
$\Delta HighLowVolat$					-0.0568	-0.0820
					(-0.1095)	(-0.1581)
R^2	1.41%	0.74%	1.93%	4.34%	4.35%	4.67%

Table 5.13: Cross-sectional regression results for trading on close (AU)

This table reports the cross-sectional regression results for Australian market. The dependent variable is *PercentOnClose*. The independent variables are in column one. Detailed variable definitions are in Table 5.1. Each observation is a daily average across all days of the sample (May 2012 – December 2018). The number of observations is 952. T-statistics are reported in parentheses. ***,**, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	Percent OnClose(1)	*Percent OnClose(2)*	*Percent OnClose(3)*	*Percent OnClose(4)*	*Percent OnClose(5)*	*Percent OnClose(6)*
LogPassive	0.0012^{***}	0.0016^{***}	0.0012^{***}	0.0012^{***}	0.0012^{***}	0.0014^{***}
	(6.8725)	(8.7708)	(7.1881)	(7.0270)	(7.5322)	(8.2155)
PercCtPoint	0.4803^{***}	0.5083^{***}	0.3789^{***}	0.3737^{***}	0.3868^{***}	0.3674^{***}
	(11.4589)	(11.8710)	(8.9664)	(8.8196)	(9.0516)	(8.8012)
OTTR	-0.0027^{*}			-0.0019	-0.0022	-0.0029^{**}
	(-1.8082)			(-1.2818)	(-1.5309)	(-2.0779)
LogMktCap	0.0135^{***}	0.0113^{***}	0.0135^{***}	0.0138^{***}	0.0147^{***}	0.0130^{***}
	(14.1191)	(11.7969)	(14.5430)	(14.5600)	(14.7404)	(11.9970)
LogTradeSize		0.0071^{***}				
		(3.8637)				
Frag			0.0657^{***}	0.0646^{***}	0.0672^{***}	0.0822^{***}
			(7.4333)	(7.1799)	(7.5355)	(8.0568)
RelSpread					0.1343^{***}	0.1492^{***}
					(3.3872)	(3.4951)
HighLowVolat						-0.2499^{***}
						(-3.3528)
R^2	70%	71%	72%	72%	73%	73%

Table 5.14: Cross-sectional regression results for trading on close (US)

This table reports the cross-sectional regression results for the US market. The dependent variable is *PercentOnClose*. The independent variables are in column one. Detailed variable definitions are in Table 5.1. Each observation is a daily average across all days of the sample (January 2000 – December 2017). The number of observations is 1,846. T-statistics are reported in parentheses. ***,**, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	Percent OnClose(1)	*Percent OnClose(2)*	*Percent OnClose(3)*	*Percent OnClose(4)*	*Percent OnClose(5)*	*Percent OnClose(6)*
PercentPassive	0.0180^{***}					
	(22.5282)					
LogPassive		0.0211^{***}	0.0137^{***}	0.0131^{***}	0.0162^{***}	0.0157^{***}
		(13.9242)	(8.4817)	(7.9113)	(9.0003)	(8.7365)
PercentDark	0.0040	0.0035	-0.0014^{***}	-0.0013^{***}	-0.0018^{***}	-0.0012^{**}
	(0.9641)	(0.9354)	(-3.5449)	(-3.1234)	(-5.0516)	(-2.0566)
OTTR		0.0013^{***}		0.0012^{***}	0.0012^{***}	0.0009^{***}
		(8.1571)		(7.6277)	(7.7057)	(5.9456)
LogMktCap		-0.0255^{***}	-0.0252^{***}	-0.0199^{***}	-0.019^{***}	-0.0240^{***}
		(-12.2341)	(-11.9124)	(-9.4580)	(-9.1865)	(-12.1029)
LogTradeSize	-0.0043					
	(-1.1472)					
Frag (Num mkts)			0.0106^{***}	0.0113^{***}	0.0112^{***}	0.0117^{***}
			(13.8199)	(16.3366)	(16.2990)	(17.3490)
RelSpread					1.1234^{***}	1.3237^{***}
					(3.8511)	(4.4258)
HighLowVolat						-1.0288^{***}
						(-7.8048)
R^2	15%	24%	23%	31%	32%	36%

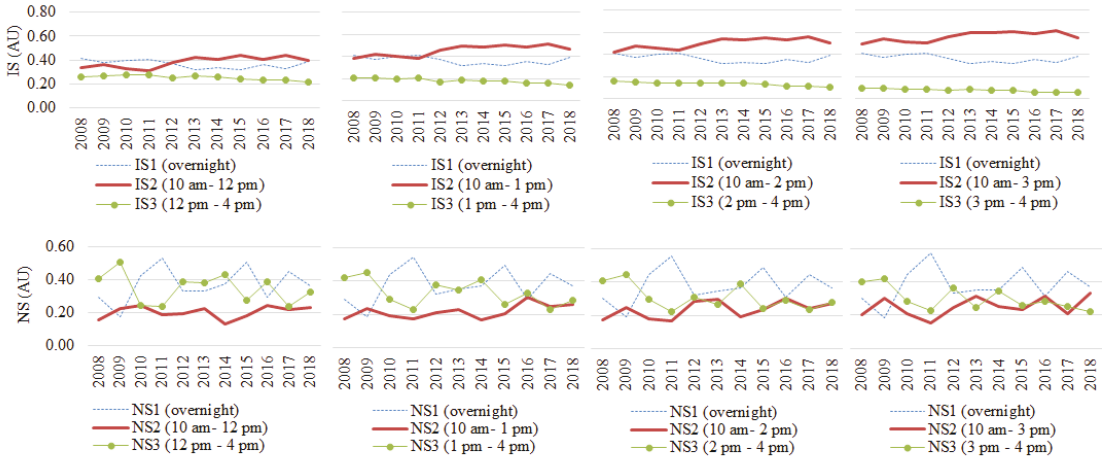
Table 5.15: Stage 2 – 2SLS regression results for passive investing (AU)

This table reports stage 2 results from 2SLS. The dependent variable is *PercentOnClose*. The fitted value of *LogPassive* is from model (1) in Table 5.4. All variables are at stock-day level. See Table 5.1 for detailed variable definitions. The sample stocks are within 100-stocks band on each side of stock number 300 by market capitalization. The sample period is May 2012 – December 2018. T-statistics are reported in parentheses.***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	Percent OnClose(1)	*Percent OnClose(2)*	*Percent OnClose(3)*	*Percent OnClose(4)*	*Percent OnClose(5)*
LogPassiveFitted	0.0020^{***}	0.0020^{***}	0.0021^{***}	0.0019^{***}	0.0021^{***}
	(10.4949)	(10.7695)	(11.2021)	(9.6756)	(11.0580)
LogMktCap	0.1740	0.3515	0.1536	0.0404	-0.0525
	(0.3688)	(0.7361)	(0.3222)	(0.0858)	(-0.1111)
$LogMktCap^2$	-0.0121	-0.0207	-0.0114	-0.0064	-0.0013
	(-0.5268)	(-0.8938)	(-0.4925)	(-0.2806)	(-0.0556)
$LogMktCap^3$	0.0003	0.0004	0.0003	0.0002	0.0001
	(0.7122)	(1.0788)	(0.6877)	(0.5046)	(0.2478)
PercentCtPoint	-0.0305^{***}			-0.0355^{***}	-0.0337^{***}
	(-9.5247)			(-11.4239)	(-10.7750)
Frag	0.1222^{***}	0.1176^{***}	0.1157^{***}	0.1161^{***}	0.1211^{***}
	(19.5108)	(19.3682)	(19.2895)	(18.5546)	(19.5452)
OTTR	0.0023^{***}	0.0024^{***}	0.0023^{***}	0.0022^{***}	0.0022^{***}
	(10.2344)	(10.5692)	(10.3871)	(10.1873)	(9.9793)
$TickToPrice \times D_{reg}$		0.2566	0.3235^{*}		
		(1.4128)	(1.7895)		
HighLowVolat			-0.2130^{***}	-0.2038^{***}	-0.2204^{***}
			(-11.6417)	(-11.5785)	(-11.8908)
LogTradeSize				-0.0061^{***}	
				(-10.9202)	
RelSpread					0.0702
					(1.5303)
Adjusted R^2	9%	9%	9%	9%	9%
Clustered std. errors	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date
Fixed effects	Stock & Date	Stock & Date	Stock & Date	Stock & Date	Stock & Date

