

© 2022 Pedro Tremacoldi-Rossi

ESSAYS IN FINANCIAL MARKETS AND INSTITUTIONS

BY

PEDRO TREMACOLDI-ROSSI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Applied and Agricultural Economics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Scott H. Irwin, Chair
Professor Dan Bernhardt
Professor Michel Robe
Assistant Professor Joe Janzen

Abstract

This dissertation consists of three chapters on financial markets and institutions. The first essay studies several dimensions of how trading automation in stock markets impacted trading firms and workers. Electronic trading migration in the early 2000s was the most important technological change in the history of stock markets. However, our understanding of its impacts on industry profits and their distribution across firms, as well as on workers that witnessed a dramatic transformation in the task content of their jobs from physical and intuitive thinking to highly skilled programming and IT knowledge, remains limited by data opacity and lack of appropriate variation. To answer these questions, I combine a novel dataset of the universe of broker-dealer firms and licensed professional traders in the US with variation to exposure to trading automation induced by the SEC's major redesign of stock trading.

I find that areas with greater exposure to trading automation, measured as higher pre-existing IT infrastructure, experience larger increases in total trading profits. In the same high-exposure areas, larger firms increase their share of industry profits, a distributional effect driven by an intensive margin response of surviving firms increasing profit levels. Consistent with higher barriers to entry and setup costs imposed by higher technology requirements, the entry of new broker-dealers decreases more in areas with higher automation, and both the largest and the smallest firms experience higher survival rates. While these larger firms represent multi-billion-dollar investment banks and other brokers that cater to institutional and large investors, very small firms are mainly specialized boutique brokers with large relative IT capital investments and more likely to generate revenue from proprietary trading activities.

I also find that stock traders experience large decreases in employment, both when comparing these workers to traders plausibly unaffected by the stock market redesign policy, including commodities traders, and when comparing stock traders in areas with differential degrees of exposure to trading automation. Traders that manage to remain employed in broker-dealers are more likely to switch markets, becoming investment advisors, bond traders, or working in other segments with lower degrees of automation relative to stocks. They are also more likely to switch between firms, usually with a firm size penalty. Consistent with trading automation being a skill-biased technical change, more skilled traders experience wage gains

over less skilled individuals. Additionally, the employment of non-traditional professionals in the industry, including computer scientists and software engineers grows, albeit being insufficient to offset losses in trader employment.

The second essay (joint with Scott H. Irwin) focuses on a crucial aspect of financial markets: liquidity measurement. We investigate how and why tools developed to estimate bid-ask spread with daily data often misbehave in several contexts and provide researchers with simple tests to empirically assess the magnitude of estimation bias from using these proxies. We first show that poorly designed horse-race studies measuring the performance of simple bid-ask spread estimators in modern stock markets lead to inflated performance results and mask the existence of a systematic relationship between higher liquidity and estimation bias. We then derive a framework that explains the determinants of the bias using a popular and recently developed spread estimator, the high-low measure, showing how these common misbehavior factors are equivalent to well-known bias sources in older spread estimators. We suggest that these biases cannot be completely corrected for in any estimator sharing the same properties and implemented with sample quantities instead of population moments, and unfortunately are likely to perform poorly in modern markets, where liquidity cost levels tend to be sufficiently low. Finally, we offer a number of tools to successfully bound the estimation bias in any empirical context.

The third essay (joint with Scott H. Irwin and Conner Naughton) studies the prevalence and effects of market manipulation by algorithmic traders. We focus on self-trading, a predatory practice that entices traders to submit orders on a certain side of the market by creating fictitious trades where both counterparties are the same trader. We show how microstructure features of electronic markets enable a much higher level of sophistication in the use of self-trading and study how different quoting strategies impact market quality and order flow.

To Mom, Dad, & Tatiana.

Nisi credideritis, non intelligetis — Saint Augustine, *De libero arbitrio*

Acknowledgments

First and foremost, I would like to thank my advisors Scott H. Irwin, Dan Bernhardt, and Michel Robe.

Scott has been a powerful intellectual presence and wonderful advisor since I began my graduate studies. From him, I learned countless lessons: Think big picture. Always have a punchline — most preferably subsumed in one picture. Deliver clear and concise writing, but not too concise. Work on what you love. He introduced me to commodities markets and many of the fundamental changes trading has undergone in the last years, which has greatly inspired my work. I am deeply grateful for his careful revision of countless of my drafts and providing detailed feedback. He continued supporting me even as my research expanded well beyond the scope of commodities markets.

Among many things, Dan taught me how to transition from graduate student to researcher. He challenged me to go beyond the obvious answer to a research question and methodically investigate alternative avenues and arrive at deeper insights. His approach to research, dedication and kindness toward students have inspired me immensely.

Michel provided guidance and support throughout my MS and PhD. I am extremely grateful for his early feedback on this dissertation's first chapter which helped not only in its framing but to define my research agenda more broadly.

I am grateful to Joe Janzen for his generosity with his time and for insightful feedback and comments. Beyond my committee, I also want to thank Phil Garcia for numerous conversations and guidance. His lectures with Scott were a highlight of my years at UIUC and a great privilege for ACE students. I also would like to thank Teresa Serra for her support over the last years. I also want to thank ACE's faculty and staff, particularly Dung Wisdom for her immense help and dedication to the department.

Many say that graduate school is a lonely journey. Luckily, I will never be able to say whether this is the case. I am lucky to have found my life partner, Tatiana Mocanu. She has been my inspiration and joy. She made me grow years in months, pushed my research in directions I never knew were possible, and supported me unconditionally as a friend, coauthor, reviewer, discussant, interviewer, and partner. This dissertation would have not been possible without her love and brilliance.

Table of Contents

List of Tables	viii
--------------------------	------

List of Figures	ix
---------------------------	----

Chapter 1 Financial Regulation and Automation Adoption: Evidence from Stock Trading Firms 1

1.1 Introduction	1
1.2 Relation to Previous Literature	6
1.3 Background	7
1.3.1 Financial Regulation: Institutional Framework	7
1.3.2 The Slow Rise of Electronic Trading	9
1.3.3 The Fast Dominance of Algorithmic Trading and Its Consequences	10
1.4 Broker-Dealer Firms and Workers	11
1.4.1 US Broker-Dealers	11
1.4.2 Financial Filings	12
1.4.3 Information Technology	13
1.4.4 Other Data	14
1.5 Anatomy of the US Trading Industry	14
1.6 Profits, Concentration, and Firm Performance	16
1.6.1 Empirical Strategy	16
1.6.2 Validating IT Exposure: Evidence from Trading Patents	17
1.6.3 Industry Profitability and Firm Performance	18
1.6.4 Survival and Entry	19
1.7 Impact on Traders	20
1.7.1 Employment	20
1.7.2 Other Broker-Dealer Workers	22
1.7.3 Skill Premium	23
1.7.4 Robustness Checks	24
1.8 Alternative Explanations	26
1.8.1 The Great Recession	26
1.8.2 Online Brokerage Services	29
1.9 Conclusion	29
1.10 Figures and Tables	31

Chapter 2 The Bias of Simple Bid-Ask Spread Estimators 52

2.1 Introduction	52
2.2 The Structure of Simple Spread Estimators	57
2.2.1 Setting	57
2.2.2 The Roll Estimator	58
2.2.3 The High-Low Estimator	59
2.3 Reevaluating Simple Bid-Ask Spread Estimators	61
2.3.1 Benchmark Stock Data	61

2.3.2	Reassessing the Performance of Simple Spread Estimators	61
2.4	Estimation Bias & Performance	64
2.4.1	Framework	65
2.4.2	Bias and Its Sources	66
2.4.3	Common Biases Account for 90% of Performance Loss	68
2.4.4	Determinants of Bias	70
2.5	Measuring Bias Empirically	72
2.5.1	Deriving Empirical Bias Bounds	73
2.5.2	Empirical Implementation	74
2.6	Conclusion	79
2.7	Figures and Table	80
Chapter 3	New Wine into Old Wineskins: Manipulation in Financial Markets	90
3.1	Introduction	90
3.2	Related Literature and Contribution	92
3.3	Data and Setting	92
3.3.1	Regulatory Background	93
3.3.2	Self-Trade Prevention Functionalities	93
3.3.3	Recent Enforcement Cases	94
3.3.4	Trading Message Data	95
3.4	Anatomy of Self-Trading in Futures Markets	96
3.4.1	Algorithm for Detecting Self-Trading	96
3.4.2	Aggregate Patterns	97
3.5	Quantifying the Impact of Self-Trading	98
3.5.1	The Economics of STP Manipulation	98
3.5.2	Market Impact	99
3.6	Figures and Tables	103
References		111
Appendix A	Appendix for Chapter 1	156
A.1	Broker-Dealer Financial Information	156
A.2	Appendix Figures and Table	158
Appendix B	Appendix for Chapter 2	169
B.1	Sample Handling & Construction	169
B.2	Proofs & Other Derivations	169
B.3	Additional Performance Evaluation in US Stocks	172
B.4	Negative Spread Estimates: Determinants & Empirical Use	174
B.5	Supplementary Analysis: Close-High-Low Estimator	180
B.6	Appendix Figures	185

List of Tables

1.1	Where do Stock Traders Work?	43
1.2	Largest Employers in the Investment Industry	44
1.3	Availability of New Technologies by Sector	45
1.4	Did Automation Increase Industry Trading Profits?	46
1.5	Did Automation Increase Firm Trading Profits?	47
1.6	Impact of Automation on Broker-Dealer Entry & Survival	48
1.7	Employment & Wage of Securities Brokers	49
1.8	Placebo and Robustness Tests	50
1.9	Great Recession & Online Brokerage Services	51
2.1	Ranking Stocks Based on Liquidity	89
3.1	Self-Trading in Futures Markets, Summary Statistics	106
3.2	Liquidity Taken By STP-Triggered Cancellations	107
3.3	Order Flow Impact	108
3.4	Order Flow Intensity	109
3.5	Trading Pressure Following Self-Trades	110
A.1	Standardizing Broker-Dealer Balance Sheet, Assets	168

List of Figures

1.1	Trends in Jobs and Technology Capital in Trading	31
1.2	Growth in US Stock Trading Activity and Employment of NYSE Floor Brokers	32
1.3	Occupation Entry and Investment Industry Exit	33
1.4	Total Assets of Top Broker-Dealer Firms	34
1.5	Range of Broker-Dealer Performance	35
1.6	Broker-Dealer Profitability and Stock Market Return	36
1.7	Trading Automation, Overall Finance, & Non-Finance Patents	37
1.8	Impact of Trading Automation on Stockbroker Employment	38
1.9	Impact on Stockbroker Employment vs. Other Investment Workers	39
1.10	Impact on Employment of Other Broker-Dealer Workers	40
1.11	Placebo Employment Regressions	41
1.12	Stock Trader Employment Rates During Key Events of the Financial Crisis	42
2.1	Distribution of Effective Bid-Ask Spreads in US Stocks	80
2.2	Cross-Sectional Correlation: HL Estimates and Effective Spread By Effective Spread Decile Size	81
2.3	Relationship Between $\sqrt{2}R$ and R^* Under Ideal Conditions	82
2.4	Cross-Sectional Correlation: Best-Performing HL & Effective Spread by Effective Spread Decile Size	83
2.5	Two Examples of Performance: Best-Performing HL Estimator	84
2.6	Bias and True Spread Under Ideal Conditions	85
2.7	Testing the Accuracy of the Empirical Approach to Sign the Small Sample Bias	86
2.8	Bounding the High-Low Empirical Bias in US Stocks	87
2.9	Stock-level Estimation Bias Bounds	88
3.1	Stylized Routing System and Measurement Points of Timestamps	103
3.2	Identifying Self-Trades in Message Data	104
3.3	Intraday Distributon of Self-Trades	105
A.1	Goldman Sachs Listed Names at BrokerCheck.com	158
A.2	Example of Goldman Sachs & Co. LLC 13-NT	159
A.3	Broker-Dealer Balance Sheet Examples, Assets	160
A.4	Two Sides of Trading Automation: Exchanges and Broker-Dealers	161
A.5	Example of Stockbroker Registration Record from BrokerCheck.com	162
A.6	IT Intensity Across Industries in 2004	163
A.7	Cross-Sectional Variation: IT Intensity and Broadband Providers	164
A.8	Face of Form X-17A-5 Submitted in Paper Format	165
A.9	Number of Form X-17A-5 Filed Every Year by Section Available	166
A.10	Evolution of Broker-Dealer Employment: US Counties	167
B.1	Cross-Sectional Correlation: Stability Across Different Sample Sizes	185
B.2	High-Low Negative Estimates Under Ideal Conditions	186

B.3	HL Performance in the Time Series and Dispersion of Effective Spread	187
B.4	Negativity Induced by the Spread Size in the HL Estimator	188
B.5	Relationship Between Spread Size, % of Negative Estimates Under Ideal Conditions	189
B.6	Relationship Between Spread Size, % of Negative Estimates Under Ideal Conditions	190
B.7	Bias, Negative Estimates, and True Spread Under Ideal Conditions	191
B.8	Cross-Sectional Correlation: CHL Estimates and Effective Spread By Effective Spread Decile Size	192
B.9	Cross-Sectional Correlation: Roll Estimates and Effective Spread By Effective Spread Decile Size	193
B.10	Cross-Sectional Correlation: Changes in HL Estimates and Changes in Effective Spreads By Effective Spread Decile Size	194
B.11	Cross-Sectional Correlation: Changes in CHL Estimates and Effective Spreads By Effective Spread Decile Size	195
B.12	Cross-Sectional Correlation: Changes in Roll Estimates and Changes in Effective Spreads By Effective Spread Decile Size	196

Chapter 1

Financial Regulation and Automation Adoption: Evidence from Stock Trading Firms

1.1 Introduction

Electronic trading changed the traditional organization of financial markets. From the 18th century to the late-1990s, equities and derivatives trading took place in exchange-based physical marketplaces, where broker-dealers met to buy and sell financial products for their own accounts or on behalf of customers. As electronic platforms started to proliferate in the early 2000s, machines gradually eliminated much of the need for direct intermediation and floor trading by replacing physical markets with computer networks. Market participants' orders are now routed to electronic systems that fully constitute the trading venue. Matching algorithms organize trade in continuous auctions markets with little human intervention.

The ability to trade electronically at increasing speeds and optimize order handling was soon associated with the rise of new types of market participants. Exchanges created new technologies to extract rents from selling speed advantages and increasingly introduced benefits to reward timely liquidity provision (Budish et al. (2019)). Firms created trading desks specialized in low latency arbitrage and sophisticated algorithms implementing hundreds of parallel strategies, which quickly dominated stock traded volume. The new landscape of stock markets required massive investments in physical infrastructure and human capital.¹ Fairness and efficiency concerns ensued. This technological arms race was seen by many as wasteful and detrimental to the majority of other investors. Michael Lewis' *Flash Boys* famously claimed that automated trading and speed "rigged" the market to benefit "insiders" with deep pockets.²

Did technology make trading more unequal? Are profits now more concentrated? Which firms and what type of traders have benefited from algorithmic markets? To answer these questions, I assemble a unique dataset with the universe of US broker-dealer firms and stock traders. With detailed information on traders' employment history, market segment, licenses, firms' performance and IT investment, I investigate the consequences of trading automation for the investment industry. My estimates put a dollar figure on

¹When Spread Networks connected Chicago to New York using a straighter fiber cable in 2010, trading firms were eager to pay \$16 million upfront to shave 1.5 milliseconds off the time their orders took to reach the exchange's electronic system. Few years later, the technology became too slow (MacKenzie (2021), Shkilko and Sokolov (2020)). Routing times today are as fast as 1 microsecond (0.001 millisecond). Spread Networks intended to raise almost \$3 billion in access fees to the cable.