Appendix 5.3. Validation checks for the price discovery results

Panel A. Australian markets



Panel B. US markets

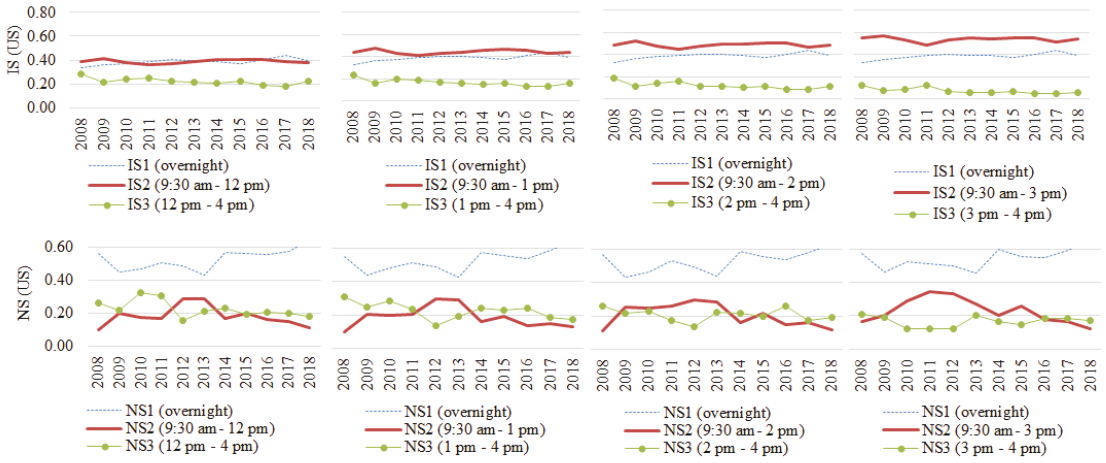
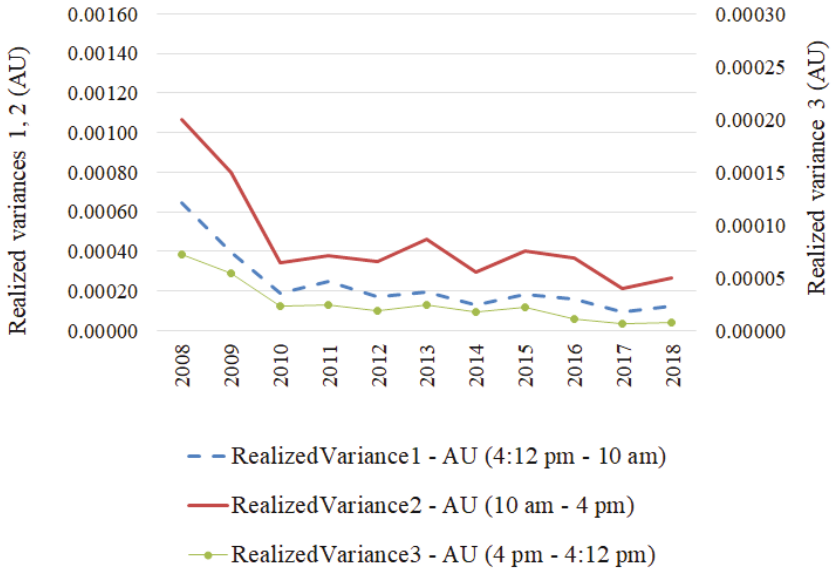


Figure 5.8: Information shares and noise shares, with shifting pre-close time

This figure plots information shares (IS) and noise shares (NS) for different splits of the 24-hour period. Going left to right, the pre-close price shifts from 12 pm to 3 pm. Panel A (B) uses prices from the consolidated data feed for Australian (US) markets. The results are from the VAR models estimated by stock-year, using 2,392 (2,063) stock-day observations for Australian (US) markets over the sample period of 2008-2018.

Panel A. Australian markets



Panel B. US markets

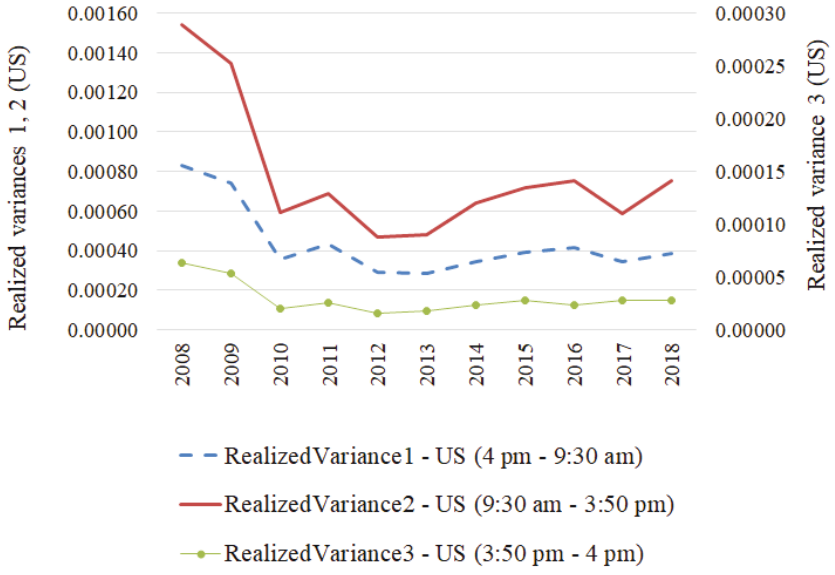
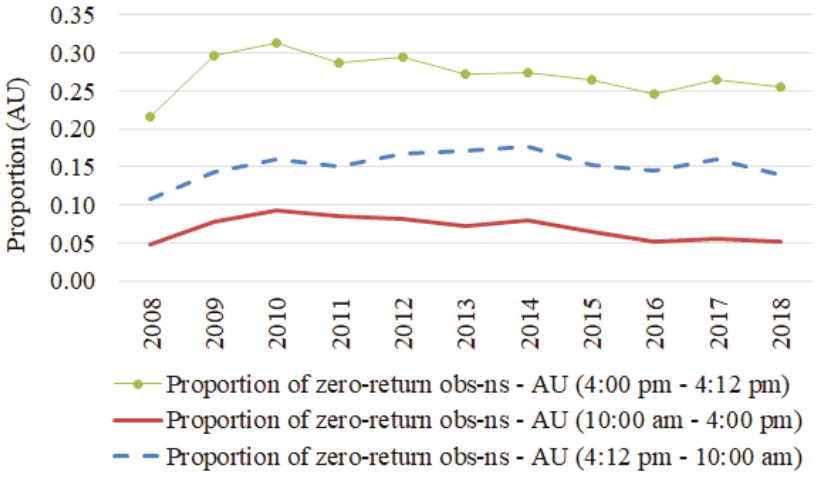


Figure 5.9: Realized variances in three phases of the trading day

Panel A (B) plots annual averages of realized variances for the Australian (US) markets. Realized variances are computed at stock-day level and averaged across stock-days in a given year. The sample covers 2,392 (2,063) stock-day observations for Australian (US) markets over the sample period of 2008–2018.

Panel A. Australian markets



Panel B. US markets

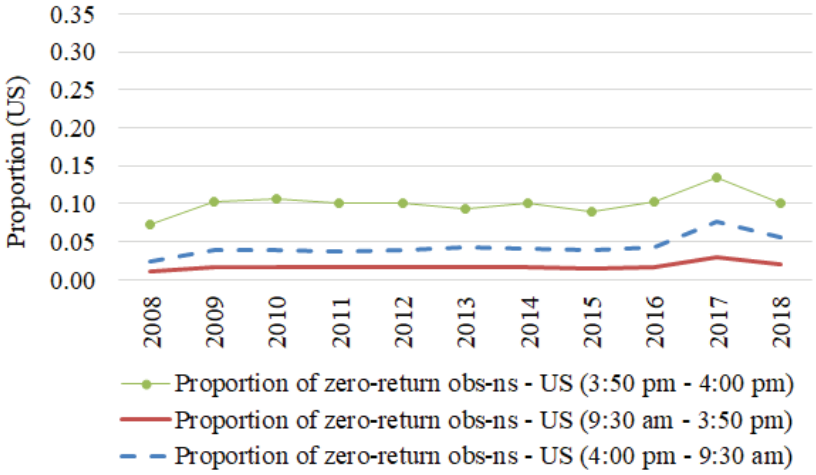
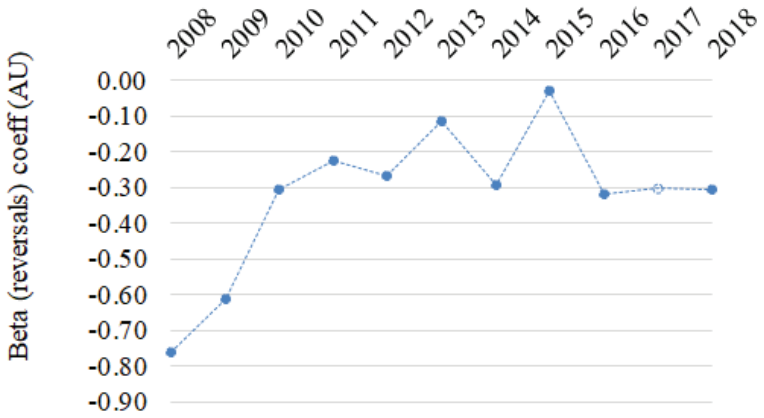


Figure 5.10: Proportion of zero-return observations in three phases of the trading day

Panel A (B) plots the average annual proportion of zero-return observations in the Australian (US) markets. The estimates are computed at stock-day level and averaged across stock-days in a given year. The proportion is the count of zero-return observations in a given phase of the trading day, divided by the count of of all return observations in this phase. The sample covers 2,392 (2,063) stock-day observations for Australian (US) markets over the period 2008–2018.

Panel A. Australian markets



Panel B. US markets

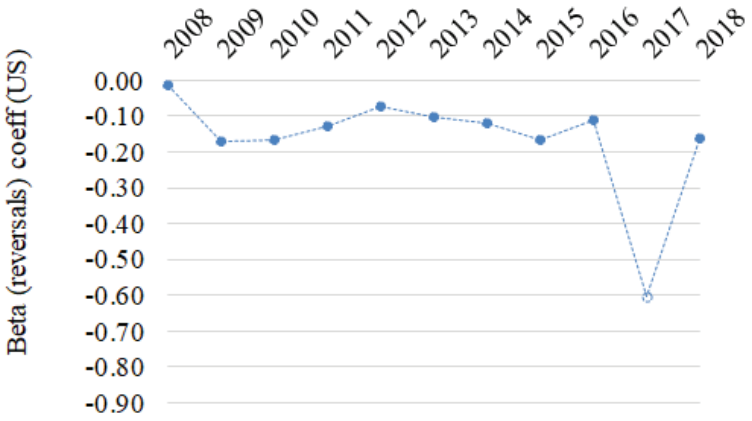


Figure 5.11: Overnight return reversals

Panel A (B) plots the beta coefficient indicating the extent of overnight return reversals in the Australian (US) markets. The reversals coefficient is β_1 from the following regression: $r_{1i,t} = \beta_0 + \beta_1 r_{3i,t-1} + \mu_i + \varepsilon_{i,t}$, where $r_{1i,t}$ is overnight log-return in stock i on day t, $r_{3i,t-1}$ is closing period log-return in stock i on day $t-1$, μ_i is stock fixed effect. Regressions are run for each year, and betas are averaged across stocks. The sample covers 2,392 (2,063) stock-day observations for Australian (US) markets over the sample period of 2008–2018.

Chapter 6

Conclusions and Future Research

Truth does not remain the same forever, but keeps changing continuously.
Volga, The Liberation of Sita.

The overarching purpose of this thesis is to examine the latest developments in microstructure of financial markets. Specifically, the thesis investigates what drives the changes in how people trade, how these changes are manifested and measured, and what effects they have on market quality. The thesis also studies the social welfare implications of these effects, and proposes regulatory actions based on the findings.

Trading in financial instruments on a centralized exchange can be traced back to 16th century Antwerp and Amsterdam. Debt instruments (promissory notes) and equities (shares in joint-stock West India companies) changed hands with the help of brokers who met clients and collected information to set prices. About two centuries later, the 1792 Buttonwood agreement gave a start to the New York Stock Exchange. Two key ingredients led to NYSE's success: consolidation of liquidity (due to being situated in the center of burgeoning American business life) and trust (NYSE limited its member circle to 24 brokers).

What has changed in financial markets since 1531 Antwerp or 1792 New York? First, financial instruments no longer exist in the physical form. What used to be a piece of paper certifying ownership of an asset, became an entry in an electronic database. Second, information travels much faster. It used to take carrier

pigeons to gain an information advantage, now it takes ultrafast fiber-optic networks. Third, intermediaries are maths-smart rather than people-smart. Broker-dealers used to derive their information advantage from personal interactions with buyers and sellers; now they derive it from sophisticated computer algorithms that analyze the data feeds faster than the rest of the market. Fourth, stock exchanges no longer rely on consolidation of liquidity as the key feature of their business model. Where a single monopolist stock exchange used to be standard, multiple competing venues emerged to deliver lower fees and faster technology, as markets are seamlessly consolidated via electronic networks.

6.1 What drives the changes in how people trade?

Technology, regulation, and innovation are three interrelated forces that drive the changes in trading. Given that all trading in the last two decades is done electronically, technological advancements mean faster data processing, and render market making a business of high-frequency traders. The modern market makers gain all the more advantage from being fast as market structure becomes more complex: providing liquidity across 13 US exchanges and over 40 alternative trading systems requires sophisticated algorithms and speed. Regulators enable this dynamic through setting the rules of the game in a way conducive to speed-based competition for liquidity provision. For example, in the US, the price-time priority and no trade-through rule dictate who gets a fill: the limit order that offers the best price (regardless of quantity), and if several offer the same price, it's the one that arrives earlier. Based on these rules, market makers that are fastest are more competitive.

6.1.1 How are market structure changes manifested in order-to-trade ratios?

Market makers are central to healthy financial markets. But the way their activity manifests itself has changed dramatically due to shifts in technology and regulation. As explained above, most modern market makers are high-frequency traders (HFTs). However, researchers hardly ever observe the trading strategy

(market making) or the latency differentials (a more direct measure of HFT) associated with a particular trader. Therefore, empirical papers proxy HFT activity by order-to-trade ratios. It is important to note that HFT strategies include, but are not limited to market making.

Order-to-trade ratios have been growing across stocks and ETFs in the past two decades. The question is, what does this trend mean? Is it beneficial to market quality or does it indicate that harmful fast trading has become more prevalent? Given that existing studies do not find that market quality deteriorates with higher order-to-trade ratios, it is of interest why that is the case. To relate order-to-trade ratios to non-harmful HFT activity (i.e., market making), Chapter 2 offers a simple theory model and tests it empirically.

The model recognizes the key features of modern liquidity provision, whereas market makers trade off the cost of information monitoring against the cost of trading at stale prices. In equilibrium, order-to-trade ratios are a function of monitoring intensity and the number of trading venues. The model-suggested drivers of order-to-trade ratios are supported in the data. Further, a calibration exercise suggests that on average, the current levels of order-to-trade ratios do not seem to warrant concern.

Across exchanges, order-to-trade ratios increase as markets fragment. On a single exchange, order-to-trade ratios are higher, if an exchange has lower market share. In the model, this effect arises due to scaling up of quoting (but not trading) activity as market makers post limit orders across venues. Therefore, if regulators levy charges in proportion to order-to-trade ratios (rather than market shares), those charges disadvantage smaller trading venues.

Overall, Chapter 2 enhances our understanding of HFT market making in fragmented markets. It shows theoretically and empirically that order-to-trade ratios are higher in more volatile conditions, for smaller stocks, and smaller trading venues. It also explains why securities with a clearly defined set of signals (e.g., exchange-traded funds, ETFs) have higher order-to-trade ratios, compared to stocks. The regulatory implication is that messaging taxes are likely to harm market making, and disproportionately so in more volatile markets, smaller stocks, smaller trading venues, and in ETFs compared to stocks. This is consistent with earlier research, which finds that market quality suffers when messaging taxes are imposed.

6.1.2 How is liquidity priced in new financial products?

Some features of financial markets have not changed between 16th century Amsterdam and 21st century Wall Street. Merchants of Venice valued the liquidity in West India joint stock companies. Equally, Wall Street hedge funds value the ability to trade their emerging markets exposure even when the underlying markets are closed. Liquidity, – broadly defined as the ability to trade a substantial position cheaply, quickly, and without moving prices, – is an elusive concept. Maureen O'Hara (1997) likens it to pornography, in that it is hard to define, but "you know it when you see it". It is undisputable that traders value liquidity, but just how much are they prepared to pay for it? This has been one of the cornerstone issues in market microstructure literature.

Chapter 3 approaches the value of liquidity from a new angle. It considers financial instruments that have started trading only in 1993, but grew to account for over 50% of overall US trading volumes in 2016. These instruments, called exchange-traded funds (ETFs), are akin to open-end mutual funds, except investors can buy and sell ETFs at any point during the trading day. As new ETFs emerge, a variety of index exposures become available to investors. Interestingly, only a handful of major ETFs account for most of the trading. Moreover, some of these hyper-liquid ETFs have direct competitors that track exactly the same index. One would expect that competition among identical investment products drives their fees to be equal. However, this is not the case. ETFs with exactly the same underlying stocks can charge very different management fees (a fixed percentage that gets deducted off the index return). Further, the more liquid the ETF, the greater the fee premium that it charges, compared to competitors. Thus, ETFs offer a unique laboratory to extract the liquidity differentials between otherwise identical products.

ETFs emerged as a financial innovation in response to the 1987 stock market crash. The idea came from the AMEX investment products team, after reading the SEC report on 1987 stock market crash. One of the suggestions in the report was to trade stocks as a basket. The S&P 500 ETF was born in 1993, and now it is the most frequently traded security in the world. In the early 2000s, other ETFs were launched, tracking major indices such as Russell, MSCI, and S&P. While ETFs provide a convenient way to invest for the long term, they are also suitable to trade index exposures. Institutional investors in particular use ETFs

for tactical allocations, portfolio transitions, hedging, and many other short-term liquidity-motivated purposes. As trading concentrated in a few ETFs favored by liquidity-motivated short-horizon investors, this dynamic led to a handful of ETFs becoming hyper-liquid.

Issuers do not have much control over ETF liquidity (except to a limited extent through creation-redemption fees), but they certainly benefit, if an ETF accumulates substantial liquidity. How so? Because more liquid ETFs can charge higher fees without losing their core clientele — the short-horizon investors. Chapter 3 formalizes this logic in a game theory model with Nash equilibrium. In equilibrium, high-fee ETFs can sustain their high fees, if they have a core clientele of fee-insensitive liquidity-hungry short-horizon investors. Competing low-fee ETFs can survive, if there are enough fee-sensitive long-horizon investors. What prevents the low-fee ETFs from winning over the lucrative short-horizon clientele is precisely liquidity. In a typical coordination problem, it is suboptimal for any individual short-horizon investor to switch to a low-fee ETF. The model shows how liquidity can create monopolistic profits for issuers of liquid ETFs.

The model of ETF liquidity from Chapter 3 directly estimates the welfare effects of hyper-liquid ETFs in oligopolistic markets. The results suggests that liquidity premia generate substantial welfare transfers from short-horizon ETF investors to issuers. Due to liquidity externalities, society as a whole experiences a welfare loss resulting from "overproduction" of the same investment product in multiple facets (i.e., same index exposure offered by multiple ETFs).