²Aquilina et al. (ming) estimate a cost levied on other traders of \$5 billion per year from competition between fast traders.

the benefits to automation: on average \$98,000 in annual firm net profits for each additional computer per worker. In the aggregate, even relatively modest investments in technology lead to an increase in industry local profits of \$1 million.

I begin by constructing a novel dataset combining two sources of broker-dealer data. The first exploits the high degree of regulation on the activity of traders and broker-dealer firms in equities markets. These requirements establish a complex and continuously updated system of licenses and registrations maintained by the Financial Industry Regulation Authority (FINRA) called Central Registration Depository (CRD). To trade professionally or provide financial services, individuals need to take and pass examinations. Because examination applications require an employer’s sponsorship, a new license captures not only a worker’s new skills but plausible changes in market segment or narrow occupational tasks. I extract all information contained in FINRA’s CRD database to construct a longitudinal linked employer-employee panel comprising over 500,000 financial workers — both currently active or that left the industry in the last 30 years — with rich employment history and location, experience, qualification, and market segment licenses. The sample provides a first-time comprehensive portrait of the US stock trading industry at a disaggregated level and without survivorship or selection biases.

The second source of information I leverage uses firm-level financial filings of broker-dealers. The SEC mandates every broker-dealer to submit form X-17A-5 annually. The filing contains detailed balance sheet information, financial income, and leverage. Contrary to commonly used financial filings (e.g., 10-K forms), X-17A-5s are not centrally held in any repository and public versions of the forms are available as scanned paper documents. This requires both search and extraction from each firm’s records individually. By linking broker-dealer names from the FINRA database to information from several regulatory systems, I extract over 80,000 unique broker-dealer financial filings spanning 2001 to 2020. I then parse and standardized the contents of each filing into a representative broker-dealer balance sheet and income statement.³

To generate variation in the adoption of automated trading by broker-dealers, I use SEC’s chief provision to redesign equities trading, Regulation National Market System (Reg NMS). Electronic trading systems for stocks had been available since the 1970s, but adoption by broker-dealers and exchanges was slow. Floor trading remained the leading form of trading even after Instinet’s green screen terminals already provided off-exchange electronic trading since the 1980s, and equities markets were long connected electronically by the Intermarket Trading System.

Reg NMS changed this paradigm for stocks by installing price-protection provisions that only apply to

³Recent work by Egan et al. (2019) and Maggio et al. (2019) also compiles some of the data from FINRA’s BrokerCheck, the user interface providing public access to CRD records. Part of the information available in forms X-17A-5 microfound national aggregate broker-dealer information used in Adrian et al. (2014), which is available from the Federal Reserve Flow of Funds.

stocks in electronic markets and with automated execution, which imposed a minimum speed requirement on participant exchanges. Initially meant to improve stock trading competition between exchanges, price integration steps in the regulation, like Rule 611 in the National Best Bid and Offer (NBBO), led to market fragmentation and raised the returns to speed by forcing orders to be executed in the market offering the best outstanding quotes (O’Hara and Ye (2011), O’Hara (2015)).

Because the policy targeted stock markets and not individual firms, mapping changes in market microstructure induced by automation to treated and untreated agents is problematic without knowing each firm’s portfolio. The second step I take for identification is to exploit spatial variation in a measure that is likely correlated with local automated trading take-up by broker-dealers. Areas with large pre-existing technology infrastructure in the financial investment industry, measured by computers per worker, face lower implementation costs of more sophisticated technology and imply the availability of basic services to sustain greater automation, such as high-speed internet. This IT intensity allows me to assign heterogeneous degrees of local exposure to electronic trading by comparing worker and firm outcomes in areas with high levels of IT infrastructure to the ones in low-IT areas before and after Reg NMS.

The relevance of the local exposure in broker-dealer firms relates to two channels. First, because of decreased adoption uncertainty induced by Regulation National Market System and if speed is profitable, firms have greater incentive to upgrade technology when barriers to development are lower. Second, if speed is a cost-saving technology and enables firms to cut commission fees, local competition for investors will influence automation take-up. While this second channel largely applies to broker-dealer firms competing for a local pool of customers and the degree of localization in the demand for brokerage services, the first channel is relevant to all broker-dealers trading in stock markets that reward speed.⁴

To validate the intensity measure, I show that areas with higher pre-existing IT exposure have much larger growth rates in patenting related to trading automation, while patents in the finance industry overall and in non-financial sectors behave similarly regardless of this infrastructure. Further, areas with high levels of IT were not accelerating technology investments prior to Reg NMS, suggesting that firms there did not anticipate the increased returns to electronic trading before the policy started to shift exchanges to automated execution.

Centering my first set of analyses on this empirical design, I show that aggregate industry profits in areas with higher automation propensity increased on average by \$1 million for each additional computer per worker as a consequence of Reg NMS. This profit growth was fueled by broker-dealers with large book sizes, whose profit participation increases 4 percentage points for the same investment in technology. In fact,

⁴Evidence by Maggio et al. (2019) shows that the demand for brokerage services is extremely local, partially because investors prefer to be close to their brokers. For evidence on the rewards to speed in financial markets, see Baron et al. (2012).

interacting the IT measure with firm total assets implies that the same technology investment in JP Morgan and Citadel is \$330,000 and \$55,000, respectively, more profitable than in a \$100 million asset broker-dealer. Higher returns to technology in larger firms made trading profits more concentrated.

Next, I find that on average firm profits increase by \$98,000. To investigate which firms benefit the most from trading automation, I show that existing broker-dealers increase net income on average by \$58,000. This implies that new trading firms experience greater benefits from automation. Supporting the conjecture of increase in local competition, firm entry falls by 24 percentage points following the implementation of Reg NMS. The impact on incumbents also reveals important patterns. Small trading firms with less than 50 stock traders and the largest brokers become more likely to survive — with large investment banks being 4 times more likely to stay in business.

These findings underscore the unequal gains from trading automation. While aggregate profits in the industry increased, larger firms became both more profitable and more likely to survive relative to mid-sized brokers. Firms like CJ Lawrence, Susquehanna International, Goldman Sachs, Credit, and Virtu derive most of their trading revenues from proprietary activities and by catering to high net worth investors. My estimates indicating the firm with few traders also become more likely to stay in business aligns with the organizational structure of many of these brokers — small, specialized venues with high investments in IT infrastructure per worker. It is also important to emphasize that some large book-sized brokers, such as Schwab, have invested heavily in increased market access to small and retail investors through online trading resources.

In the next part of the paper, I turn to changes within broker-dealers. Algorithms not only replace repetitive tasks with automated processes, but by large substitute human decision making, operating largely unsupervised for extended periods of time. Trading algorithms make decisions based on pre-established rules that operate by autonomously reading market signals inferred from order flow data. These tools require intensive back-office resources for maintenance and fine-tuning, and take several thousands of person-hours to be developed and tested until becoming operational.⁵

While these new technologies expanded the set of tasks and occupations in broker-dealers and increased the demand for mathematical and programming knowledge, potential labor reinstatement (Acemoglu and Restrepo (2019)) of stock traders did not outweigh displacement. I find that higher exposure to automation decreases the employment probability of stock traders working at the time of the full implementation of Reg NMS by 2 percentage points. I then exploit an additional design to identify employment effects, where I compare stock traders to financial workers employed at broker-dealers working in other market segments. Specifically, I consider bond traders and investment advisors, whose markets had much lower

⁵In *Flash Boys*, Michael Lewis reports that in 2008, Goldman Sachs had over 60 million lines of proprietary trading code.

degrees of automation than stocks. Stock traders become 10-20 percentage points less likely than these other professionals to be employed. When considering all other workers in the investment industry, stockbrokers experience employment losses of similar magnitude.

I supplement the analysis by investigating the role of trading automation in complementing the skills of certain traders who remain employed. Indeed, I find support for electronic trading being a skill-biased technology by comparing wages of individuals in the investment industry in areas with higher exposure to automation. Contrary to the previous analyses, this step pools stock traders with other workers in the trading sector, and includes composition changes by tracking repeated cross-sections of individuals. My estimates show a 9 percentage-point increase in wages of college-educated traders relative to unskilled individuals, translating into \$10,000 annually at the mean for each additional unit of technology investment.

My different empirical designs rest on the identifying assumption that without Regulation NMS, the trend in the outcomes I study would have remained unchanged between treated and control units. Given that Reg NMS became fully operational in October 2007, the potential effects of the Great Recession correlated with treatment intensity are a direct threat to identification. I carefully consider multiple avenues where the financial crisis could impact my results and conduct a series of tests to analyze the importance of alternative explanations.

Particularly, I conduct two main lines of investigation. The first considers that the Great Recession directly drives my estimates through mechanical employment losses and deterioration of broker-dealer balance sheets. Excluding the big 5 investment banks — including Lehman Brothers — has no effects on my results. Alternatively, excluding New York or other large financial centers from my sample only alters marginally the magnitude of the estimates.

Since my specifications allow for worker reallocation between firms after separation, firm idiosyncratic or localized shocks only affect the estimated results if effects of the Great Recession correlate with 2004 IT levels. To test this hypothesis, I use a placebo group of workers unaffected by Reg NMS but that I assume are just as exposed to the Great Recession as stock traders. These comprise individuals employed in all finance areas, including real estate agents, insurance agents, and bankers. The employment effect of this group in higher IT areas is positive, indicating that the the Great Recession is unlikely to drive my results via shocks to the financial sector.

Next, I test for the effects of the financial crisis on the demand for brokerage services. Losses in employment in other sectors could depress household stock holdings and decrease assets under management for brokers. All my results remain unchanged when adding measures of local demand for brokerage services to the baseline model, including local income per capita and stock market wealth following Chodorow-Reich et al. (2021). Further controlling for local home price changes, which absorbs shocks to household balance

sheets, also has no impact on my estimates.

Although I cannot rule out the importance of additional channels through which the financial crisis bias the difference-in-differences estimates, taken together these tests suggest that areas with higher exposure to trading automation were not particularly affected by the Great Recession. Finally, I also test the sensitivity of my estimates to the rise in online discount brokerage services. Platforms such as E*trade allowed physical investors the ability to trade using their personal computers and phones. Controlling for local use of online brokerage services also exerts minimal influence on the baseline results.

The rest of the paper proceeds as follows. Section 1.2 discusses the paper’s contribution to the literature. Section 1.3 introduces the regulatory and technology background in the US trading industry, focusing on the relevant aspects of the policy to my context and its consequences to stock trading. Section 1.4 describes raw information sources and construction steps of data sets of trading firms and workers. Section 1.5 shows some patterns of these new data. Section 1.6 studies which broker-dealer firms benefited from trading automation. Section 1.7 studies the impacts of automation on stock traders and other groups of workers in trading firms, including quants, clerical occupations, and traders in different market segments. Section 1.8 assesses the role of the Great Recession and online brokerage services in explaining my results through a battery of tests. Section 1.9 concludes.

1.2 Relation to Previous Literature

This paper contributes to several strands in the finance and economics literature.

Broadly, my paper complements existing theoretical work showing how new technologies in financial markets shape the investment industry and market participants (Budish et al. (2015), Farboodi et al. (2020), Mihet (2020)). By causally linking trading automation to winners and losers in new financial markets, I show that despite increasing profits in the investment industry, technology investments had several distributional consequences. Trading profits became more concentrated in broker-dealers with larger asset sizes, where technology improvements translate into higher revenues related to smaller firms. These brokers, along with firms employing few traders that specialize in proprietary activities and wealth management experienced increased survival rates, with fewer entrants due to automation.

I also go beyond market microstructure effects (Stoll (2006), Menkveld (2016)) and show how the largest technological change financial markets have ever undergone impacted the investment industry. Several finance papers have looked at how migration to electronic trading affected liquidity provision, quality, and efficiency. Work by Jain (2005), Tse and Zabolina (2001), Raman et al. (2017), and many others shows that electronic markets increased pricing efficiency and liquidity. Other studies revealed how electronic markets

changed market participants (e.g., Raman et al. (2017) for oil markets), which suggests trader composition shifts with unknown consequences to firm and worker outcomes.

My empirical findings further connect a more theoretical-based strand of the literature on the structure of skill and pay in the financial intermediation industry (e.g., Bond and Glode (2014), Reshef and Philippon (2012) and Glode and Lowery (2016)). By providing empirical evidence on employment losses and higher skill wage premia for stockbrokers caused by electronic trading, my analysis underpins skill-biased technologies as a main driver of intensive-margin adjustments in broker-dealers. Trading automation not only substitutes away unskilled work and eliminates repetitive tasks, but also requires high skill work to be operated, impacting the skill composition of traders.⁶

By using the implementation of Reg NMS to flesh out the timing of trading automation in stock markets, this paper naturally stresses the role of financial regulation in directly shaping firm behavior. In particular, it relates to work by O'Hara and Ye (2011) and Haslag and Ringgenberg (2016) who study the effects of Reg NMS and market fragmentation on trading costs and market quality. They find that execution quality and liquidity costs have improved, with the differential in Haslag and Ringgenberg (2016) that only large stocks reaped the benefits of market fragmentation. I contribute to this body of literature by showing how the policy negatively affected trader employment by influencing firms to shift activities towards capital stock.

1.3 Background

I begin by introducing key regulatory changes in the equities trading industry since 2000, in particular Regulation National Market System. I then briefly discuss several dimensions illustrating the dominance of electronic trading in modern financial markets. This discussion sets the stage to highlight the sharp changes to the structure of stock trading beginning in 2005.

1.3.1 Financial Regulation: Institutional Framework

Since the creation of the National Market System (NMS) by the Securities Acts Amendments of 1975, US securities markets have been partially integrated in an attempt by policy makers to provide greater transparency and foster competition between trading venues. The NMS created a central marketplace where quotes once scattered across different exchanges could be publicly displayed. In the 2000s, the SEC introduced new regulation as a response to the development of new market mechanisms and electronic stock

⁶This is also related to a large strand in the literature investigating the impacts of technological change on labor markets (e.g., Autor et al. (2003), Acemoglu and Restrepo (2019), Acemoglu and Restrepo (2020), Acemoglu and Autor (2011), Aghion et al. (2020)).

exchanges which rendered the 1975 national market system outdated.

Regulation ATS. Implemented in 2000, SEC Regulation ATS extended the reach of existing rules to include automated trading systems (ATS) under the national market system. Automated trading systems are one type of algorithmic trading and use sophisticated computer programs to generate and execute buy and sell orders. Reg ATS offered ATSs the option to register as an exchange or a broker-dealer, which initiated the integration of these automated off-exchange systems with traditional financial centers.

Regulation NMS. Following Reg ATS, Reg NMS was passed in 2005 and became fully effective in October 2007, after major exchanges requested additional time to adapt to the policy provisions. The set of rules intended to promote “*efficient and innovative trading services*” and “*more efficient pricing*” of stocks by fostering “*vigorous competition among markets*”, in addition to “*integrated competition among orders*”.⁷ The dramatic redesign of the national market system centered in a new requirement for brokers to execute customer orders at the best available quote (top-of-book). The most important and contentious provision of Reg NMS, the Order Protection Rule (Rule 611), known as “trade through” rule, mandates intermediaries to direct orders to the venue posting the best available quote, which is protected from trading through (executing the order at a suboptimal price).

Only automated quotes — offering immediate display in the NMS, now consolidated under the NBBO (National Best Bid/Offer) — were subject to the Reg NMS price protection. Receiving this benefit was seen by exchanges as essential to attract order flow (Spatt (2018)), which along with other provisions in the regulation profoundly changed stock trading and the investment industry in the US.

Reg NMS is widely acknowledged as one of the main causes of market fragmentation and the rise of speed-driven competition in US financial markets. First, because Reg NMS only protects electronic quotes, stock exchanges had an incentive to adopt electronic limit order books (Chao et al. (2018)), boosting execution speed as automated order entry and execution is inherently faster than manual quoting. Hendershott and Moulton (2011) argue that NYSE introduced its Hybrid Market (which expanded trading automation and consequently reduced latency) to meet the criteria of Rule NMS. Pagnotta and Philippon (2018) show that order execution speed in the exchange decreased over 10-fold within weeks following automation.

Second, since Rule 611 routes trading volume to venues offering lower quotes, it mechanically increases off-exchange volume and thereby fragmentation. Off-exchange ATSs, such as electronic communication networks (ECNs) and dark pools, now account for 44 of 57 US trading venues and much larger market shares. By 2013, dark pools had a two-fold increase in market share from 7% in 2008 (Kwan et al. (2015)). Traditional exchanges saw sharp drops in market share: equity market volume held by NYSE and NASDAQ decreased from 70% and 50% in 2006, respectively, to 20% and 30% in 2011 (Pagnotta and Philippon

⁷Text of SEC’s Regulation NMS. More details in <https://www.sec.gov/rules/final/34-51808.pdf>.

(2018)).

Under market fragmentation, quotes are scattered across a larger number of markets and search costs increase. Biais et al. (2015) sets up a model in which fragmentation cost reduction induces investment firms to overinvest in speed-driven technology. Competition between exchanges also lead to significant increases in trading speed. In order to attract volume and reduce execution latency, exchanges cut transaction fees (Colliard and Foucault (2012)) and invested in new infrastructure to make trading faster (Pagnotta and Philippon (2018)). But these new capabilities could only be explored by trading firms undertaking significant investments in IT hardware and automation software.

Overall, Reg NMS fostered speed-driven competition in two ways. By imposing a minimum speed requirement (electronic trading and automated order entry), the legislation rendered the existence of purely floor-based exchanges economically unattractive. Second, price integration policies, like Rule 611, result in market fragmentation (Pagnotta and Philippon (2018)), which make latency a central market structure issue. US exchanges are for-profit companies that generate revenue out of fees accruing from trading volume. If trade volume is directed to cheaper venues, exchanges lose market share and trading revenue. Taken together, these two facts increased the returns to trading speed and lowered adoption uncertainty, likely contributing to the rise of algorithmic trading, in particular of high-frequency traders (HFTs).⁸

1.3.2 The Slow Rise of Electronic Trading

Electronic financial markets are not particularly a recent innovation. Launched in 1969, Instinet was the first and only electronic communications network in Wall Street, competing with traditional floor-based exchanges. Only after the diffusion of PCs in the 1980s, non-physical trading started to fully develop, in part because of regulation following the crash of 1987 pushing exchanges to adopt automated execution systems.

Transitioning to electronic markets, however, was a slow process. Traditional exchanges operated under secular systems of memberships, where trading advantages come from member benefits — liquidity externalities attracted volume which begat volume. Floor-brokers pushed back automation, fearing both skill obsolescence and loss of trading fee revenues, as most exchanges were then non-profit organizations owned by physical traders. As late as 2003, NYSE’s chairman Richard Grasso vigorously defended open outcry trading. Following Grasso’s ousting, Chief Executive John Thain pushed the exchange to further develop electronic-based trading systems. NYSE merged with Chicago’s Archipelago Exchange in 2005, then, a leading ECN, following Nasdaq’s purchase of Instinet. NYSE’s merge (which also resulted in the company going public

⁸With time, exchanges developed new ways to exploit electronic trading and the presence of algorithmic traders to increase profit beyond volume-based trading fees, including offering co-location slots to reduce speed latency, and several complex fee schema based on quoting activity.

and becoming for-profit) was part of the hybrid market initiative. The move still faced opposition of several floor brokers — self-named “Independent Broker Action Committee” — who claimed that automation would *“leave floor brokers out of major trades, limiting their access to liquidity and threatening their livelihoods”*. Thain responded that *“while a small number of people, although vocal, are opposing”* automation, *“the vast majority of brokers are very supportive”*. He also referred to floor specialists as a “bottleneck”, eroding the exchange’s market share and slowing down trading.

The late adoption of electronic platforms by major exchanges (NYSE’s initial steps in 2001 and 2004, NYMEX in 2006, and ICE in 2008) shows the uncertainty and challenges surrounding the adoption of electronic markets. A notorious example of skepticism with respect to electronic trading adoption in the late 1990s not exclusive to stock markets is the case of the former futures exchange CBOT. In 1997, the legendary exchange took a gamble against its rising competitor — CME’s electronic platform Globex, introduced in 1992 — by building a \$300 million, 65,000 ft. trading floor in Chicago. Called the “Taj Mahal”, the new physical floor represented the belief of many in the investment industry that physical trading offered advantages to automation and would out-compete electronic marketplaces. In December 1998, chairman Patrick Arbor was voted out of office by CBOT members because of his attempt to offer the exchange’s derivatives contracts at Eurex’s electronic platform. Ultimately, CBOT was purchased by CME in 2007. Open outcry trading in the exchange’s floor shut down in 2015, following the trend from competing venues, like ICE, who went all-electronic in 2012.

1.3.3 The Fast Dominance of Algorithmic Trading and Its Consequences

The previous discussion provided background on the importance of Reg NMS in fostering speed-based competition and lowering development costs of algorithms and of the structure necessary to fast trading. Because the policy raised the returns to automated execution, it is likely to have influenced the timing and intensity of electronic trading adoption by broker-dealers and exchanges.

After Reg NMS came into effect, trading automation sharply dominated US stock markets. I illustrate the speed and impact of this movement in two ways. First, Figure 3.2 shows real-side patterns of the technology arms race in stock trading firms in the last decade. It plots investment in IT capital stock by broker-dealers over time — these include computers, mainframes, storage capacity, data technologies, as well as expenditures on software development and adoption. The data show a three-fold increase in IT investments from 2005 to 2010, when it reached \$30 billion a year. The trend was accompanied by a fast decline in stock trading employment, which added 40,000 jobs in 2000-2007 and eliminated 60,000 since then. Employment levels in 2015 were similar to the 1990s, when the financial sector represented a much smaller

share of the US economy (Philippon (2015)).⁹

This accelerated investment in IT capital had profound effects on stock markets. Figure 1.2 compares the daily number of traded stocks in the US by market participant segment. High-frequency trading alone — the most sophisticated form of automation — accounted for over 60% of US daily equity volume in 2009 compared to 20% in 2005, according to data from the Tabb Group and Credit Suisse displayed in the figure’s first panel. The pace of volume growth is even more striking when contrasted with the evolution in active and passive money managers and investors between 1999 and 2009. Without HFT presence, total traded volume in US stocks grew from just below 1 billion shares a day to 4 billion over the entire decade. In six years, volume traded by high-frequency traders increased from virtually zero to over 6 billion daily shares. The second part of Figure 1.2 plots floor broker employment at NYSE, which precipitously fell after the exchange introduced its hybrid system.

1.4 Broker-Dealer Firms and Workers

1.4.1 US Broker-Dealers

Workers and firms in the US securities investment industry are required to be registered and obtain activity licenses from the Financial Industry Regulation Authority (FINRA). The self-regulatory agency oversees broker-dealer operations and conduct, organizes and applies examinations that enable individuals to work in certain market segments or job titles. A firm initiating sales activities in equities markets must file Form BD (Uniform Application for Broker-Dealer Registration). Employee registration records are kept through routine updates to Form U4 (Uniform Application for Securities Industry Registration or Transfer), which must be amended within 30 days of material changes (e.g., new employment, license, disciplinary action).

These records are publicly available and contain rich information on worker and firm-level data spanning decades. FINRA maintains the Central Registration Depository (CRD) which includes all securities brokers, traders, and firms currently active or that were ever registered to operate. The database also includes workers not necessarily executing stock trades or selling equities investment products, such as investment advisers, commodities traders, and several other trading professionals that also require licensing. In sum, if an individual deals with investment-related activities and has a broker-dealer as an employer, her records are available at CRD.

FINRA grants access to CRD through the search website BrokerCheck.com. The platform allows the public to retrieve records on a specific broker-dealer or trader by searching their name, registration code, or

⁹These are primarily the type of workers represented in the UBS trading floor picture circa 2005 in Figure A.4. These traders would send orders and communicate with floor brokers, for example, at NYSE’s trading floor, whose employment time series I show in Figure 1.2.

location. One example of the platform output is shown in Figure (A.5).¹⁰ To assemble the dataset, I scrape all broker-dealer records in FINRA’s CRD and homogenize each individual’s record into a comprehensive sample containing over 1.2 million individuals currently active, inactive (no longer registered), and who left the industry in the 1990s. The latter group is discarded as broker-dealers employed only prior to 1990 have incomplete records.¹¹

For each worker, I observe complete employment history by employer’s name, litigation history, years of experience, licenses, and full name. With workers’ employment records, I construct a broker-dealer longitudinal panel tracking firm location and workforce by financial occupation or job title. There are about 10,000 unique broker-dealer firms registered in the database over 1990-2019 and over 5,000 continuously operating from 2001 to 2015, which I use as my estimating period.

I complement the detailed trader information from FINRA with data from the Current Population Survey (CPS) and 3% samples of Census American Community Survey (ACS), which are less prone to sampling error than annual 1% waves and provide information on trader wages and skills, and allow me to investigate how unlicensed professionals working at broker-dealers (e.g., software engineers, data scientists) responded to trading automation adoption. Only workers of ages 18 to 64 and earning work wage are considered. Census Public Use Microdata Area (PUMA) geographies are mapped onto counties by allocating the share of population in each PUMA to a county.

1.4.2 Financial Filings

As mandated by Section 17 of the Securities Exchange Act of 1934, Rule 17a-10(a)(1), registered broker-dealers are required to submit form X-17A-5, also known as FOCUS (Financial and Operational Combined Uniform Single) Report, every fiscal year. The form is mandatory even for privately-held broker dealers and generally contains a statement of financial condition (balance sheet), financial income and operation, as well as calculation of the Uniform Net Capital Rule (Rule 15c3-1). Although some information is not publicly available, such as detailed holdings position, publicly available records always contain broker-dealer balance sheet, leverage computation, and often statement of income, as well as market value of investments, wage bill, and capital or operational expenditures in technology.

Contrary to other financial data readily available to the public, including from forms 10-K and 8-K, until 2017, X-17A-5 documents were submitted only in paper and not indexed in SEC’s Index Files. This

¹⁰Although these data are freely accessible online and are intended to offer greater transparency for regulators and customers who wish to consult professional records, the database remains largely unexplored. In finance, Egan et al. (2019) and Maggio et al. (2019) use some of FINRA’s BrokerCheck records to study financial adviser misconduct and institutional investor demand for brokerage services.

¹¹For example, details on employment history for these traders are limited, or just the last employment firm is recorded. For all individuals active after 1999, FINRA requires any changes to employment, registration, or licenses, to be reported and updated in the system within 30 days.

makes scraping and handling these forms a much more difficult task than parsing financial information from most other filings. The SEC publicly stores only PDF scans of paper filings along all other documents filed by a certain entity.

To retrieve and process financial filings from US broker-dealers, I employ a lengthy process described in detail in the Appendix. The approach first exploits broker-dealer names obtained from the FINRA records, which are then matched to SEC’s Central Index Key (CIK) codes. Next, I screen under which entity names or filing codes a broker-dealer files its relevant documents, retrieving each filing’s Document Control Number (DCN) path. This final step allows me to access and download each filing.

Finally, after obtaining over 80,000 annual X-17A-5 forms in PDF or image format spanning 2001-2020, I convert these files into readable text and extract financial information by homogenizing different reporting standard across the firms. The resulting sample and information available for balance sheet and statement of income are displayed in Figure A.9.

1.4.3 Information Technology

I obtain firm-level IT data from the marketing company Aberdeen. The CI Technology Database (formerly offered by Harte Hanks) contains detailed information on the hardware and software used in over 400,000 business sites in the US. For each company in 2004, one observes granular location, employment (overall and desk workers), industry code, company name, revenue, number of programmers, storage capacity, mainframe machines, PCs, internet access type, and IT variables along several other dimensions. Similar to Bloom et al. (2012) and Beaudry and Lewis (2014), I construct firm j ’s IT intensity measure as the number of computers per desk worker, IT_j^{fin} .

Using only firms with NAICS industry codes 523 (Securities, Commodity Contracts, and Other Financial Investments and Related Activities) and 525 (Funds, Trusts, and Other Financial Vehicles), indexed as fin , I calculate local historical IT intensity as

$$\overline{IT}_c^{2004} \equiv \frac{\sum_{j \in j(c)} \omega_j^{fin} IT_j^{fin}}{\sum_{j \in j(c)} \omega_j^{fin}}$$

where ω_j^{fin} is total employment at firm j and $j(c)$ are fin companies in county c . More details on patterns of the IT intensity measure are provided in the next section. Figure (A.6) shows that the IT intensity in the investment industry was already relatively high compared to other sectors, with an average of about one computer per worker.

1.4.4 Other Data

To lend further credibility to the main pre-policy IT measure, I construct a second pre-Reg NMS intensity measure based on local broadband availability. The data come from the Federal Communications Commission (FCC), which provides internet access information from Forms 477. The data is available for most US ZIP codes, and tracks the number of high-speed (200 kilobits per second or faster) internet providers with at least one subscriber per ZIP code. Access to high-speed internet is a minimum infrastructure requirement for firms trading electronically, especially those with high-frequency database and back-end services, such as co-location and access to price feeds from exchanges.

I map 5-digit ZIP codes onto 2010 census counties by assigning the latitude and longitude of the ZIP code centroid to a unique county. In each county, I aggregate the number of different high-speed internet providers into an alternative intensity measure $\widetilde{IT}_c^{2004} = \frac{\sum_{z \in z(c)} \psi_z providers_z}{\sum_{z \in z(c)} \psi_z}$, where $z(c)$ gives the set of ZIP codes contained in county c and ψ_z are allocation factors based on the fraction of the county area occupied by each ZIP code.

In certain specifications, I control for potential factors affecting the demand for brokerage services. Personal income per capita data comes from the US Bureau of Economic Analysis (BEA) and dividends and labor income from IRS Statistics of Income data. To construct local stock market wealth, I follow the same methodology in Chodorow-Reich et al. (2021). Specifically, I retrieve Form 1040 aggregate dividend amount in a county c and income from wages and earnings to construct local stock wealth as

$$\eta_{ct} = \left(\frac{D_{ct}}{w_{ct} L_{ct}} \right) P_t^{S\&P\ 500}$$

where $P_t^{S\&P\ 500}$ is the inverse of the dividend yield (price/dividend ratio) on the S&P 500.

1.5 Anatomy of the US Trading Industry

This section presents both suggestive evidence of the effects of Regulation NMS and details of the US broker-dealer data. I start with descriptive facts in the data.