Overall, Chapter 3 emphasizes the "public good" nature of liquidity and demonstrates how liquidity externalities (akin to network externalities) result in oligopolistic fee-setting behaviour by issuers. The economic intuition of Nash equilibrium applies in the context of market microstructure, and especially so in the context of "public goods", such as liquidity and price discovery. Although in financial markets, it might be impractical to subsidize the producers of positive externalities and tax the producers of negative ones, at a minimum, regulators should recognize where externalities exist when introducing new market structure rules.

6.1.3 How informative are prices set by different trading mechanisms?

The workings of 19th century Paris Bourse inspired Leon Walras to develop a model of price formation that assumes the tatonnement process. In the model, multiple traders voice how much they are prepared to buy or sell at different price levels. The process — called a Walrasian auction — results in a single equilibrium price, at which demand meets supply. Although economic history literature has since disputed the historic accuracy of the tatonnement process in 19th century Paris Bourse, it is undisputable that call auctions, similar to those envisioned by Walras, play a pivotal role in modern markets.

Broadly, two different trading mechanisms exist — a continuous session, with multiple bilateral interactions between buyers and sellers, and a call auction, with a single multilateral interaction. Trading in auctions has increased substantially in recent years, partly because of growing popularity of passive funds, which use auction prices as benchmarks. At the same time, trading in the continuous session has thinned out, leading some to question the need for continuous trading altogether. The findings in Chapter 5 suggest that the continuous trading session remains the key driver of price discovery, despite volumes shifting towards closing auctions. Furthermore, closing prices tend to reflect less information over recent years, despite being formed by greater trading volumes.

To study the informativeness of prices set by different trading mechanisms (e.g., continuous trading vs auctions), Chapter 4 develops a price discovery methodology for sequential trading. The methodology relies on permanent-temporary variance decompositions from vector autoregressive models (VARs), but goes beyond existing studies in considering different levels of noise in different trading phases. Although motivated by the study of auction price informativeness, the methodology can be applied to multiple settings, in which a single security trades in different markets or time zones sequentially.

Overall, the findings from Chapter 5 imply that closing auctions are sought after for liquidity reasons, but do not contribute substantially to the information content of prices. This suggests that the shift to passive investing has real effects on price discovery: closing prices contribute little to price formation.

6.2 Work ahead

The finance profession is notorious for failing to predict any of the past financial crises. A deeper understanding of the statistics of random numbers, however, would suggest that any such prediction exercise is vain. This does not mean that academic finance is socially useless. To prepare one's house for adverse weather conditions, one should consider turning to an engineer rather than a psychic. Similarly, market microstructure research can help make "engineering choices" to help the financial system withstand poor weather and storms. Are regulatory rules conductive to fair and efficient markets? Can financial products have unintended consequences in adverse conditions? Are financial intermediaries well-capitalized? Are market participants' interactions changing over time? Asking these questions is on the research agenda for future work.

Future research could address one of the core issues in market microstructure: the lack of forward-looking regulations. Regulators, like academics, are often criticized for lagging behind the industry in addressing the structural changes in financial markets. It's only after market disruptions, the argument goes, that new regulations are put in place to address the problem. The effects of those regulations are then tested in academic papers, which adds another round of lags. In the meantime, financial markets move on to brew up an entirely different collection of challenges, often to catch regulators, once again, off-guard.

In the world of fast-to-market financial innovations, it is crucial to break the cycle of backward-looking regulations and lagging-behind academic papers. This thesis lays the groundwork for several directions of forward-looking research. Building on the model of order-to-trade ratios, future work could seek to improve detection of market manipulation activity, as opposed to legitimate market making. Building on the research of ETF liquidity, one direction is modelling the liquidity crash risk of ETFs. Another direction is addressing the lack of clarity on actively-managed ETF regulations. Building on the novel methodology of price discovery in sequential trading, new empirical applications can consider market linkages financial products that trade around the clock, including futures, commodities, and currencies. Finally, building on the existing work on closing auctions, potential next steps involve an international study of the interplay between regulatory frameworks (such as MiFID II) and trading activity in closing auctions.

6.2.1 Towards market surveillance 2.0

Spoofing and layering cases are on the rise around the world, affecting both advanced markets such as the US, and developing markets like Malaysia. Order-to-trade ratios have traditionally been used as an early warning indicator of suspicious activity. However, in modern high-frequency markets, it is non-trivial to discern the true causes of high order-to-trade ratios. Building on the model proposed in this thesis, future work can use training samples of market-making activity and market manipulation activity to calibrate the model as a market surveillance mechanism. With the appropriate training sample, the model can offer a powerful regulatory tool to monitor markets in real time, hence offering an advanced warning mechanism of financial misconduct.

6.2.2 Taking the pulse of ETF liquidity

Market participants have long voiced concerns about the effects of ETFs on liquidity in related instruments, particularly in underlying stocks. While such effects are mostly positive in normal market conditions, the effects in market downturns are unknown. A potential solution is to extend the model of market making in ETFs to encompass the multi-asset links between ETFs, the underlying basket, and the hedging instruments used by ETF market makers. As the effects of ETFs in market downturns still remain to be seen, the model can offer forward-looking insights into potential liquidity risks.

To monitor liquidity risks successfully, one has to consider the exact mechanisms through which illiquidity spirals can propagate. In adverse conditions, ETFs can only be as liquid as their underlying constituents. In other words, if market makers cannot hedge their exposure with related instruments, such as derivatives on the underlying index, ETF options, etc., ETFs become less liquid than the underlying stocks. Because multiple counterparties rely on ETF liquidity more than that of underlying stocks, increased demand for primary market liquidity can fuel illiquidity spirals in underlying instruments.

Another forward-looking model of ETF market making can consider actively-managed ETFs. Providing liquidity in these instruments introduces extreme adverse selection risk on external market makers, if ETF holdings are not disclosed.

ETF issuers of active products currently take several approaches: (i) disclose holdings with a time lag, employing external market makers, (ii) disclose holdings in real time, but employ an internal market making model. Comparing the ETF liquidity effects of these two approaches can offer insights for regulators, who currently do not have a clear framework for requirements on internal vs external market making.

6.2.3 Information and noise around the clock

Trading reveals information. Filtering out the noise from trading prices gives a better signal about changes in true price. How do we know which instruments / markets / trading mechanisms are better at uncovering that signal? The methodology developed in this thesis can help answer this question. While motivated by price discovery in closing auctions for equity trading, this methodology can be readily applied in other settings. For example, trading signals from forex and futures markets reflect information on macroeconomic news. The sequential price discovery methodology can help understand which markets contribute signals with greater information-to-noise ratios.

6.2.4 Closing auctions, MiFID II, and price informativeness

European markets experienced some of the most dramatic shifts of trading towards closing auctions. Why is it the case that Europe, for once, is ahead of the US in this shift? While this thesis studies the determinants of trading on close within a given market (Australia), it is of interest to examine the cross-country determinants of this shift. Such investigation would help us understand why some markets experience greater shifts towards the close, compared to others. The implications for price discovery might also differ, especially given that Europe's introduction of MiFID II affected multiple aspects of trade execution (e.g., limits on dark trading) and price discovery (e.g., significant cuts in equity research).

6.2.5 . . . the game without end

"The game of science is, in principle, without end. He who decides one day that scientific statements do not call for any further test, and that they can be regarded

as finally verified, retires from the game," wrote Karl Popper in his 1959 book "The Logic of Scientific Discovery". Market structure is continuously evolving, and therefore any empirical investigation of it is necessarily incomplete. This thesis offers but a glimpse of recent changes. Further work should offer insights for problems of the future.

Bibliography

Abner, D. J. (2013), *Visual guide to ETFs*, Vol. 580, John Wiley & Sons.

Admati, A. R. (1991), 'The informational role of prices: A review essay', *Journal of Monetary Economics* **28**(2), 347–361.

Admati, A. R. and Pfleiderer, P. (1988), 'A theory of intraday patterns: Volume and price variability', *Review of Financial Studies* **1**(1), 3–40.

Agapova, A. (2011), 'Conventional mutual index funds versus exchange-traded funds', *Journal of Financial Markets* **14**(2), 323–343.

Agarwal, V., Hanouna, P., Moussawi, R. and Stahel, C. W. (2018), 'Do ETFs increase the commonality in liquidity of underlying stocks?', Working paper.

Aitken, M., Chen, H. and Foley, S. (2017), 'The impact of fragmentation, exchange fees and liquidity provision on market quality', *Journal of Empirical Finance* **41**, 140–160.

Aitken, M., Comerton-Forde, C. and Frino, A. (2005), 'Closing call auctions and liquidity', *Accounting & Finance* **45**(4), 501–518.

AMF (2019), 'Growing importance of the closing auction in share trading volumes', `https://www.amf-france.org/technique/multimedia?docId=workspace://SpacesStore/a16f34e9-6cae-4bd3-adc6-0fb96a559b9e_en_1.0_rendition`. [Online; accessed 20-October-2019].

Amihud, Y. (2002), 'Illiquidity and stock returns: cross-section and time-series effects', *Journal of Financial Markets* **5**(1), 31–56.

Amihud, Y. and Mendelson, H. (1980), 'Dealership market: Market-making with inventory', *Journal of Financial Economics* **8**(1), 31–53.

Amihud, Y. and Mendelson, H. (1986), 'Asset pricing and the bid-ask spread', *Journal of Financial Economics* **17**(2), 223–249.

Amihud, Y. and Mendelson, H. (1987), 'Trading mechanisms and stock returns: An empirical investigation', *Journal of Finance* **42**(3), 533–553.

Amihud, Y. and Mendelson, H. (1991), 'Liquidity, maturity, and the yields on US treasury securities', *Journal of Finance* **46**(4), 1411–1425.

Amihud, Y., Mendelson, H. and Murgia, M. (1990), 'Stock market microstructure and return volatility: Evidence from Italy', *Journal of Banking & Finance* **14**(2-3), 423–440.

Anadu, K., Kruttli, M., McCabe, P. E., Osambela, E. and Shin, C. (2019), 'The shift from active to passive investing: Potential risks to financial stability?', Working paper.

Anand, A., Irvine, P., Puckett, A. and Venkataraman, K. (2011), 'Performance of institutional trading desks: An analysis of persistence in trading costs', *Review of Financial Studies* **25**(2), 557–598.

Anand, A. and Subrahmanyam, A. (2008), 'Information and the intermediary: Are market intermediaries informed traders in electronic markets?', *Journal of Financial and Quantitative Analysis* **43**(1), 1–28.

Anand, A. and Venkataraman, K. (2016), 'Market conditions, fragility, and the economics of market making', *Journal of Financial Economics* **121**(2), 327–349.

Appel, I. R., Gormley, T. A. and Keim, D. B. (2016), 'Passive investors, not passive owners', *Journal of Financial Economics* **121**(1), 111–141.

Argenziano, R. (2008), 'Differentiated networks: Equilibrium and efficiency', *The RAND Journal of Economics* **39**(3), 747–769.

ASIC (2012), 'ASIC Market Supervision Update Issue 45', `https://www.asic.gov.au/about-asic/corporate-publications/newsletters/asic-market-supervision-update/asic-market-supervision-update-previous-issues/asic-market-supervision-update-issue-45/`.

Asparouhova, E., Bessembinder, H. and Kalcheva, I. (2013), 'Noisy prices and inference regarding returns', *Journal of Finance* **68**(2), 665–714.

Ates, A. and Wang, G. H. (2005), 'Information transmission in electronic versus open-outcry trading systems: An analysis of US equity index futures markets', *Journal of Futures Markets* **25**(7), 679–715.

Bacidore, J. M. and Lipson, M. L. (2001), 'The effects of opening and closing procedures on the NYSE and NASDAQ', Working paper.

Baillie, R. T., Booth, G. G., Tse, Y. and Zabotina, T. (2002), 'Price discovery and common factor models', *Journal of Financial Markets* **5**(3), 309–321.

Balchunas, E. (2016), *The institutional ETF toolbox: How institutions can understand and utilize the fast-growing world of ETFs*, John Wiley & Sons.

Baldauf, M. and Mollner, J. (2016), 'Fast traders make a quick buck: The role of speed in liquidity provision', Working paper.

Ball, C. A. and Chordia, T. (2001), 'True spreads and equilibrium prices', *Journal of Finance* **56**(5), 1801–1835.

Barberis, N. and Thaler, R. (2003), 'A survey of behavioral finance', *Handbook of the Economics of Finance* **1**, 1053–1128.

Barclay, M. J., Hendershott, T. and McCormick, D. T. (2003), 'Competition among trading venues: Information and trading on electronic communications networks', *Journal of Finance* **58**(6), 2637–2665.

Barclay, M. J. and Warner, J. B. (1993), 'Stealth trading and volatility: Which trades move prices?', *Journal of Financial Economics* **34**(3), 281–305.

Battalio, R. H. (1997), 'Third market broker-dealers: Cost competitors or cream skimmers?', *Journal of Finance* **52**(1), 341–352.

Battig, C. and Chelley-Steeley, P. L. (2010), 'The impact of the closing call auction: an examination of effects in london', *Applied Financial Economics* **20**(4), 303–315.

Ben-David, I., Franzoni, F. and Moussawi, R. (2018), 'Do ETFs increase volatility?', *Journal of Finance* **73**(6), 2471–2535.

Bertsimas, D. and Lo, A. W. (1998), 'Optimal control of execution costs', *Journal of Financial Markets* **1**(1), 1–50.

Bhattacharya, A. and O'Hara, M. (2018), 'Can ETFs increase market fragility? effect of information linkages in ETF markets', Working paper.

Biais, B., Foucault, T. and Moinas, S. (2011), 'Equilibrium high-frequency trading', Working paper.

Biais, B., Glosten, L. and Spatt, C. (2005), 'Market microstructure: A survey of microfoundations, empirical results, and policy implications', *Journal of Financial Markets* **8**(2), 217–264.

Biais, B., Hillion, P. and Spatt, C. (1999), 'Price discovery and learning during the preopening period in the Paris Bourse', *Journal of Political Economy* **107**(6), 1218–1248.

Biais, B. and Pouget, S. (2000), 'Microstructure, incentives, and the discovery of equilibrium in experimental financial markets', Working paper.

Black, F. (1986), 'Noise', *Journal of Finance* **41**(3), 528–543.

Blume, M. E. and Stambaugh, R. F. (1983), 'Biases in computed returns: An application to the size effect', *Journal of Financial Economics* **12**(3), 387–404.

Boehmer, E., Fong, K. Y. and Wu, J. (2013), 'Algorithmic trading and changes in firms' equity capital', Working paper.

Boehmer, E., Jennings, R. and Wei, L. (2006), 'Public disclosure and private decisions: Equity market execution quality and order routing', *Review of Financial Studies* **20**(2), 315–358.

Boehmer, E., Li, D. and Saar, G. (2018), 'The competitive landscape of high-frequency trading firms', *Review of Financial Studies* **31**(6), 2227–2276.

Bogousslavsky, V. (2016), 'Infrequent rebalancing, return autocorrelation, and seasonality', *Journal of Finance* **71**(6), 2967–3006.

Booth, G. G., So, R. W. and Tse, Y. (1999), 'Price discovery in the German equity index derivatives markets', *Journal of Futures Markets* **19**(6), 619–643.

Boulatov, A., Bernhardt, D. and Larionov, I. (2016), 'Predatory and defensive trading in a dynamic model of optimal execution by multiple traders', Working paper.

Boussetta, S., Daures Lescourret, L. and Moinas, S. (2017), 'The role of pre-opening mechanisms in fragmented markets', Working paper.

Box, T., Davis, R. L. and Fuller, K. P. (2018), 'ETF competition and market quality', *Financial Management* .

Branch, B. S. and Ma, A. J. (2012), 'Overnight return, the invisible hand behind intraday returns?', *Journal of Applied Finance* **22**(2).

Brennan, M. J. and Subrahmanyam, A. (1996), 'Market microstructure and asset pricing: On the compensation for illiquidity in stock returns', *Journal of Financial Economics* **41**(3), 441–464.

Brogaard, J. and Garriott, C. (2019), 'High-frequency trading competition', *Journal of Financial and Quantitative Analysis* **54**(4), 1469–1497.

Brogaard, J., Hagströmer, B., Nordén, L. and Riordan, R. (2015), 'Trading fast and slow: Colocation and liquidity', *Review of Financial Studies* **28**(12), 3407–3443.

Brogaard, J., Hendershott, T. and Riordan, R. (2014), 'High-frequency trading and price discovery', *Review of Financial Studies* **27**(8), 2267–2306.

Brogaard, J., Nguyen, T. H., Putnins, T. J. and Wu, E. (2019), 'What moves stock prices? The role of news, noise, and information', *Review of Financial Studies, Forthcoming* .

Brunnermeier, M. K. and Pedersen, L. H. (2005), 'Predatory trading', *Journal of Finance* **60**(4), 1825–1863.

Budish, E., Cramton, P. and Shim, J. (2015), 'The high-frequency trading arms race: Frequent batch auctions as a market design response', *The Quarterly Journal of Economics* **130**(4), 1547–1621.

Business Insider (2018), 'Fidelity slashes fees as funds battle for investors', `https://www.businessinsider.com/ap-fidelity-slashes-fees-as-funds-battle-for-investors-2018-8/?r=AU&IR=T`. [Online; accessed 02-August-2018].

Buti, S., Rindi, B. and Werner, I. M. (2017), 'Dark pool trading strategies, market quality and welfare', *Journal of Financial Economics* **124**(2), 244–265.

Cabrera, J., Wang, T. and Yang, J. (2009), 'Do futures lead price discovery in electronic foreign exchange markets?', *Journal of Futures Markets* **29**(2), 137–156.

Cai, C. X., Hudson, R. and Keasey, K. (2004), 'Intraday bid-ask spreads, trading volume and volatility: Recent empirical evidence from the London Stock Exchange', *Journal of Business Finance & Accounting* **31**(5-6), 647–676.

Caivano, V., Ciccarelli, S., Di Stefano, G., Fratini, M., Gasparri, G., Giliberti, M., Linciano, N. and Tarola, I. (2012), 'High frequency trading: Definition, effects, policy issues', Working paper.

Campbell, J. Y. (1991), 'A variance decomposition for stock returns', *The Economic Journal* **101**(405), 157–179.

Campbell, J. Y. and Shiller, R. J. (1988), 'Interpreting cointegrated models', *Journal of Economic Dynamics and Control* **12**(2-3), 505–522.