Table 1.1 presents descriptive statistics for the top 10 counties of stock trader employment in 2004 and 2009. As expected, New York accounts for the majority of equities trading jobs in the US, with about a quarter of total employment before Reg NMS and 20% as of 2009. Other important financial centers like Chicago and Los Angeles tend to concentrate fewer broker-dealer jobs than areas in New Jersey, which has a large number of trading firms, data centers, and overall investment industry infrastructure. The pre-policy

IT intensity considerably varies across counties. Boston and New York ranked considerably higher than other top employing areas in use of PCs per worker. Since a disproportionate number of trading firms and broker-dealer individuals in the sample are in New York, later in the paper I test for the sensitivity of my main estimates to excluding the city from the analysis, as well as other top financial centers.¹²

Table 1.2 shows similar employment trends but now focusing on top broker-dealer employers. These are large investment banks, such as Morgan Stanley, JP Morgan, Merrill Lynch, and RBC Capital Markets, as well as large asset managers like Schwab, Fidelity, and Edward Jones. While the vast majority of these employers considerably decreased stock trader employment between 2004 and 2015, some firms, like JP Morgan, added about 1,000 traders. However, when compared to the expansion in total personnel, the ratio of stock traders to workers decreased from 64% to 19% — a similar trend in all other brokers. Other investment workers include any employee directly involved in the process of sales, research, management, or in advisory capacity of equities products and investment services. These do not include, necessarily, computer technicians, software developers or programmers, contract lawyers, and office personnel, unless these individuals were ever involved in the activities mentioned.

Taken together, several pieces of descriptive evidence indicate large declines in stock trader employment, even as workers in other market segments or occupations within broker-dealers did not necessarily experience the same trends. Figure 1.3 takes a closer look at entry and exit of various investment worker groups to shed light on the relative contribution of the two margins driving these aggregate patterns.

During the 2003-2007 period, increasing numbers of individuals became stock traders, a similar trend followed by investment advisors. After 2007, the entry of both groups gradually declined until becoming stable at relatively low levels. Entry in municipal bond trading is much less pronounced, trending slightly upward in the last 20 years. Commodity traders saw larger employment entry until 2012, also slowing down after that. The second panel of the figure shows annual counts of workers exiting the investment industry according to their last occupation recorded. These are net of separation effects and occupational or market segment switching within and between broker-dealer firms.

Starting in 2005, stock traders quit the industry at an increasing pace. In 2008, over 20,000 traders left broker-dealers, remaining around 15,000 a year since then, more than three times the rate of the early 2000s. Other market segments also experienced large numbers of departure.

Figure 1.4 plots data compiled from forms X-17A-5 compiled from 2002 to 2015. These are the top 30 broker-dealer firms in the period in total asset size. Not surprisingly, large investment banks including JP Morgan, Morgan Stanley, Credit, BNP Paribas, and UBS have massive balance sheets — all above \$100 billion on average. Indeed, balance sheet size is extremely concentrated in the US trading industry. Across

¹²For a spatial visualization of the evolution in stock trader employment, see (Figure A.10).

over 4,500 broker-dealers, these 30 firms alone account for over 50% of the \$4 trillion in total assets in the sector.

Figure 1.5 shows average net income of selected firms at the top and end of the performance distribution. Earnings distribution in the cross-section and time-series is very volatile. In the short cross-section illustrated, Edward Jones made on average \$200 million in net profits while MacQuare Capital lost \$100 million. In the time series, performance (and financial assets) may oscillate anywhere from hundreds of millions to several billion dollars year-over-year. Since by far the most important determinant of broker-dealer performance is some version of financial result from trading profits, it is not surprising that balance sheet in the industry presents so much variation.

1.6 Profits, Concentration, and Firm Performance

1.6.1 Empirical Strategy

I quantify the causal effects of trading automation on broker-dealers using a difference-in-differences with continuous treatment design (Card (1992)). Reg NMS fostered speed-driven competition by imposing a minimum trading speed requirement on stock exchanges and increasing the returns to fast trading. Exploiting spatial variation in local IT intensity in the finance industry, my main goal is to instrument for technology take-up, which is endogenous to firms, according to the following logic. Because technology upgrading and migration to electronic trading is facilitated by pre-existing infrastructure, I expect markets with higher IT capital stock in financial firms to be more affected by the policy.

Baseline model. In my baseline specification, I run the following model for broker-dealer firms:

$$y_{fct} = \alpha_0 + \delta (IT_c^{2004} \times \text{Post RegNMS}_t) + \gamma X_{fct} + \lambda_t + \mu_c + \varepsilon_{fct} \quad (1.1)$$

where several outcomes at the firm-level f located in county c in year t are regressed on: the pre-policy IT exposure variable, IT_c^{2004} , which is interacted with an indicator for the post-Reg NMS period, $\text{Post RegNMS}_t \equiv \mathbb{1}\{t \geq 2007\}$. To assess aggregate trends, in some regressions I use county-level outcomes, i.e., $\sum_{f \in c} y_{fct}$.

I include time-varying controls in X_{fct} that could be omitted variables otherwise left out of the model, such as local income per capita and local stock market wealth. Year fixed effects λ_t capture aggregate shocks affecting all counties with the same incidence, e.g., S&P 500 returns and the business cycle. To the extent that stock market fluctuations could exert differential effects in areas with higher investor density, the covariate for stock market wealth captures not only local consumption wealth effects as in Chodorow-Reich et al. (2021), but also the possibility that households may reduce risky asset holdings to smooth negative

labor income shocks. County-fixed effects μ_c are further included to control for unobserved determinants of employment for each county, also absorbing the level of pre-policy IT use.

Identifying assumption. In model (1.1), the key identifying assumption is that without Regulation NMS, the trend in the outcome would have remained unchanged independent of the IT intensity.

1.6.2 Validating IT Exposure: Evidence from Trading Patents

The IT exposure measure I construct intends to capture the existing level of local infrastructure in the finance industry. The relevance from assigning treatment with the measure relies on the assumption that trading automation take-up correlates with this infrastructure. One useful way to test for this assumption is to consider whether electronic trading and automation services also increase in line with local exposure.

Many trading algorithms and back-end systems are developed by specialized companies who sell these products and services to broker-dealers. This end-consumer market provides incentives for developers of trading engines and systems to file for patents to protect intellectual rights. Areas with large generation of electronic trading-related patents are likely to have a thicker market for automated trading tools, as close location to electronic network development firms is important for broker-dealers. Computer systems in trading firms undergo maintenance often. Proximity also facilitates fine-tuning and customization of automated trading strategies and off-the-shelf agency algorithms, which can be easily adapted to a firm’s particular needs.¹³

To determine patents related to trading automation, I search through all patent description records from 1976 through 2018 in the United States Patent and Trademark Office (USPTO) database, and select those related to electronic trading based on a combination of terms inventors used to link the patent use to electronic markets. I retrieve the year the patent was granted, the name and geocoordinate of the company listed as “assignee”, the investor’s name and home address. To benchmark the measure against other patents, I also construct local measures of patents related to any financial industry, and all non-finance patents. Appendix (A.1) gives further detail on the retrieval process.

To illustrate what technology type is represented by trading automation patents, consider the example of patent US7844541B2, issued in 2009 to the Chicago firm Trading Technologies International, Inc. The patent describes a “*system and method for quick quote configuration*”, allowing traders “*to quickly configure the quoting side of a trading tool, without experiencing the normal delays associated with conventional methods of quoting*” by “*automatically work[ing] an order to buy or sell a tradable object*”. The firm provides trading

¹³I use the year the patent is eventually granted and not when its application started to capture the market availability of the technology (supply-side) rather than the innovation timing. It is important to acknowledge that patents in this context do not incorporate proprietary algorithms developed within broker-dealers, which are well-kept trade secrets. Accused of stealing proprietary trading code from Goldman Sachs, former employee Sergey Aleynikov was prosecuted for years.

tools, data flow, and other services to support electronic platforms.

The first panel in Figure 1.7 shows growth rates in local patents for the trading automation, finance, and non-finance measures. Patents are aggregated from address to the county level. Trading automation patenting experienced an accelerated increase following Reg NMS. The growth rate was over 2-fold that of the overall finance industry and several orders of magnitude faster than non-finance patents, which in the counties with broker-dealers remained relatively stable over 1980-2010. The second panel in the figure decomposes the trading automation patent measure according to the IT exposure of their area. Areas with more pre-existing IT infrastructure account for almost the entire growth in trading automation patents.

These aggregate patterns are suggestive, but do not link the growth in the availability of automated trading technology to Reg NMS and IT exposure. To formally establish this link, I regress

$$\ln(\text{Patents}_{ct}^g) = \alpha_0 + \pi(IT_c^{2004} \times \text{Post RegNMS}_t) + \theta_t + \omega_c + u_{ct} \quad (1.2)$$

where the outcome measures the aggregate number of patents in group $g = \{\text{trading automation, finance, non-finance}\}$ from inventors living in county c .

Table 1.3 compares the difference-in-differences estimates across each patent group. While trading automation patenting increases by 14.7% with each additional computer per worker, there is no relationship between areas with more availability of these patents and patents in overall finance or non-finance sector.

1.6.3 Industry Profitability and Firm Performance

Table 1.4 begins by investigating how trading automation affected industry profits. Column (1) calculates county-level aggregate net income of all broker-dealers with complete records. The measure allows for both entry and exit of firms, and only discards brokers without eventual records despite being active — either because the SEC did not make the operating income part of form X-17A-5 public or the paper document copy was unreadable by OCR. The difference-in-differences estimate shows that for each additional computer per worker, local aggregate broker-dealer profits increased by \$1,009,000 a year. This is a crucial result: electronic markets made the industry in areas more likely to automate more profitable.

Column (2) considers all broker-dealers with available net income information before calculating county aggregates. The intuition here is that if missing values in column (1) are systematically associated with more or less profitable firms, the estimated effect may incorporate measurement error. When incorporating firms that do not have net income information despite having filed their forms, the estimated increase in aggregate profits is 40% larger, suggesting that the first estimate is, if anything, understating the profit growth.

The next result looks at whether larger profits were distributed equally to all firms. In column (3), I

calculate the share of local profits at the above-median largest broker-dealers (defined by total assets, which for broker-dealers largely represent financial assets). The estimated effect shows an important distributional consequence of automation. By increasing one computer per worker in a local area, largest brokers increase their profit share by 4 percentage points. Taken together, the two effects show that the trading automation boosted investment profits, but larger firms benefited more than small broker-dealers.

Broker-dealer profit. The next set of results turns to firm-level net income to compare how firm profitability responded to automation. Column (1) of Table 1.5 indicates that one additional computer per worker increases firm net income by \$98,000. This specification pools firms within the same county before and after Reg NMS, incorporating both composition changes and firms potentially not directly affected by the regulation’s introduction. Adding year dummies has little impact on the estimate. In order to understand whether the increase in industry profits stems from entry of more profitable firms, or whether automation fostered the profitability of the incumbent firms, I estimate a specification with broker-dealer fixed effects. The estimate in Column (3) indicates that incumbent firms benefited from automation, however, the coefficient is smaller in magnitude by half, increasing existing firms’ profits by \$53,000. This is estimated off brokers reporting net income at least once before and after 2007, therefore more cleanly estimating the intensive margin response to the policy. Comparing the two estimates reveals that, while incumbent firms became more profitable, a large source of increase in the industry profitability came from more efficient entrants or less profitable firms being driven out of business. Columns (5) through (8) replicate the previous estimates interacting the policy term, $IT_c^{2004} \times \text{Post RegNMS}_t$, with firm’s total assets. The triple interaction shows that larger brokers, increased profits even more relative to others in more exposed areas.

1.6.4 Survival and Entry

Next, I analyze how the policy affected industry composition, namely, how it affected broker-dealer survival and entry rates. Table 1.6 begins by analyzing how likely, on average, broker-dealer firms were to stay in business in the next 5-years. Across all firms and counties, brokers in high IT areas were about 2 percentage points more likely to survive. Decomposing this pooled coefficient according to firm size reveals several patterns. First, both the smallest and largest firms account for the higher average survival rates following automation. Employers with more than 1,000 traders were as much as 8 percentage points more likely to stay operating. Second, mid-sized firms with contingents of stock traders in line with local branches and small-operated targeting primarily financial advising were just as likely to survive regardless of automation.

1.7 Impact on Traders

This section focuses on the impact of trading automation and new technologies in the investment industry on financial workers. I combine my baseline empirical design, which exploits local variation in IT intensity to instrument for automation adoption following Reg NMS, with a standard binary difference-in-differences approach comparing stock traders to investment workers plausibly unaffected by trading automation.

1.7.1 Employment

I first modify the baseline model (1.7.1) by running worker-level regressions of the type:

$$\Pr(\text{Employed} = 1)_{ict} = \alpha_0 + \delta (IT_c^{2004} \times \text{Post RegNMS}_t) + \gamma X_{ict} + \lambda_t + \mu_c + \varepsilon_{ict}$$

where all variables are as previously defined, except for the outcome which is now a dummy equal to one if worker i is employed in year t . The specification above leverages the granularity of FINRA’s CRD records, which enables me to construct a detailed employment path stemming from a technology shock in the following manner. I consider all stock traders employed in 2007 in county c and track their employment status over time at any firm f or location. I then consider a stock trader as leaving the occupation (or “unemployed” as a stock trader) if the individual i) obtains a new license to work in a different occupation or market segment at t^* , for which I assign $\{\text{Employed}_{ict}\}_{t \geq t^*} = 0$, or if she ii) is currently not employed by any broker-dealer in t^{**} , $\text{Employed}_{ict=t^{**}} = 0$.

The first condition imposes that when switching roles or market segments, the trader leaves permanently, regardless of her actual employment status. The second allows for re-entry in any year in case the stock trader becomes employed again. Overall, this approach allows for both firm and location reshuffling, only pinpointing workers to their local of work to be treated by $IT_c^{2004} \times \text{Post RegNMS}_t$. Finally, this procedure also defines the estimating sample.

The specification above directly tests for the impact of trading automation on stock trading employment. My second, complementary approach compares outcomes of these workers to other investment professionals plausibly unaffected by Reg NMS. I focus on market segments and jobs within broker-dealers less exposed to electronic trading or task automation. Specifically, I compare stock traders to investment advisors and municipal bond traders, two markets with low penetration of automation by electronic systems, and to all non-stock traders combined to avoid potential selection bias.

Financial or investment advising (IA) is closely related to the role of stock trading — in fact, many traders also hold IA licenses. Pure investment advisors focus solely on client-relationships, most commonly

acting as sophisticated financial planners. They provide asset allocation recommendations to clients, usually using or conducting research analysis on equities products, including stocks. More recently, the emergence of robo-advisors — particularly popular among millennial investors — which offer low-cost ETFs, automatic portfolio rebalancing, and investment analysis, has also exposed IAs to technological displacement. However, robo-advisors were only introduced in 2008, and assets managed through these services were virtually zero until 2014, according to data from the Aite Group. To construct the investment advisor control group, I consider individuals holding an IA license (granted by passing the Series 65 exam, known as “*Uniform Investment Adviser Law Exam*”) who never hold a joint stock trading license.

The second group I consider are municipal bond traders, who must obtain a license from Series 50 through 54. Municipal bond trading occurs in a much different environment than stock trading. Markets are opaque, decentralized, and largely voice-based, with low penetration of electronic trading relative to stocks or derivatives. Focusing on IAs and muni bond traders also has a practical reason: these are among the best well-defined groups in FINRA’s examination and license system that allow me to exclude tasks related to stock trading. To provide as broad as possible comparison group, the third control set of workers I consider excludes all individuals ever holding a stock trading license and pools all workers in the investment industry.

If trading automation is a labor-saving technology, particularly in cognitive and manual tasks that can be subsumed by automatic rules (Autor et al. (2003)), one expects low-skill employment to respond negatively to greater adoption of electronic trading. The magnitude of this extensive-margin adjustment depends on the skill composition of affected workers and whether electronic trading is a skill-biased technology. For example, if algorithms and greater execution speed make the traders able to use these tools more productive, low-skill traders are more likely to be fired and wages of highly-skilled workers will increase. In the previous sections, I showed that the volume of stocks traded in the US — a rough output measure — increased dramatically since 2005. This growth was almost entirely attributed to automated orders.

Baseline model. Baseline results are displayed in Figure 1.8. I begin by investigating the presence of pre-trends in stock trader employment at various different levels of IT intensity. In light of the identifying assumption for the baseline specification, the presence of parallel pre-trends across different intensity values provides credibility for the design. The figure shows that traders employed in 2007 had similar trends of employment in areas with different degrees of pre-existing IT infrastructure. The second panel in the figure plots dynamic estimates of the baseline model,

$$\Pr(Employed = 1)_{ict} = \alpha_0 + \sum_{k=2001}^{2015} \delta_k (IT_c^{2004} \times \mathbb{1}\{k = t\}) + \gamma X_{ict} + \lambda_t + \mu_c + \varepsilon_{ict},$$

which interacts IT_c^{2004} with a full set of year dummies. To plausibly attribute any effects on employment to the introduction of Reg NMS, the IT intensity measure should correlate with post-policy trends in the outcome, but not prior to 2007: we expect $\hat{\delta}_k$ for $k < 2007$ to be non-significant.

Results show that stock traders experienced large declines in the probability of remaining in the occupation as a result from automation adoption. Just shortly after Reg NMS is fully implemented, stock traders are 2% less likely to be employed for each additional PC per worker. To put this magnitude in perspective, consider that based on the raw employment probability in the figure’s first panel, traders working in 2007 were almost as likely to be working back to 2001. In other words, immediately following Reg NMS, stock trader employment declines by more than the accumulated pre-level of separation.

Control workers. Figure 1.9 plots coefficients of a similar dynamic difference-in-differences model, now exploiting groups of workers employed by broker-dealers as control individuals to stock traders. Across investment advisors, municipal bond traders, and all other non-stock traders, employment of stockbrokers declined between 10 and 20 percentage points following Reg NMS. Pre-2007 estimates show no statistical difference between employment of these groups and stock traders, confirming the role of the policy in structurally changing the nature of stock markets to traders. Note that these regressions compare workers in the same county, similar to the IT intensity design. Overall, results remain unchanged when controlling for the demand of local stock brokerage services, or including broker-dealer fixed effects.

1.7.2 Other Broker-Dealer Workers

As broker-dealers replace stock traders with electronic systems and automated execution, do they increasingly employ other professionals? To analyze the extensive labor response of trading automation adoption, I turn to other occupational groups employed in the investment industry. Using county-level employment of five groups of workers, I run the following regressions

$$\ln(employment)_{ct}^o = \alpha_0 + \delta(IT_c^{2004} \times \text{Post RegNMS}_t) + \gamma X_{ct} + \lambda_t + \mu_c + \varepsilon_{fct}$$

which use log employment levels of five worker groups o , Quants, Financial Analysts, Investment Advisors, Tax and Audit, and Brokerage Clerks.

The first group, Quants, includes computer scientists, programmers, developers, computer hardware and software engineers, and IT support specialists. With higher investment in IT infrastructure and the need for development, maintenance, and support, broker dealers may increase the demand for within-firm presence of these professionals. The second group, Financial Analysts, also usually involves highly-educated individuals considered “quants”, but whose skills are employed in financial research, modelling, and risk

and portfolio analysis. New data technologies in the investment industry likely expand the role of these professionals. Investment advisors represents the same occupational group that I observe with precision and granularity from FINRA’s records, and we should expect null employment effects. I include this group as a check for using Census employment data in the investment industry, similar to tax and audit professionals.

The final group, Brokerage Clerks, provides an important related occupation to stock trading. The US Bureau of Labor Statistics describes brokerage clerks as performing duties including “*writing orders for stock purchases or sales, computing transfer taxes, verifying stock transactions, accepting and delivering securities, tracking stock price fluctuations, and keeping records of daily transactions and holdings*”. The need for these routine manual tasks supporting stock trading was likely reduced by electronic markets. Taxes are computed automatically, trading orders are no longer written by hand, transactions are recorded in the electronic tape, just like daily positions and holdings are netted electronically.

Figure 1.10 shows that trading automation slightly increased the employment of quants and financial analysts in broker-dealers, although estimates are imprecise and not statistically significant. One reason may be that firms with greater exposure to automation have higher propensity to outsource these services. Another reason is that, while Census employment counts are useful for this exercise, they are likely to suffer from large sampling error in narrow occupational cells, such as “Software Developers, Applications and Systems Software” in investment firms.

The employment of brokerage clerks decreases dramatically following Reg NMS. For each additional computer per worker, the employment of workers performing manual routine-based tasks related to stock trading reduces by almost 50%. This is a greater employment elasticity to automation compared to stock traders and in line with the decline of clerical employment in other industries due to automation.

1.7.3 Skill Premium

The financial sector has experienced a dramatic increase in pay and education level since the 1980s. Most of this trend is driven by a subset within finance, including traders. Reshef and Philippon (2012) show that while credit intermediation, insurance, and trading services had a wage premium of about 50% over the nonfarm private sector in 1970, average trading wages were 4-fold those in the wide economy by 2010 and credit and insurance wage premia remained relatively constant.

Various papers have proposed theoretical explanations for this surge in financial compensation, ranging from overinvestment in expertise (Glode et al. (2012)), to concentration of workers in the high end of the skill distribution (Glode and Lowery (2016), Célérier and Vallée (2019)) in financial firms and competition for talent (Thanassoulis (2012)). Trading in electronic markets involves a considerably higher level of skill and

technical ability than trading did until the late 1990s. If electronic trading complements the skills of highly educated traders, one would expect the wage skill premium between this group and stockbrokers without college degree to widen after Reg NMS. I test formally for the presence of a skill-biased channel by running a difference-in-difference-in-differences (DDD) regression:

$$\begin{aligned}\ln(Wage_{ict}) = & \alpha_0 + \delta_0 \text{Post RegNMS}_t + \delta_1 (\text{Post RegNMS}_t \times \text{Skilled}_i) \\ & + \delta_2 (IT_c^{2004} \times \text{Post RegNMS}_t) + \delta_3 (IT_c^{2004} \times \text{Post RegNMS}_t \times \text{Skilled}_i) \\ & + \gamma X_{ict} + \lambda_t + \mu_c + \varepsilon_{ict}\end{aligned}$$

where the coefficient δ_3 in the reduced-form equation measures how Reg NMS affected high-skill (college or more) brokers in areas with historical intensive IT use differentially from low-skill workers. The outcome is log hourly wage and personal characteristics in X_{ict} include gender, experience (linear and squared), and number of children. If electronic trading is a skill-biased technology, we expect the wage differential to be positive, $\delta_3 > 0$.

The DDD model uses pooled census data instead of a longitudinal firm-worker panel as in the baseline model. To properly identify δ_3 , I rely on the stronger assumption that within county and netting out controlled observable characteristics, brokers in the repeated cross section do not differ systematically over time. The estimate is reported in column 4 of Table 1.7. The positive effect shows that Reg NMS increased hourly wages of high-skill workers in areas with an additional computer per broker by 9% compared to low-skilled workers. This translates as roughly \$10,000 annually at the mean.

1.7.4 Robustness Checks

This subsection conducts a series of tests to verify the plausibility of the parallel trends and other possible model assumption violations, and whether the baseline model is robust to a different intensity measure. An analysis of the significance of alternative explanations, including the Great Recession and online brokerage platforms is delayed until Section (1.8).

In additional results, I interact county and time effects to allow for flexible trends potentially driving local broker employment. Second, I remove year dummies from (1.7.1) and add a linear time trend and a trend interacted with IT_c^{2004} . This specification tests whether the policy still affects employment after allowing for a common linear trend across counties. The estimate in column 2 of (1.7) is remarkably similar to the coefficient from the baseline model. Next, (1.3) interacts the linear trend with county fixed effects, in

the following specification:

$$Employed_{ict} = \alpha_{0c} + \mu_c t + \delta (IT_c^{2004} \times \text{Post RegNMS}_t) + \gamma X_{ict} + \lambda_t + \varepsilon_{ict} \quad (1.3)$$

which allows counties with different levels of IT intensity to follow different trends. In other words, (1.3) nets out changes in local broker-dealer employment caused by Reg NMS from each county's (potential) own employment trend. Results are displayed in column 3 and again sustain the inference from my main estimates. Taken together, these specifications tests formally confirm that the parallel trends assumption is unlikely to be violated in my baseline model.

Alternative intensity measure. As an additional check, I use the broadband availability measure $\widetilde{IT_c^{2004}}$. While this measure is less relevant than the IT intensity using computers per desk worker, it tests for the importance of *minimum* technology in providing heterogeneous levels of local exposure to Reg NMS. Many broker-dealers occupy the same buildings to take advantage of pre-existing high-speed internet infrastructure. Back-end communication systems and networks in trading firms to access price feeds and co-location rights offered by exchanges require advanced internet services, which might make electronic trading adoption in areas with larger high-speed internet easier. Results in column 1 are in line with the baseline estimates.

Placebo period. Next, I verify whether pre-policy differences between firms in high IT areas and the ones in low IT counties might be driving the baseline employment results. A simple test for this is to “ignore” the introduction of Reg NMS by replicating the baseline model prior to 2005. Over 2001-2004, I consider 2004 as a placebo policy intervention (i.e. “Post RegNMS”_t $\equiv \mathbb{1}\{t \geq 2004\}$) and run DiD regressions up until Reg NMS became fully effective. Columns (1) through (3) in Table 1.8 confirm that potential observed and unobserved heterogeneity between broker-dealers in areas with varying degrees of IT intensity are unlikely to drive the employment effects I attribute to Reg NMS.

Anticipatory IT investment. A potential concern with the IT intensity measure is that if the county-level number of PCs per worker would reflect firms' ongoing shift toward electronic trading, values of IT_c^{2004} might correlate with unobserved trends in technology upgrading. This would in turn exacerbate the effects on employment attributed to Reg NMS in my baseline model, since a county with high PC-worker ratio could already be automating trading at a faster rate.

The CI Technology Database offers a variable tracking the growth of a specific technology for every firm, which I denote by ΔIT_j^{fin} . For example, if firm j had 30 computers per worker in 2004, $\Delta IT_j^{fin} > 0$ indicates whether the firm had been increasing the number of computers until the observed 30 units. I test whether the trend in IT upgrading pre-2004 predicts employment growth in columns (4) and (5) of

Table 1.8. The non-significant coefficients indicate that local trends in IT upgrading are uncorrelated with future employment changes.

1.8 Alternative Explanations

This section discusses and tests the role of two potentially confounding factors to my empirical design. First, the Great Recession exerted differential effects in local labor and housing markets, which could affect the demand for brokerage services or the supply-side of trading services to a greater extent in areas with higher pre-policy IT intensity. Second, the rise of online discount brokerage services that started in the early 2000s could also affect the demand of customers for retail and perhaps even institutional brokerage firms.

1.8.1 The Great Recession

The economic downturn during 2007-2009 impacted in particular housing markets and highly-leveraged financial institutions, to a great extent because of mortgage-backed securities and CDO holdings. At a high level, employment losses, credit crunch, and deterioration of household balance sheets during the Great Recession could all affect the demand or supply of brokerage services. One of the controls included in the baseline model — local stock market wealth — is a first attempt to control for spatial heterogeneity in shocks to individuals’ income from stock market holdings relative to labor wages. The analysis below intends to capture several specific channels through which the Great Recession could bias my main results.

Omitted variable bias in the treatment intensity measure. The first test investigates the possibility that the pre-Reg NMS IT infrastructure measure correlates with unobservable characteristics that made the effects of the financial crisis particularly acute in more exposed areas.¹⁴ Under the assumption that this potential omitted variable bias is not exclusive to broker-dealers, estimates in the first panel of Figure 1.11 provide exhaustive placebo-type regressions. Each coefficient captures how local employment in a given sector changes in areas with larger trader employment losses after Reg NMS. Although I keep the intervention year in these regressions as 2007, since Reg NMS has no impact on other sectors the “after” period may be interpreted as the confounding effect of the Great Recession picked up by the Reg NMS dummy, treating different areas given the IT measure.

Across almost 180 sectors, only 13 estimates are statistically significant. This indicates that areas with larger automation exposure in the investment industry do not disproportionately experience employment changes in other sectors, either because of local effects from the Great Recession or other systematic events

¹⁴Perhaps high pre-existing IT investments were part of firm conduct that included over-leverage, even though IT levels in 2004 negatively correlate with prior trends in investment.

around the implementation of Reg NMS. Another interpretation of these regressions is that there is no evidence that counties with high IT exposure have any particular employment drivers.

Did the Great Recession accelerate trading automation? The second panel of Figure 1.11 tests a more subtle possibility. Suppose that Reg NMS indeed led broker-dealers to automate and substitute labor for capital, resulting in stock trader employment losses, but this process was accelerated or intensified by the Great Recession. The possibility of recessions accelerating routine-biased technological change (Hershbein and Kahn (2018)) would imply that part of the effect or timing I attribute to Reg NMS stems from unrelated recessionary forces. To test this conjecture, I calculate industry-specific IT measures for each county and verify whether, for example, areas with higher IT levels in the healthcare sector experience faster employment changes in the industry post-2007. Out of 23 broad sectors, only one has a *positive* employment pattern after 2007 in counties with higher IT infrastructure in 2004. Taken together, these results point to a lack of evidence that industry-specific IT infrastructure measures correlate with subsequent changes in employment during the Financial Crisis.

Other financial workers. Given that Reg NMS should only directly affect broker-dealer individuals trading stocks, one could test to whether other financial workers plausibly unaffected by the provision also experienced similar employment losses during 2001-2015. The idea with this exercise is to a class of financial workers that could be affected *only through* the Great Recession to test whether the “post-policy” period is really picking up the effects from the financial crisis. I use local employment (at the county level) in the financial sector excluding security, commodity brokerage, and investment companies. The data use employed individuals between ages 18 and 64 working in the following industries: banking, saving institutions, credit unions, credit agencies, and insurance companies from March supplements of Current Population Surveys (CPS) from 2001-2015. Note that while this exercise is complementary to the difference-in-differences results in the previous section that compare stock traders to investment advisors, bond traders, and all other financial workers employed in broker dealers, here I consider all workers in any financial service.

Results in column (1) of Table 1.9 indicate that employment in the finance-wide industry actually increased post-Reg NMS in counties with higher historical presence of computers per worker. Of course, these effects should not be attributed to Reg NMS. Instead, they point out to a lack or even positive trend in employment of non-stockbroker workers during 2007-2015 relative to the 2001-2006 period. There is no reason to expect differential effects of the Great Recession for these workers compared to the Series 7 individuals used in the baseline analysis given that the Reg NMS channel through electronic trading is shut off for, say, real estate or insurance agents.

Key Financial Crisis events. Leveraging the high frequency and regular updates of FINRA records, an additional stringent analysis I conduct inspects employment trends following key events of the Great

Recession. Figure 1.12 displays monthly changes in employment status of stock traders who were working in October 2007, when Reg NMS is fully implemented. These raw changes in employment intend to provide a visual inspection of potential clusters following, for example, the bankruptcy of Lehman Brothers.¹⁵

The figure shows that from 2006 to October 2007, monthly changes in stock trader employment followed relatively constant rates of 2%. The period incorporates many of the early signs of the financial crisis, including New Century's bankruptcy early 2007. From October 2007 to March 2008, prior to the failure of Bear Stearns, monthly employment changes increase, achieving almost 10% in February 2007. While the move also captures seasonality during the month, its magnitude is still larger than subsequent February observations. After that, around the collapse of IndyMac, Lehman Brothers, and Citigroup, and the historical Dow Jones drop, stock trader employment continues to decline at rates ranging from 2%-6% per month, without clear clusters other than typical seasonality. From 2010 and beyond, monthly changes in employment status of traders employed in October 2007 continue to decline at similar rates.