Cao, C., Ghysels, E. and Hatheway, F. (2000), 'Price discovery without trading: Evidence from the NASDAQ preopening', *Journal of Finance* **55**(3), 1339–1365.

Cao, C., Hansch, O. and Wang, X. (2009), 'The information content of an open limit-order book', *Journal of Futures Markets* **29**(1), 16–41.

Capelle-Blancard, G. (2017), 'Curbing the growth of stock trading? Order-to-trade ratios and financial transaction taxes', *Journal of International Financial Markets, Institutions and Money* **49**, 48–73.

Chakravarty, S., Gulen, H. and Mayhew, S. (2004), 'Informed trading in stock and option markets', *Journal of Finance* **59**(3), 1235–1257.

Chen, H., Choi, P. M. S. and Hong, Y. (2013), 'How smooth is price discovery? Evidence from cross-listed stock trading', *Journal of International Money and Finance* **32**, 668–699.

Chinco, A. and Fos, V. (2019), 'The sound of many funds rebalancing', Working paper.

Chordia, T. (1996), 'The structure of mutual fund charges', *Journal of Financial Economics* **41**(1), 3–39.

Chordia, T., Roll, R. and Subrahmanyam, A. (2000), 'Commonality in liquidity', *Journal of Financial Economics* **56**(1), 3–28.

Chordia, T., Roll, R. and Subrahmanyam, A. (2008), 'Liquidity and market efficiency', *Journal of Financial Economics* **87**(2), 249–268.

Chowdhry, B. and Nanda, V. (1991), 'Multimarket trading and market liquidity', *Review of Financial Studies* **4**(3), 483–511.

Cliff, M., Cooper, M. J. and Gulen, H. (2008), 'Return differences between trading and non-trading hours: Like night and day', Working paper.

CNBC (2019), 'Vanguard Group eliminates trading fees on almost all ETFs — including funds from most of its rivals', `https://www.icifactbook.org/`. [Online; accessed 02-July-2019].

Colliard, J.-E. and Foucault, T. (2012), 'Trading fees and efficiency in limit order markets', *Review of Financial Studies* **25**(11), 3389–3421.

Colliard, J.-E. and Hoffmann, P. (2017), 'Financial transaction taxes, market composition, and liquidity', *Journal of Finance* **72**(6), 2685–2716.

Comerton-Forde, C. and Putnins, T. J. (2013), 'Stock price manipulation: Prevalence and determinants', *Review of Finance* **18**(1), 23–66.

Comerton-Forde, C. and Rydge, J. (2006), 'Call auction algorithm design and market manipulation', *Journal of Multinational Financial Management* **16**(2), 184–198.

Committee on Capital Markets Regulation (2016), 'The US equity markets. A plan for regulatory reform.', `https://www.capmktsreg.org/wp-content/uploads/2018/10/The-US-Equity-Markets.pdf`.

Conrad, J., Wahal, S. and Xiang, J. (2015), 'High-frequency quoting, trading, and the efficiency of prices', *Journal of Financial Economics* **116**(2), 271–291.

Copeland, T. E. and Galai, D. (1983), 'Information effects on the bid-ask spread', *Journal of Finance* **38**(5), 1457–1469.

Cordi, N., Foley, S. and Putnins, T. J. (2015), 'Is there an optimal closing mechanism?', Working paper.

Covrig, V., Ding, D. K. and Low, B. S. (2004), 'The contribution of a satellite market to price discovery: Evidence from the Singapore exchange', *Journal of Futures Markets* **24**(10), 981–1004.

Cushing, D. and Madhavan, A. (2000), 'Stock returns and trading at the close', *Journal of Financial Markets* **3**(1), 45–67.

Da, Z. and Shive, S. (2018), 'Exchange traded funds and asset return correlations', *European Financial Management* **24**(1), 136–168.

Dahlström, P., Hagströmer, B. and Nordén, L. L. (2018), 'Determinants of limit order cancellations', Working paper.

Dannhauser, C. D. (2017), 'The impact of innovation: Evidence from corporate bond exchange-traded funds', *Journal of Financial Economics* **125**(3), 537–560.

De Jong, F. (2002), 'Measures of contributions to price discovery: A comparison', *Journal of Financial Markets* **5**(3), 323–327.

Degryse, H., De Jong, F. and van Kervel, V. (2015), 'The impact of dark trading and visible fragmentation on market quality', *Review of Finance* **19**(4), 1587–1622.

Degryse, H., Karagiannis, N., Tombeur, G. and Wuyts, G. (2018), 'Two shades of opacity: Hidden orders versus dark trading'.

Dobrev, D. and Schaumburg, E. (2017), 'High-frequency cross-market trading: Model free measurement and applications', *Perspectives* .

Doidge, C., Karolyi, G. A. and Stulz, R. M. (2017), 'The US listing gap', *Journal of Financial Economics* **123**(3), 464–487.

Duffie, D., Gârleanu, N. and Pedersen, L. H. (2005), 'Over-the-counter markets', *Econometrica* **73**(6), 1815–1847.

Easley, D., Kiefer, N. M., O'Hara, M. and Paperman, J. B. (1996), 'Liquidity, information, and infrequently traded stocks', *Journal of Finance* **51**(4), 1405–1436.

Easley, D., Michayluk, D., O'Hara, M. and Putnins, T. J. (2020), 'The active world of passive investing', Working paper.

Economides, N. (1996), 'The economics of networks', *International Journal of Industrial Organization* **14**(6), 673–699.

Edelen, R. M. (1999), 'Investor flows and the assessed performance of open-end mutual funds', *Journal of Financial Economics* **53**(3), 439–466.

Egginton, J. F., Van Ness, B. F. and Van Ness, R. A. (2016), 'Quote stuffing', *Financial Management* **45**(3), 583–608.

Ellul, A., Shin, H. S. and Tonks, I. (2005), 'Opening and closing the market: Evidence from the London Stock Exchange', *Journal of Financial and Quantitative Analysis* **40**(4), 779–801.

Financial Times (2017), 'ETFs are eating the US stock market', `https://www.ft.com/content/6dabad28-e19c-11e6-9645-c9357a75844a`. [Online; accessed 25-January-2017].

Financial Times (2018), 'The 30 minutes that have an outsized role in US stock trading', `https://www.ft.com/content/9e1f05b4-43e7-11e8-803a-295c97e6fd0b`. [Online; accessed 25-April-2018].

Foley, S. and Putnins, T. J. (2016), 'Should we be afraid of the dark? Dark trading and market quality', *Journal of Financial Economics* **122**(3), 456–481.

Fong, K. and Zurbruegg, R. (2003), 'How much do locals contribute to the price discovery process?', *Journal of Empirical Finance* **10**(3), 305–320.

Forte, S. and Pena, J. I. (2009), 'Credit spreads: An empirical analysis on the informational content of stocks, bonds, and CDS', *Journal of Banking & Finance* **33**(11), 2013–2025.

Foucault, T., Kadan, O. and Kandel, E. (2005), 'Limit order book as a market for liquidity', *Review of Financial Studies* **18**(4), 1171–1217.

Foucault, T., Kadan, O. and Kandel, E. (2013), 'Liquidity cycles and make/take fees in electronic markets', *Journal of Finance* **68**(1), 299–341.

Foucault, T., Kozhan, R. and Tham, W. W. (2016), 'Toxic arbitrage', *Review of Financial Studies* **30**(4), 1053–1094.

Foucault, T. and Menkveld, A. J. (2008), 'Competition for order flow and smart order routing systems', *Journal of Finance* **63**(1), 119–158.

Foucault, T., Röell, A. and Sandås, P. (2003), 'Market making with costly monitoring: An analysis of the SOES controversy', *Review of Financial Studies* **16**(2), 345–384.

Frazzini, A., Israel, R. and Moskowitz, T. J. (2018), 'Trading costs', Working paper.

French, K. R. and Roll, R. (1986), 'Stock return variances: The arrival of information and the reaction of traders', *Journal of Financial Economics* **17**(1), 5–26.

Fricke, C. and Menkhoff, L. (2011), 'Does the "Bund" dominate price discovery in Euro bond futures? Examining information shares', *Journal of Banking & Finance* **35**(5), 1057–1072.

Friederich, S. and Payne, R. (2015), 'Order-to-trade ratios and market liquidity', *Journal of Banking & Finance* **50**, 214–223.

Friewald, N., Jankowitsch, R. and Subrahmanyam, M. G. (2012), 'Illiquidity or credit deterioration: A study of liquidity in the US corporate bond market during financial crises', *Journal of Financial Economics* **105**(1), 18–36.

Frijns, B., Gilbert, A. and Tourani-Rad, A. (2010), 'The dynamics of price discovery for cross-listed shares: Evidence from Australia and New Zealand', *Journal of Banking & Finance* **34**(3), 498–508.

Garman, M. B. (1976), 'Market microstructure', *Journal of Financial Economics* **3**(3), 257–275.

Glosten, L., Nallareddy, S. and Zou, Y. (2016), 'ETF trading and informational efficiency of underlying securities', Working paper.

Glosten, L. R. and Milgrom, P. R. (1985), 'Bid, ask and transaction prices in a specialist market with heterogeneously informed traders', *Journal of Financial Economics* **14**(1), 71–100.

Goldstein, M. A., Kwan, A. and Philip, R. (2018), 'High-frequency trading strategies', Working paper.