These aggregate patterns show that during most of the Financial Crisis, stock broker employment declined at relatively constant rates, without large movements following key supply-side events. The largest short-term employment decline occurs immediately after the full implementation of Reg NMS. Despite some of this movement being attributable to seasonality in the beginning of the year, there is little descriptive evidence that stock trader employment trends align with other events during the same period. For example, the drop in employment in January-February 2011 of 4.5%, well after the official end of the Great Recession, has the same magnitude as in the beginning of 2009.

Investment banks. To formalize the points above, next I replicate the baseline stock trader employment model by excluding large investment banks — Goldman Sachs, Merrill Lynch, Morgan Stanley, JP Morgan, and Lehman Brothers. Dropping the *Big 5* in column (2) of Table 1.9 produces no changes in the estimate of interest, even though these institutions were particularly affected the the crisis (Longstaff (2010), Ivashina and Scharfstein (2010), Aragon and Strahan (2012)). Of course, the effects of the balance sheet deterioration from these banks spilled over and contaminated various financial institutions. While completely controlling for such effects is not possible, the robustness of the baseline results to excluding those banks plausibly affected the most by the financial crisis lends further credibility to the role of Reg NMS in accounting for the employment dynamics of stockbrokers.

Omitting New York. Given that New York accounts for nearly 25% of stock traders in the US and the area was among the ones with highest pre-policy IT intensity levels, one might worry idiosyncratic factors related to the financial center drive my results. In column (3) of Table 1.9, I omit traders working

¹⁵Even if such clusters exist, my empirical designs are only potentially contaminated by these events to the extent that their intensity correlates with 2004 levels of IT infrastructure or that they did not affect investment advisors, bond traders, and other financial workers.

at New York brokers, which result in estimated effects nearly identical to the ones when NYC is included.

Housing crisis and household balance sheet. Previous work by Mian et al. (2013) has shown that the 2006-2009 housing collapse affected consumption differentially across the US depending on the local composition of household balance sheets. One potential implication of this decline in housing net worth is that perhaps the demand for risky assets decreased to mitigate income effects. I use a repeat-sales home price index from the Federal Housing Finance Agency (FHFA) to answer whether the negative effects on stockbroker employment remain after controlling for the deterioration of home values. The estimate in column (4) remains unchanged relative to the baseline model.

1.8.2 Online Brokerage Services

To conclude the empirical analysis, an additional test checks the possibility that growth in local demand for online brokerage services impacted broker-dealer firm employment. Online brokerage firms, usually discount brokers, offer online infrastructure for individual investors to have access to financial markets without relying on traditional broker dealer firms. Analyzing the impact of this type of online financial service is challenging if such platforms do not provide granular data on costumer accounts over time and across space. Consider for instance E*Trade, one of the first and major online brokers in the US. E*Trade only publishes the aggregate annual number of active accounts in the US. Measuring the influence E*Trade on traditional broker-dealers requires the spatial allocation of these aggregate numbers.

I circumvent this issue by exploiting spatial variation in relative search traffic for “E*Trade” using Google Trends to allocate US E*Trade accounts to each county in my sample. Other papers have used Google Trends in distinct contexts to instrument for local demand or awareness for certain services, such as AirBnb (Barron et al. (2018)). For each year, Google assigns a value of 100 to the MSA with the largest number of searches of the term “E*Trade” and then ranks other MSAs relative to the top search area. This ranking generates spatial variation in the intensity of the “interest” for the searched term, which I then use to allocate aggregate brokerage accounts available from E*Trade 10-K forms to each county in my analysis. In column (5) of Table 1.9, including this local measure of demand for online brokerage services as an additional control in the baseline specification does not affect the estimated effects of Reg NMS.

1.9 Conclusion

The migration to electronic markets and trading automation in the early 2000s was a crucial technological change in the history of stock markets. The trading industry made investments in IT physical infrastructure and software of over \$100 billion and markets became dominated by algorithms and speed competition. This

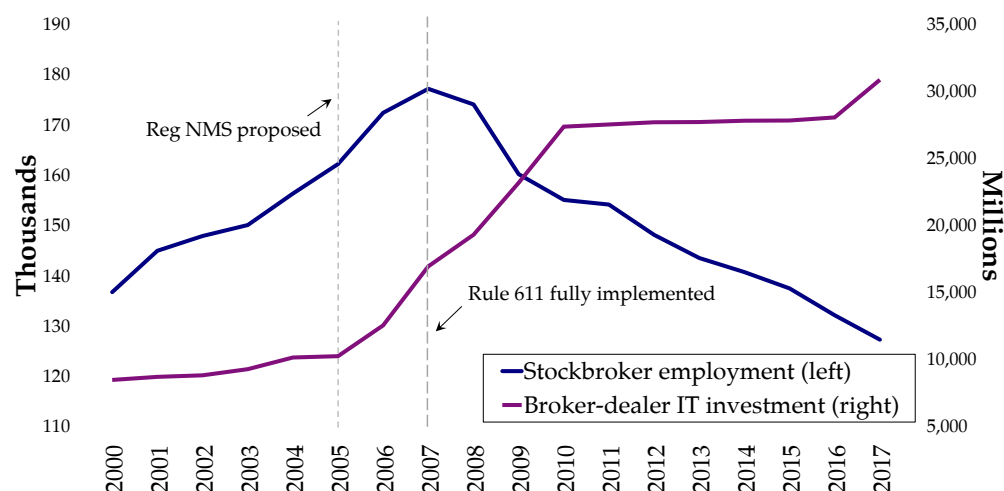
paper studies how automation affected broker-dealer firms and stock traders, investigating winners and losers in new financial markets.

I estimate an average return to automation to trading firms of \$98,000 in annual net profits for each additional computer per worker. In the aggregate, even relatively small investments in technology lead to an increase in industry local profits of \$1 million. Returns to automation accrue unevenly. The same technology investment in broker-dealers with total assets above \$80 billion are anywhere from \$50,000 to \$380,000 more profitable than in a \$100 million asset firm. As a consequence, profit concentration in large firms increases 4 percentage points. Automation also affected the composition of trading firms along two margins. First, fewer broker-dealers enter the investment industry. In addition, large and small employers, who are more likely to cater to high net worth individuals and engage in proprietary trading, experience higher survival rates.

I also document large displacement effects on stock traders, whose job task content changed dramatically with electronic trading. I find that these financial professionals become from 2 to 20 percentage points less likely to remain employed, even when compared to other workers in market segments that did not experience automation. This large elasticity of employment with respect to computer capital aligns with extensive-margin adjustments of labor when workers perform cognitive and manual tasks that can be automated (Autor et al. (2003)).

Overall, my findings offer evidence supporting that large trading firms, who provide services to a small portion of investors, appear to have benefited from trading automation. An avenue for future research is to investigate whether small investors are necessarily worse-off. New trading and data technologies increased market access and cheapened fees investors face. If the firms who failed to capture the benefits from automation predominantly served this group of individuals, investors may be better off by switching to different brokers or taking advantage of new investment products made possible by electronic markets.

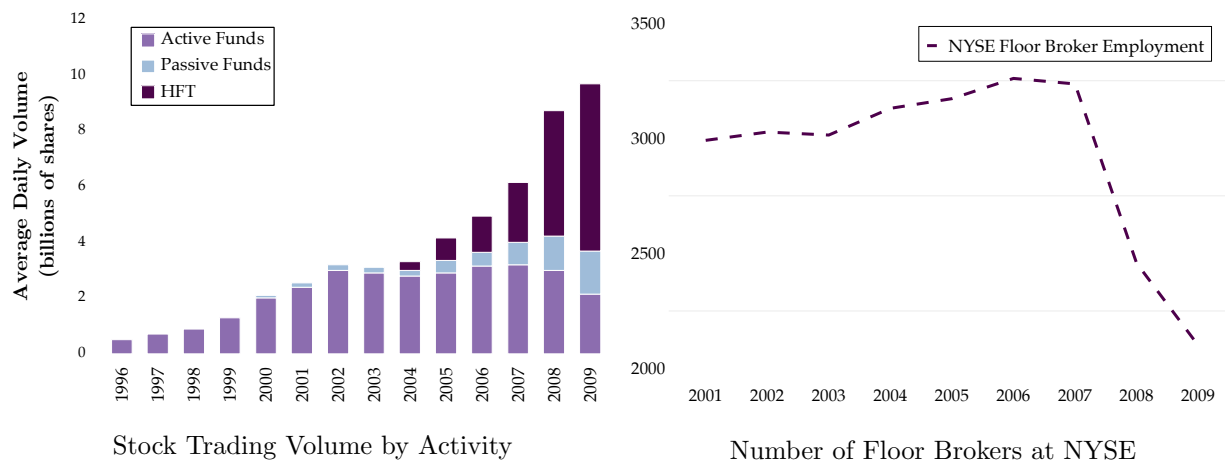
1.10 Figures and Tables



Following Reg NMS, IT investment in US trading firms increased three-fold, while stock trader employment fell by almost 60,000 jobs.

Notes: This figure compares the evolution of IT capital stock (in the securities, commodities, and investments industry) and employment of stock traders in US broker-dealers. The employment series aggregates all employed individuals registered in FINRA's Central Registration Depository (CRD) with an active registration trading stocks in a given year. Dashed lines indicate Regulation NMS's rollout period, with full implementation in 2007. The IT capital stock series is constructed following in Reshef and Philippon (2012), except that the series plotted is not normalized by other sectors.

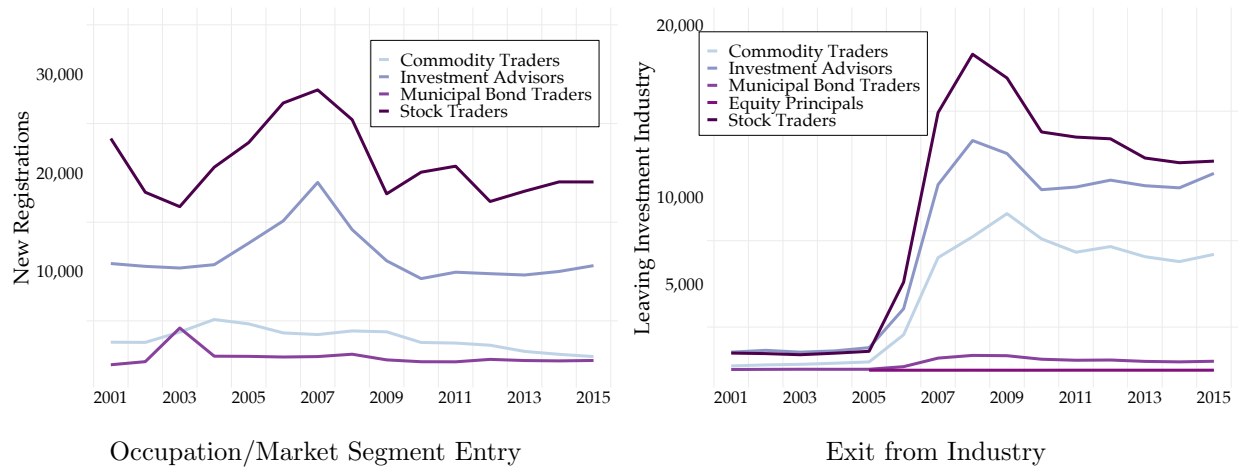
Figure 1.1: Trends in Jobs and Technology Capital in Trading



Left: High-frequency trading (HFT) accounted for over 60% of daily volume traded in US stocks in 2009, more than doubling total market output since 2004. **Right:** Following the introduction of the Hybrid Market at NYSE, floor broker employment sharply declined.

Notes: The figure on the left shows data compiled by Tabb Group and Credit Suisse on the distribution of average stock traded volume by market participant segment. Active and passive funds include money managers and investors. High-frequency traders are a subset of algorithmic trading (AT), and provide an underestimate of the contribution of electronic submission and automated quoting and executing, some of which are not considered “high-frequency”. US daily traded volume in 2009 was over 10 billion shares. The figure on the right plots data extracted from FINRA’s CRD database, as described in the main text. The series is constructed using individuals holding a valid Series 25 license, known as “NYSE Trading Assistant Exam”, which was required from trading assistants (post and booth clerks) at NYSE’s trading floor. FINRA records show when the individual had employment terminated, left the securities industry, or changed market segment or occupation. The series thus tracks annual counts of unique NYSE floor brokers. For a visual depiction, see Figure A.4.

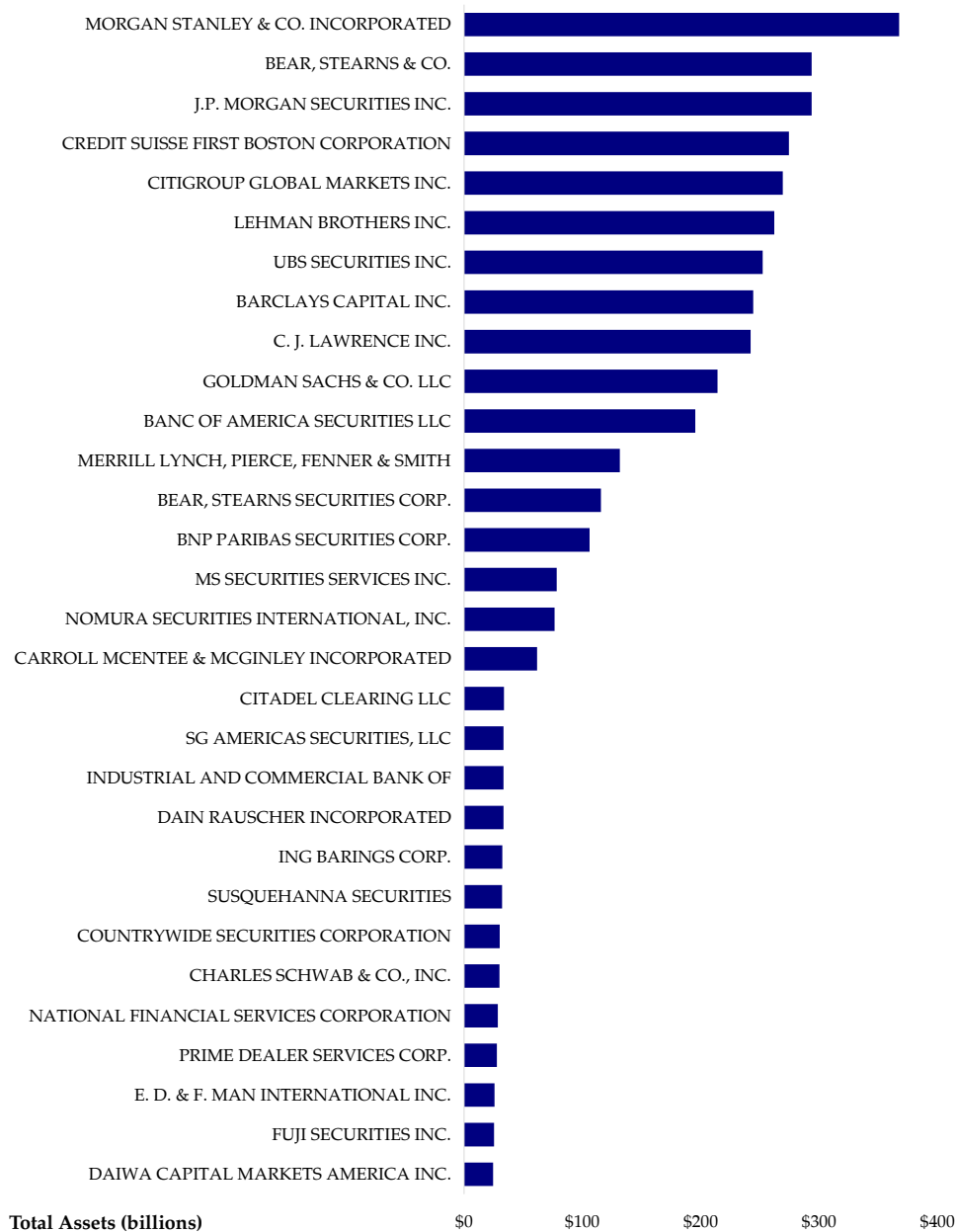
Figure 1.2: Growth in US Stock Trading Activity and Employment of NYSE Floor Brokers



Left: 20,000 new stock traders have entered the industry a year since 2009. **Right:** Stock traders left the investment industry at historically high numbers starting in 2005.

Notes: These figures compare industry entry and exit of stock traders to several categories of other workers in the investment industry. Entry is measured by the year an individual obtains the necessary license to trade or work in the referred market segment, combined with a valid employment status. Exit is measured as the year an individual permanently leaves any segment in equities industry, where I assign the type of trader to the most recent license she obtained before leaving the sector. Data come from FINRA's CRD repository. Details on sample construction and definitions of each group of trader can be found in the main text.

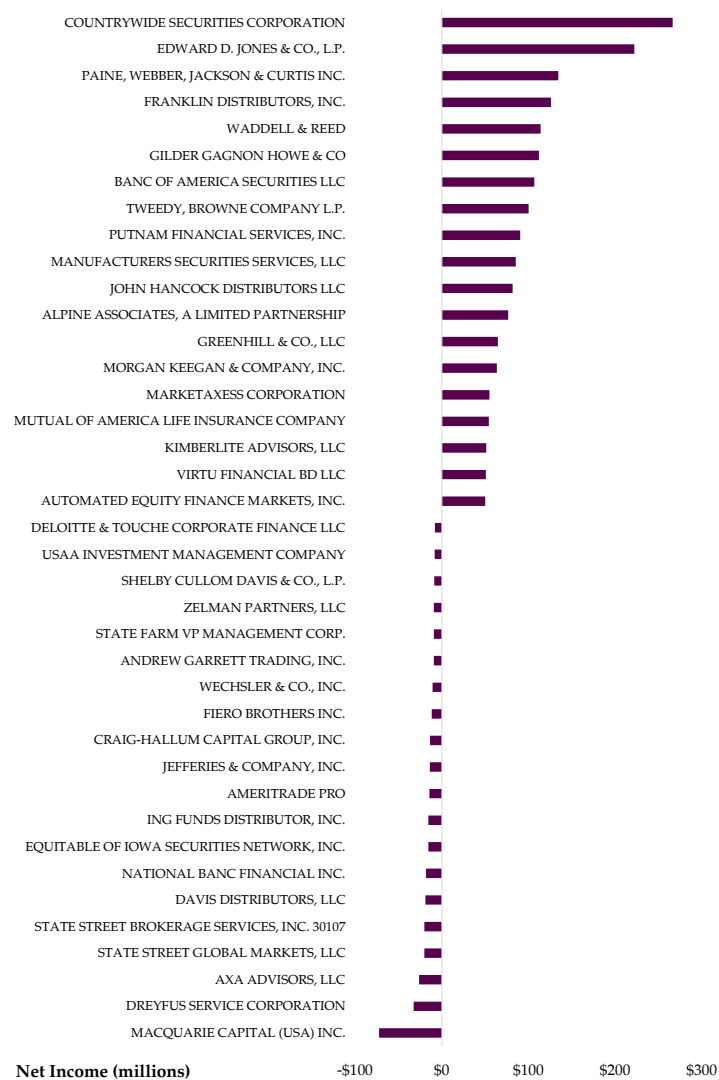
Figure 1.3: Occupation Entry and Investment Industry Exit



Book size of the trading division from large investment firms tops \$300 billion, largely corresponding to cash on hand, market positions, and receivables from clearing broker and customers.

Notes: The figure shows total asset values for the 30 largest broker-dealers in US equities markets. Bars represent the average of annual book sizes from 2002–2014. Micro data is constructed from forms X-17A-15 submitted by every broker-dealer every year, extracted, processed, and parsed following the details in the main text. For comparison, the total value of assets for broker-dealers was \$4 trillion over the period.

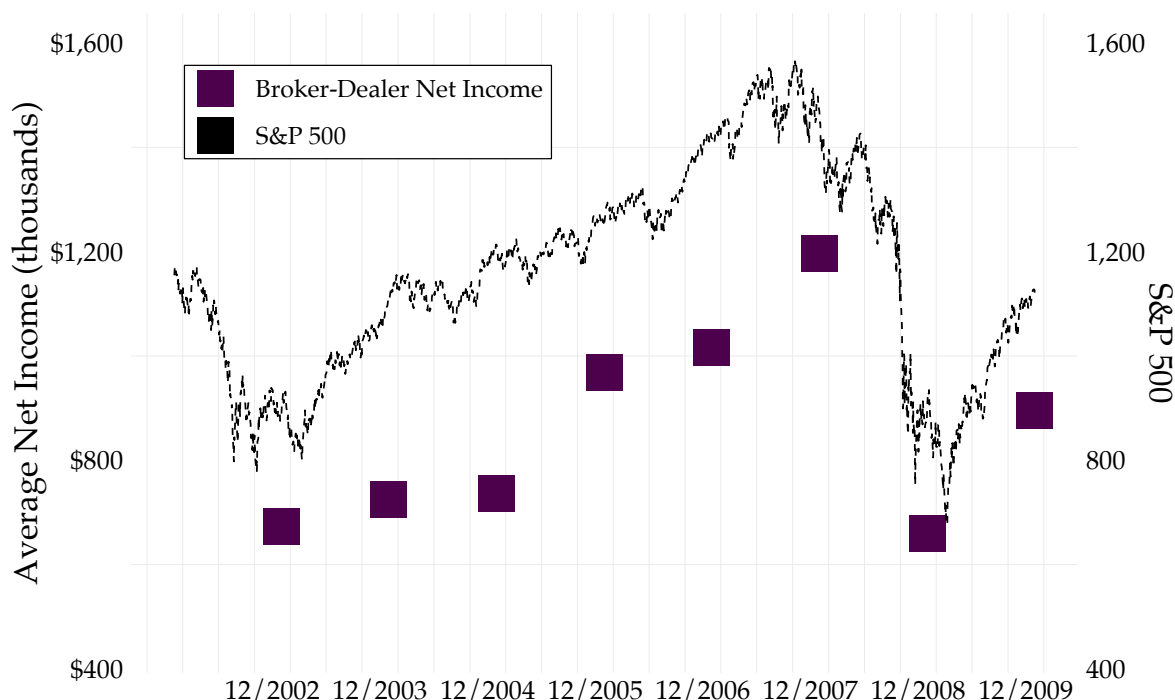
Figure 1.4: Total Assets of Top Broker-Dealer Firms



The average distribution of broker-dealer net income ranges from a loss of over \$100 million to \$300 million in profit.

Notes: The figure shows net income of selected broker-dealers with the best average gains and losses from 2002—2014. Micro data is constructed from forms X-17A-15 submitted by every broker-dealer every year, extracted, processed, and parsed following the details in the main text. Common lines in broker-dealers statement of operations/income include commissions, investment brokerage fees, advisory and transaction services, private placements, incentive fees, research revenue, handling and processing fees, interest and dividends as revenue sources, and commissions, employee compensation and benefits, clearing and transaction costs, advertising, communications and technology, membership and clearing fees, travel and entertainment, regulatory fees and expenses, and office and occupancy expenses as main sources of operating expenses.

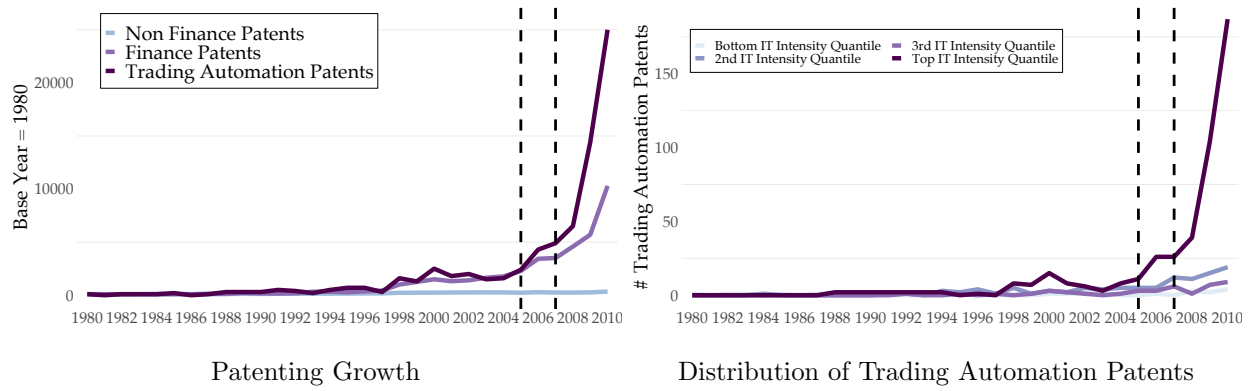
Figure 1.5: Range of Broker-Dealer Performance



Average broker-dealer performance is highly correlated with market returns.

Notes: The figure plots annual averages of broker-dealer net income from individual X-17A-5 filings extracted, processed, and parsed following the details in the main text. Common lines in broker-dealers statement of operations/income include commissions, investment brokerage fees, advisory and transaction services, private placements, incentive fees, research revenue, handling and processing fees, interest and dividends as revenue sources, and commissions, employee compensation and benefits, clearing and transaction costs, advertising, communications and technology, membership and clearing fees, travel and entertainment, regulatory fees and expenses, and office and occupancy expenses as main sources of operating expenses. The other series shows daily closings of the S&P 500.

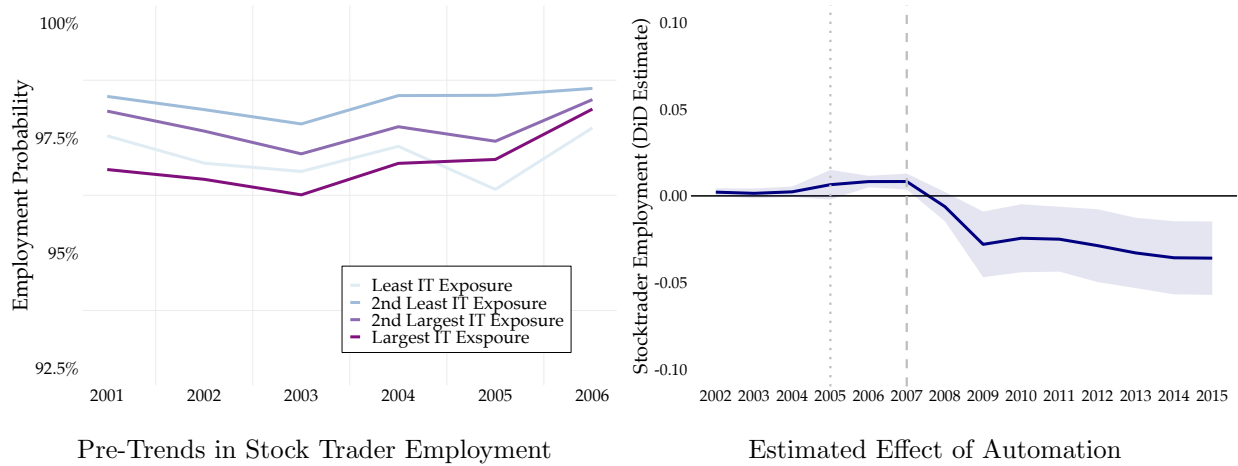
Figure 1.6: Broker-Dealer Profitability and Stock Market Return



Left: *Trading automation patenting grew much faster than patents in overall finance or in other sectors.*
Right: *The growth in trading automation patenting occurred mostly in counties with high pre-existing IT infrastructure.*

Notes: The left panel shows the growth of i) trading automation patents, ii) finance industry patents, and iii) non-finance sector patents in the US with 1980 as the base year for counties with broker-dealer firms. The figure on the right zooms in on trading automation patents and plots growth trends according to different levels of local IT exposure.

Figure 1.7: Trading Automation, Overall Finance, & Non-Finance Patents



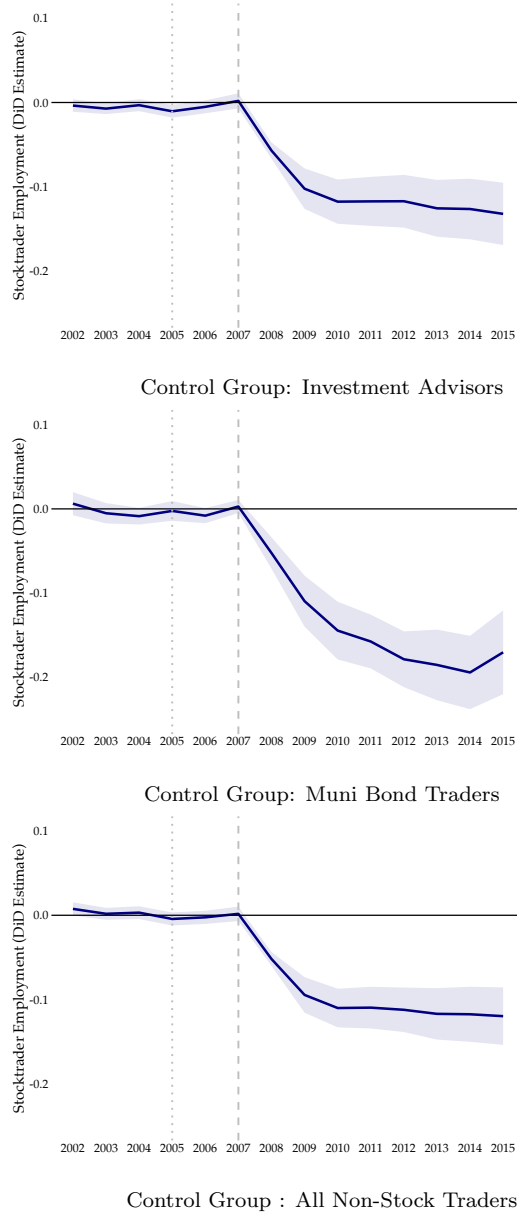
Left: *Employment propensities of stock traders in areas with high IT exposure followed similar trends relative to traders in areas with low exposure.* **Right:** *For each additional PC/worker before Reg NMS, stock traders become 2 p.p. less likely to be employed on average.*

Notes: The first panel shows pre-trends in the probability of employment of stock traders who worked in 2007. Each line represents the average outcome in areas with different degrees of exposure to trading automation. The second plot shows dynamic difference-in-differences estimates of the model

$$\Pr(\text{Employed} = 1)_{ict} = \alpha_0 + \sum_{k=2001}^{2015} \delta_k (IT_c^{2004} \times \mathbb{1}\{k = t\}) + \gamma X_{ict} + \lambda_t + \mu_c + \varepsilon_{ict}$$

which compares the probability of employment of stock traders exposed to different degrees of IT infrastructure. Shaded areas plot 95% confidence intervals obtained from standard errors clustered at the county level. The first dashed line represents the introduction of Reg NMS in 2005, while the second shows the regulation's full implementation in 2007. Data come from FINRA's Central Central Registration Depository (CRD).