Gomber, P. and Haferkorn, M. (2015), High frequency trading, *in* 'Encyclopedia of Information Science and Technology, Third Edition', IGI Global, pp. 1–9.

Gomber, P., Sagade, S., Theissen, E., Weber, M. C. and Westheide, C. (2016), 'Spoilt for choice: Order routing decisions in fragmented equity markets', Working paper.

Gomber, P., Sagade, S., Theissen, E., Weber, M. C. and Westheide, C. (2017), 'Competition between equity markets: A review of the consolidation versus fragmentation debate', *Journal of Economic Surveys* **31**(3), 792–814.

Gonzalo, J. and Granger, C. (1995), 'Estimation of common long-memory components in cointegrated systems', *Journal of Business & Economic Statistics* **13**(1), 27–35.

Grossman, S. J. and Miller, M. H. (1988), 'Liquidity and market structure', *Journal of Finance* **43**(3), 617–633.

Hagströmer, B. and Menkveld, A. J. (2017), 'A network map of information percolation', Working paper.

Hamm, S. J. (2014), 'The effect of ETFs on stock liquidity', Working paper.

Harris, L. E. (1993), 'Consolidation, fragmentation, segmentation, and regulation', *Journal of Finance* **48**(3), 1092–1093.

Harris, L. and Gurel, E. (1986), 'Price and volume effects associated with changes in the S&P 500 list: New evidence for the existence of price pressures', *Journal of Finance* **41**(4), 815–829.

Hasbrouck, J. (1991*a*), 'Measuring the information content of stock trades', *Journal of Finance* **46**(1), 179–207.

Hasbrouck, J. (1991*b*), 'The summary informativeness of stock trades: An econometric analysis', *Review of Financial Studies* **4**(3), 571–595.

Hasbrouck, J. (1993), 'Assessing the quality of a security market: A new approach to transaction-cost measurement', *Review of Financial Studies* **6**(1), 191–212.

Hasbrouck, J. (1995), 'One security, many markets: Determining the contributions to price discovery', *Journal of Finance* **50**(4), 1175–1199.

Hasbrouck, J. (2018), 'High-frequency quoting: Short-term volatility in bids and offers', *Journal of Financial and Quantitative Analysis* **53**(2), 613–641.

Hasbrouck, J. (2019), 'Rejoinder on: Price discovery in high resolution', *Journal of Financial Econometrics* .

Hasbrouck, J. and Saar, G. (2013), 'Low-latency trading', *Journal of Financial Markets* **16**(4), 646–679.

Hasbrouck, J. and Seppi, D. J. (2001), 'Common factors in prices, order flows, and liquidity', *Journal of Financial Economics* **59**(3), 383–411.

Haslag, P. H. and Ringgenberg, M. (2017), 'The causal impact of market fragmentation on liquidity', Working paper.

Heath, D., Macciocchi, D., Michaely, R. and Ringgenberg, M. (2020), 'Do index funds monitor?', Working paper.

Hendershott, T. and Jones, C. M. (2005), 'Island goes dark: Transparency, fragmentation, and regulation', *Review of Financial Studies* **18**(3), 743–793.

Hendershott, T., Jones, C. M. and Menkveld, A. J. (2011), 'Does algorithmic trading improve liquidity?', *Journal of Finance* **66**(1), 1–33.

Hendershott, T., Livdan, D. and Rösch, D. (2019), 'Asset pricing: A tale of night and day', Working paper.

Hendershott, T. and Mendelson, H. (2000), 'Crossing networks and dealer markets: Competition and performance', *Journal of Finance* **55**(5), 2071–2115.

Hendershott, T. and Menkveld, A. J. (2014), 'Price pressures', *Journal of Financial Economics* **114**(3), 405–423.

Heston, S. L., Korajczyk, R. A. and Sadka, R. (2010), 'Intraday patterns in the cross-section of stock returns', *Journal of Finance* **65**(4), 1369–1407.

Ho, T. and Stoll, H. R. (1981), 'Optimal dealer pricing under transactions and return uncertainty', *Journal of Financial Economics* **9**(1), 47–73.

Hoffmann, P. (2014), 'A dynamic limit order market with fast and slow traders', *Journal of Financial Economics* **113**(1), 156–169.

Holden, C. W., Jacobsen, S. E. and Subrahmanyam, A. (2014), 'The empirical analysis of liquidity', *Foundations and Trends in Finance* **8**, 263–365.

Hortaçsu, A. and Syverson, C. (2004), 'Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds', *The Quarterly Journal of Economics* **119**(2), 403–456.

Huang, R. (2000), 'Price discovery by ECNs and NASDAQ market makers', Working paper.

Huang, S., O'Hara, M. and Zhong, Z. (2019), 'Innovation and informed trading: Evidence from industry ETFs', Working paper.

Huberman, G. and Halka, D. (2001), 'Systematic liquidity', *Journal of Financial Research* **24**(2), 161–178.

Investment Company Institute (2019), '2019 Investment Company Fact Book', `https://www.cnbc.com/2018/07/02/vanguard-slashing-costs-on-nearly-all-etfs-even-rival-schwab.html`. [Online; accessed 05-July-2019].

Israeli, D., Lee, C. M. and Sridharan, S. A. (2017), 'Is there a dark side to exchange traded funds? An information perspective', *Review of Accounting Studies* **22**(3), 1048–1083.

Jegadeesh, N. (1990), 'Evidence of predictable behavior of security returns', *Journal of Finance* **45**(3), 881–898.

Johann, T., Putnins, T. J., Sagade, S. and Westheide, C. (2019), 'Quasi-dark trading: The effects of banning dark pools in a world of many alternatives'.

Jones, C. M. (2013), 'What do we know about high-frequency trading?', Working paper.

Jørgensen, K., Skjeltorp, J. and Ødegaard, B. A. (2018), 'Throttling hyperactive robots — Order-to-trade ratios at the Oslo Stock Exchange', *Journal of Financial Markets* **37**, 1–16.

Kandel, E., Rindi, B. and Bosetti, L. (2012), 'The effect of a closing call auction on market quality and trading strategies', *Journal of Financial Intermediation* **21**(1), 23–49.

Katz, M. L. and Shapiro, C. (1985), 'Network externalities, competition, and compatibility', *American Economic Review* **75**(3), 424–440.

Korajczyk, R. A. and Murphy, D. (2018), 'High-frequency market making to large institutional trades', *Review of Financial Studies* **32**(3), 1034–1067.

Körber, L., Linton, O. B. and Vogt, M. (2013), 'The effect of fragmentation in trading on market quality in the UK equity market', Working paper.

Krause, T., Ehsani, S. and Lien, D. (2014), 'Exchange-traded funds, liquidity and volatility', *Applied Financial Economics* **24**(24), 1617–1630.

Krishnamurthy, A. (2002), 'The bond/old-bond spread', *Journal of Financial Economics* **66**(2-3), 463–506.

Kurov, A. and Lasser, D. J. (2004), 'Price dynamics in the regular and e-mini futures markets', *Journal of Financial and Quantitative Analysis* **39**(2), 365–384.

Kurzweil, R. (2005), *The singularity is near: When humans transcend biology*, Penguin.

Kyle, A. S. (1985), 'Continuous auctions and insider trading', *Econometrica* **15**(35), 1315–1335.

Lehmann, B. N. (1990), 'Fads, martingales, and market efficiency', *The Quarterly Journal of Economics* **105**(1), 1–28.

Lehmann, B. N. (2002), 'Some desiderata for the measurement of price discovery across markets', *Journal of Financial Markets* **5**(3), 259–276.

Lepone, A. and Sacco, A. (2013), 'The impact of message traffic regulatory restrictions on market quality: Evidence from Chi-X Canada', Working paper.

Lettau, M. and Madhavan, A. (2018), 'Exchange-traded funds 101 for economists', *Journal of Economic Perspectives* **32**(1), 135–54.

Levine, R. (2005), 'Finance and growth: Theory and evidence', *Handbook of Economic Growth* **1**, 865–934.

Li, F. W. and Zhu, Q. (2016), 'Synthetic shorting with ETFs', Working paper.

Li, W. (2019), 'High frequency trading with speed hierarchies', Working paper.

Liu, W.-M. (2009), 'Monitoring and limit order submission risks', *Journal of Financial Markets* **12**(1), 107–141.

Lockwood, L. J. and Linn, S. C. (1990), 'An examination of stock market return volatility during overnight and intraday periods, 1964–1989', *Journal of Finance* **45**(2), 591–601.

Longstaff, F. A., Mithal, S. and Neis, E. (2005), 'Corporate yield spreads: Default risk or liquidity? New evidence from the credit default swap market', *Journal of Finance* **60**(5), 2213–2253.

Lou, D., Polk, C. and Skouras, S. (2019), 'A tug of war: Overnight versus intraday expected returns', *Journal of Financial Economics* **134**(1), 192–213.

Lyle, M. R. and Naughton, J. P. (2018), 'How does algorithmic trading improve market quality?', Working paper.

Madhavan, A. (1995), 'Consolidation, fragmentation, and the disclosure of trading information', *Review of Financial Studies* **8**(3), 579–603.

Madhavan, A. N. (2016), *Exchange-traded funds and the new dynamics of investing*, Oxford University Press.

Madhavan, A. and Sobczyk, A. (2016), 'Price dynamics and liquidity of exchange-traded funds', *Journal of Investment Management* **14**(2), 1–17.

Mahanti, S., Nashikkar, A., Subrahmanyam, M., Chacko, G. and Mallik, G. (2008), 'Latent liquidity: A new measure of liquidity, with an application to corporate bonds', *Journal of Financial Economics* **88**(2), 272–298.

Malamud, S. (2016), 'A dynamic equilibrium model of ETFs', Working paper.

Malceniece, L., Malcenieks, K. and Putnins, T. J. (2019), 'High frequency trading and comovement in financial markets', *Journal of Financial Economics* .

Malinova, K., Park, A. and Riordan, R. (2016), 'Taxing high frequency market making: Who pays the bill', Working paper.

Marshall, B. R., Nguyen, N. H. and Visaltanachoti, N. (2013), 'ETF arbitrage: Intraday evidence', *Journal of Banking and Finance* **37**(9), 3486–3498.

McInish, T. H. and Wood, R. A. (1990), 'An analysis of transactions data for the Toronto Stock Exchange: Return patterns and end-of-the-day effect', *Journal of Banking & Finance* **14**(2-3), 441–458.

Medrano, L. A. and Vives, X. (2001), 'Strategic behavior and price discovery', *RAND Journal of Economics* pp. 221–248.

Mendelson, H. (1987), 'Consolidation, fragmentation, and market performance', *Journal of Financial and Quantitative Analysis* **22**(2), 189–207.

Menkveld, A. J. (2013), 'High frequency trading and the new market makers', *Journal of Financial Markets* **16**(4), 712–740.

Menkveld, A. J. (2014), 'High-frequency traders and market structure', *Financial Review* **49**(2), 333–344.

Menkveld, A. J. (2016), 'The economics of high-frequency trading: Taking stock', *Annual Review of Financial Economics* **8**, 1–24.

Menkveld, A. J., Yueshen, B. Z. and Zhu, H. (2017), 'Shades of darkness: A pecking order of trading venues', *Journal of Financial Economics* **124**(3), 503–534.

Morck, R., Yeung, B. and Yu, W. (2000), 'The information content of stock markets: Why do emerging markets have synchronous stock price movements?', *Journal of Financial Economics* **58**(1-2), 215–260.

Morris, S. and Shin, H. S. (2006), 'Heterogeneity and uniqueness in interaction', *The Economy As an Evolving Complex System, III: Current Perspectives and Future Directions* **3**, 207.

Muravyev, D., Pearson, N. D. and Broussard, J. P. (2013), 'Is there price discovery in equity options?', *Journal of Financial Economics* **107**(2), 259–283.

Nagel, S. (2012), 'Evaporating liquidity', *Review of Financial Studies* **25**(7), 2005–2039.

Narayan, S. and Smyth, R. (2015), 'The financial econometrics of price discovery and predictability', *International Review of Financial Analysis* **42**, 380–393.

Nash, J. (1951), 'Non-cooperative games', *Annals of Mathematics* pp. 286–295.

O'Hara, M. (1997), *Market microstructure theory*, Wiley.

O'Hara, M. and Ye, M. (2011), 'Is market fragmentation harming market quality?', *Journal of Financial Economics* **100**(3), 459–474.

O'Hara, M. (2014), 'High-frequency trading and its impact on markets', *Financial Analysts Journal* **70**(3), 18–27.

O'Hara, M. (2015), 'High frequency market microstructure', *Journal of Financial Economics* **116**(2), 257–270.

Pagano, M. (1989), 'Trading volume and asset liquidity', *The Quarterly Journal of Economics* **104**(2), 255–274.

Pagano, M. and Röell, A. (1996), 'Transparency and liquidity: A comparison of auction and dealer markets with informed trading', *Journal of Finance* **51**(2), 579–611.

Pagano, M. S., Peng, L. and Schwartz, R. A. (2013), 'A call auction's impact on price formation and order routing: Evidence from the NASDAQ stock market', *Journal of Financial Markets* **16**(2), 331–361.

Pagano, M. S. and Schwartz, R. A. (2003), 'A closing call's impact on market quality at Euronext Paris', *Journal of Financial Economics* **68**(3), 439–484.

Pagnotta, E. S. and Philippon, T. (2018), 'Competing on speed', *Econometrica* **86**(3), 1067–1115.

Pascual, R., Pascual-Fuster, B. and Climent, F. (2006), 'Cross-listing, price discovery and the informativeness of the trading process', *Journal of Financial Markets* **9**(2), 144–161.

Popper, K. (1959), *The logic of scientific discovery*, Routledge.

Putnins, T. J. (2013), 'What do price discovery metrics really measure?', *Journal of Empirical Finance* **23**, 68–83.

Putnins, T. J. and Barbara, J. (2020), 'Heterogeneity in how algorithmic traders impact institutional trading costs', Working paper.

Quartz (2013), '96.8% of trades placed in the US stock market are cancelled.', `https://qz.com/133695/96-8-of-trades-placed-in-the-us-stock-market-are-cancelled/`. [Online; accessed 05-July-2017].

Reuters (2019), 'Last orders: Rise of closing auctions stirs worries in European stock markets', `https://fr.reuters.com/article/idUSKCN1V70M8`. [Online; accessed 18-August-2019].

Rosu, I., Sojli, E. and Tham, W. W. (2020), 'Quoting activity and the cost of capital', Working paper.

Schnitzlein, C. R. (1996), 'Call and continuous trading mechanisms under asymmetric information: An experimental investigation', *Journal of Finance* **51**(2), 613–636.

Shleifer, A. and Vishny, R. W. (1997), 'The limits of arbitrage', *Journal of Finance* **52**(1), 35–55.

Stoll, H. R. (1978), 'The supply of dealer services in securities markets', *Journal of Finance* **33**(4), 1133–1151.

Subrahmanyam, A. and Zheng, H. (2016), 'Limit order placement by high-frequency traders', *Borsa Istanbul Review* **16**(4), 185–209.

Tabb Forum (2019), 'MiFID II in September: All eyes on auctions', `https://tabbforum.com/opinions/mifid-ii-in-september-all-eyes-on-auctions/`. [Online; accessed 19-October-2019].

The Economist (2019), 'The stockmarket is now run by computers, algorithms and passive managers', Print edition, October 9, 2019.

The Trade (2018), 'Liquidity is for Closers', `https://www.thetradenews.com/thought-leadership/liquidity-is-for-closers/`. [Online; accessed 25-September-2018].

The Wall Street Journal (2018), 'Market cheats getting caught in record numbers.', `https://www.wsj.com/articles/u-s-market-manipulation-cases-reach-record-1540983720`. [Online; accessed 02-September-2019].

Theissen, E. and Westheide, C. (2017), Call of duty: Designated market maker participation in call auctions, Technical report.

US Securities and Exchange Commission (2013), 'US SEC Market Structure research highlights', `https://www.sec.gov/marketstructure/research/highlight-2013-01.html#.XkIlWGgzaUk`. [Online; accessed 12-July-2017].

US Securities and Exchange Commission (2019), 'Release Nos. 33-10695; IC-33646; File No. S7-15-18', https://www.sec.gov/rules/final/2019/33-10695.pdf. [Online; accessed 23-December-2019].

Van Kervel, V. (2015), 'Competition for order flow with fast and slow traders', *Review of Financial Studies* **28**(7), 2094–2127.

Van Kervel, V. and Menkveld, A. J. (2019), 'High-frequency trading around large institutional orders', *Journal of Finance* **74**(3), 1091–1137.

Vayanos, D. and Wang, T. (2007), 'Search and endogenous concentration of liquidity in asset markets', *Journal of Economic Theory* **136**(1), 66–104.

Walras, L. (1874), *Elements of Pure Economics*, Routledge.

Wang, J. and Yang, M. (2011), 'Housewives of Tokyo versus the gnomes of Zurich: Measuring price discovery in sequential markets', *Journal of Financial Markets* **14**(1), 82–108.

Wang, X. and Ye, M. (2017), 'Who provides liquidity and when: An analysis of price vs. speed competition on liquidity and welfare', Working paper.

Wermers, R. and Xue, J. (2015), 'Intraday ETF trading and the volatility of the underlying', Working paper.

Wood, R. A., McInish, T. H. and Ord, J. K. (1985), 'An investigation of transactions data for NYSE stocks', *Journal of Finance* **40**(3), 723–739.

Xu, L., Yin, X. and Zhao, J. (2019), 'Differently motivated ETF trading activities and the volatility of the underlying index', Working paper.

Yan, B. and Zivot, E. (2010), 'A structural analysis of price discovery measures', *Journal of Financial Markets* **13**(1), 1–19.

Yang, L. and Zhu, H. (2019), 'Back-running: Seeking and hiding fundamental information in order flows', *Review of Financial Studies, Forthcoming* .

Yao, C. and Ye, M. (2018), 'Why trading speed matters: A tale of queue rationing under price controls', *Review of Financial Studies* **31**(6), 2157–2183.

Ye, G. (2011), *High-frequency trading models*, Wiley Online Library.

Yin, X. (2005), 'A comparison of centralized and fragmented markets with costly search', *Journal of Finance* **60**(3), 1567–1590.

Yueshen, B. Z. (2017), 'Uncertain market making', Working paper.

Zhu, H. (2014), 'Do dark pools harm price discovery?', *Review of Financial Studies* **27**(3), 747–789.