Figure 1.8: Impact of Trading Automation on Stockbroker Employment



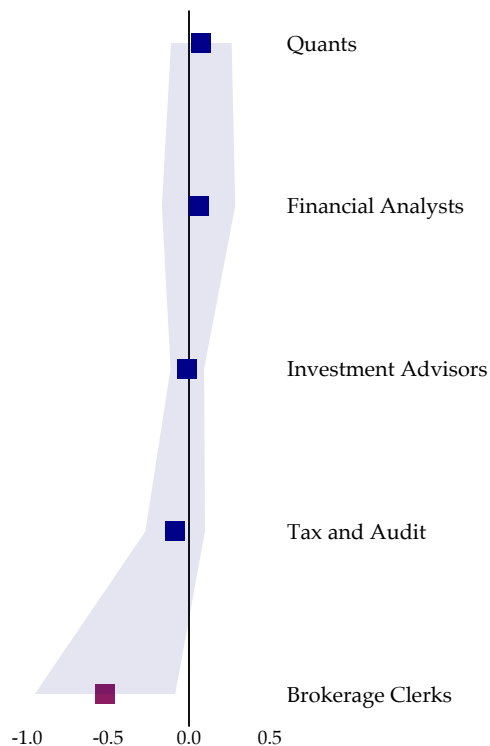
Following Reg NMS, stock traders were 10-20 percentage points less likely to be employed than investment advisors, municipal bond traders, or any other financial professional not trading stocks working for broker-dealers

Notes: These figures show DiD estimates of the model

$$\Pr(\text{Employed} = 1)_{ict} = \alpha_0 + \sum_{k=2001}^{2015} \delta_k (\text{Stock Trader}_{ict} \times \mathbb{1}\{k = t\}) + \gamma X_{ict} + \lambda_t + \mu_c + \varepsilon_{ict}$$

which compares the probability of employment of stock traders to other financial workers employed at broker-dealers in the same area. Shaded areas plot 95% confidence intervals obtained from standard errors clustered at the county level. The first dashed line represents the introduction of Reg NMS in 2005, while the second shows the regulation's full implementation in 2007. Data come from FINRA's Central Registration Depository (CRD). For definitions of treatment and control groups, see main text.

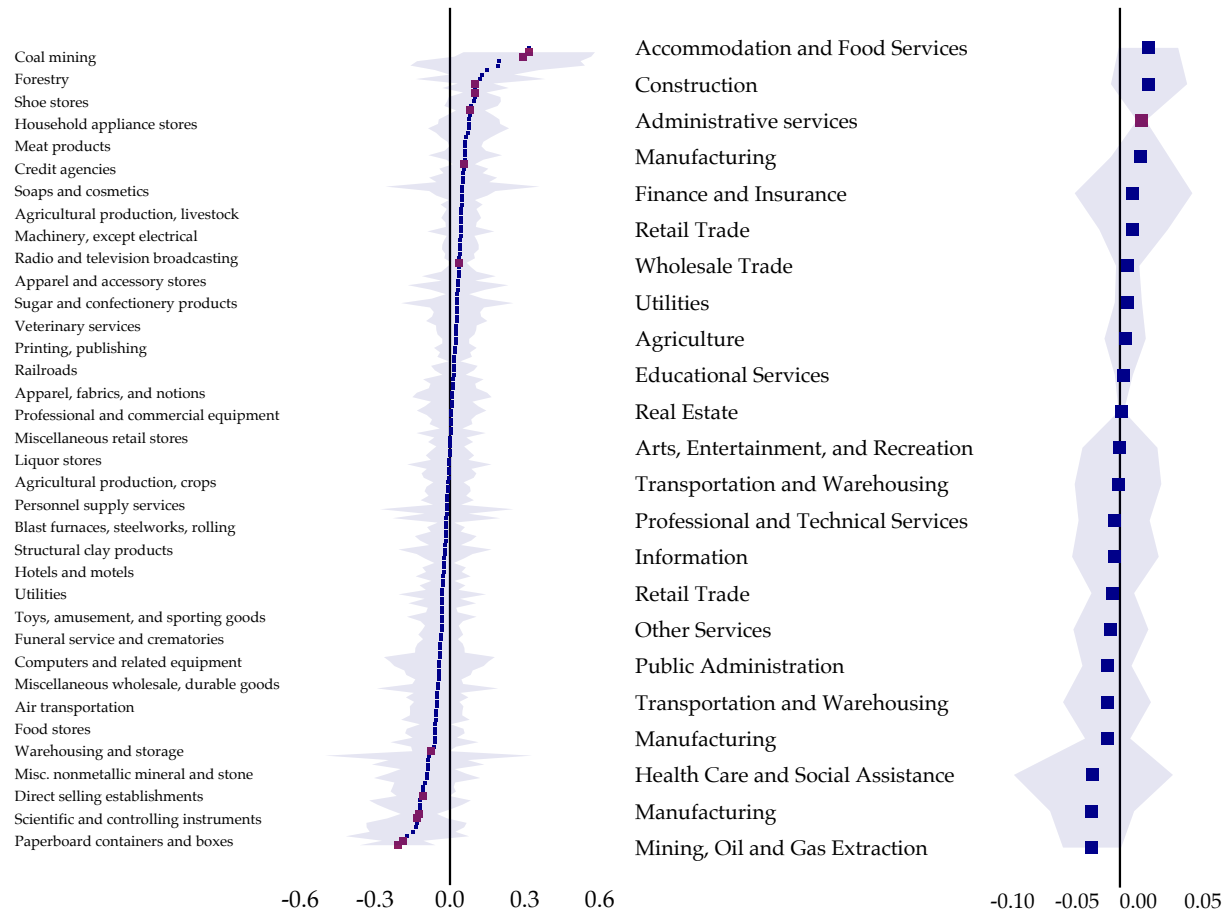
Figure 1.9: Impact on Stockbroker Employment vs. Other Investment Workers



Broker-dealer employment of other workers remained relatively constant following trading automation, except for workers performing clerical tasks related to manual stock trading

Notes: These figures show difference-in-differences estimates of county-level log employment of five occupational groups — quants, broadly encompassing computer and software engineers, programmers, and other IT professionals, financial analysts, investment advisors, tax and audit professionals, and brokerage clerks — on the 2004 IT infrastructure and Reg NMS.

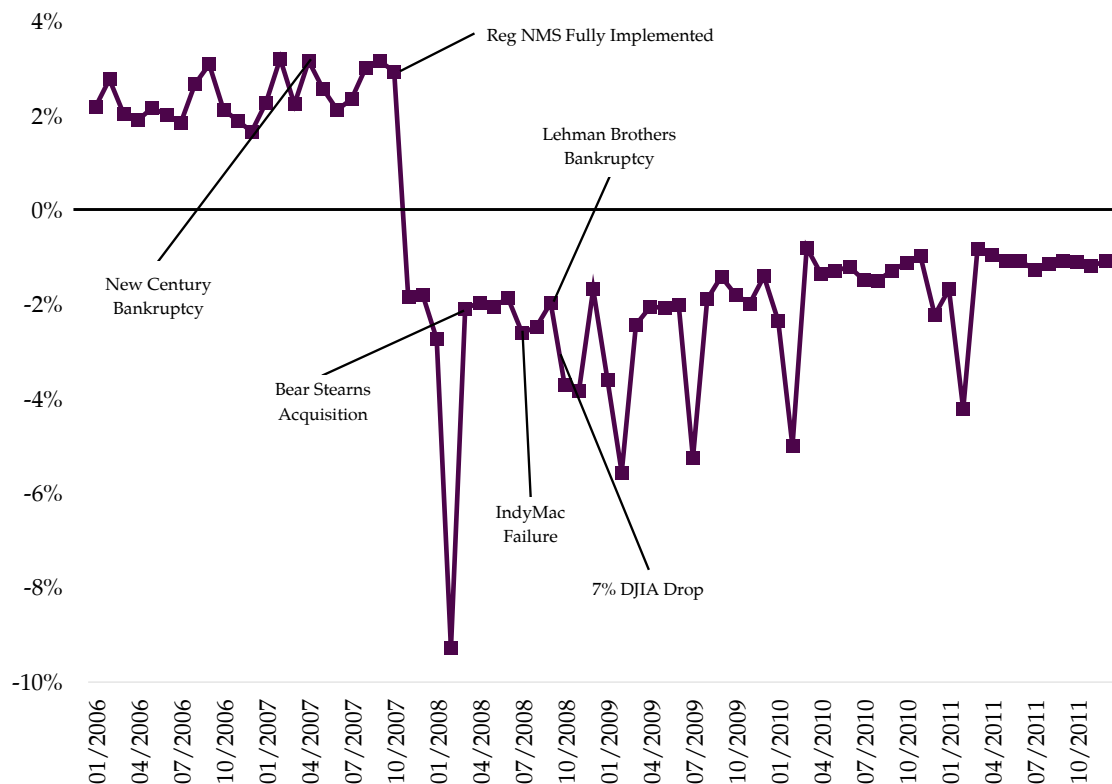
Figure 1.10: Impact on Employment of Other Broker-Dealer Workers



Employment patterns across several industries show no association between pre-existing IT infrastructure and the Great Recession

Notes: The figure on the left plots placebo-style coefficients across 180 sectors from regressions of county-level (log) employment in each sector on the treatment variable $(IT_c^{2004} \times \text{Post RegNMS}_t)$, where Post RegNMS_t is a dummy equal to one after 2006 and IT_c^{2004} is the PC per worker measure in the investment industry in 2004. The second figure runs similar regressions, except that I calculate an industry-specific IT intensity within each county for all $s = 1, \dots, 23$ broad sectors depicted, $IT_{c,s}^{2004}$. Only selected labels are plotted in the first figure. Estimates in purple are significant at 5%. County-level employment comes from the 2000 Census and annual American Community Surveys (ACS) annual samples after that. The estimation period is 2000-2015. Standard errors are clustered at the county-level.

Figure 1.11: Placebo Employment Regressions



Employment rates of stock traders working in October 2007 change on average 2% month-over-month prior to the full implementation of Reg NMS and 6% after October. Changes in employment are seasonal and relatively stable around key events of the Great Recession. The largest short-term decrease in employment follows the implementation of the regulation

Notes: The figure shows monthly changes of stock trader employment in the US. The sample tracks the employment status of stock traders working in October 2007, when Reg NMS became fully active. Data come from FINRA's Central Central Registration Depository (CRD).

Figure 1.12: Stock Trader Employment Rates During Key Events of the Financial Crisis

Table 1.1: Where do Stock Traders Work?

County	Major City	Stock Trader Employment	% Overall Employment	IT intensity rank (Out of 287)
<i>Panel A. 2004</i>				
New York, NY	New York City	55,276	24%	3
Atlantic, NJ	Atlantic City	17,056	7%	117
Hennepin, MN	Minneapolis	11,671	5%	44
Hudson, NJ	Jersey City	9,385	4%	30
Los Angeles, CA	Los Angeles	7,444	3%	145
Cook, IL	Chicago	7,154	3%	52
Suffolk, MA	Boston	6,527	3%	5
Providence, RI	Providence	5,471	2%	40
San Mateo, CA	San Mateo	4,872	2%	56
Fulton, GA	Atlanta	4,532	2%	4
<i>Panel B. 2009</i>				
New York, NY	New York City	44,636	20%	3
Atlantic, NJ	Atlantic City	15,691	7%	117
Hudson, NJ	Jersey City	10,177	4%	30
Dauphin, PA	Harrisburg	10,020	4%	80
Atlantic, NJ	Atlantic City	8,728	4%	117
Hennepin, MN	Minneapolis	7,472	3%	44
Los Angeles, CA	Los Angeles	7,247	3%	145
Lake, IL	Waukegan	6,614	3%	230
Hampden, MA	Springfield	5,286	2%	20
Suffolk, MA	Boston	4,914	2%	5
Cook, IL	Chicago	4,670	2%	52

Notes: This table shows the largest employment centers of stock traders in the US and the pre-existing IT infrastructure in those areas. The IT intensity rank measures the average number of computer per office worker in firms across the entire investment industry in 2004. Stock traders are individuals currently registered and employed to trade stocks professionally, holding a valid Series 7 license without licenses to work in other market segments.

Table 1.2: Largest Employers in the Investment Industry

	Stock Traders		All Investment Workers	
	2004	2015	2004	2015
EDWARD D. JONES & CO., L.P.	8,952	2,592	10,831	18,835
SALOMON SMITH BARNEY, INC.	7,733	3,290	22,184	8,127
MERRILL LYNCH, PIERCE, FENNER & SMITH, INC.	9,029	5,183	34,211	35,045
AMERICAN EXPRESS FINANCIAL ADVISORS	5,209	3,591	14,946	14,854
IDS LIFE INSURANCE COMPANY	4,981	*	14,439	*
A. G. EDWARDS & SONS, INC.	4,170	**	8,823	**
GOLDMAN SACHS & CO.	3,318	3,720	5,810	8,859
WELLS FARGO ADVISORS, LLC	3,225	3,881	14,654	28,688
CHARLES SCHWAB & CO., INC.	2,865	2,168	6,744	8,263
LEHMAN BROTHERS INC.	2,790	***	4,107	***
UBS FINANCIAL SERVICES INC.	2,732	1,659	11,305	13,250
FIDELITY BROKERAGE SERVICES, INC.	2,629	4,698	5,580	15,242
CREDIT SUISSE FIRST BOSTON LLC	2,244	1,771	3,768	4,421
MORGAN STANLEY & CO., INC.	4,740	2,102	17,257	4,614
J.P. MORGAN SECURITIES LLC	4,917	5,986	7,733	31,246
DEUTSCHE BANK SECURITIES INC.	1,863	1,703	3,113	4,287
UBS SECURITIES LLC	1,749	991	2,548	2,211
RBC CAPITAL MARKETS, LLC	1,374	1,578	3,505	5,732
LPL FINANCIAL LLC	1,323	2,400	6,633	20,083
INVESTMENT MANAGEMENT & RESEARCH	1,301	872	4,714	5,706
MSI FINANCIAL SERVICES, INC.	1,078	971	9,103	8,441
PRINCOR FINANCIAL SERVICES CORPORATION	1,053	433	4,180	4,104
NORTHWESTERN MUTUAL INVESTMENT SERVICES, LLC	999	1,058	7,075	10,394
H.D. VEST INVESTMENT SECURITIES, INC.	985	860	5,091	5,108
WACHOVIA CAPITAL MARKETS, LLC	963	1,610	1,455	3,400

Notes: This table shows computations of two groups of professionals at the largest broker-dealer employers. The first group, stock traders, include individuals licensed to trade stocks professionally (holding a Series 7 exam and valid license) who do not hold licenses to act in different market segments or occupations in the investment industry. The second group counts all individuals registered to work in the investment industry directly related to the sale or research on equities products or related activities related to investing in broker-dealers. *IDS Life Insurance merged and was renamed RiverSource Life in 2008. **AG Edwards & Sons was acquired by Wells Fargo Advisors in 2008. ***Lehman Brothers was liquidated in 2008 and its US trading division purchased by Barclays PLC also in 2008. Merrill Lynch's employment also counts Banc of America Investment Services and Banc of America Securities which were independent brokers prior to 2009. JP Morgan, Goldman Sachs, Morgan Stanley, and other entities with multiple subsidiaries have employment figures combined and displayed under one firm in the table. No employee counts include office or any personnel not directly involved with the activities described. These were the largest broker-dealer employers of stock traders in 2004.

Table 1.3: Availability of New Technologies by Sector

	Trading Automation	Fin	Non-Fin
	(1)	(2)	(3)
Post RegNMS $\times IT_c^{2004}$	0.137*** (0.034)	0.097 (0.061)	0.016 (0.106)
Year FE	X	X	X
County FE	X	X	X
# Counties	107	227	323

Notes: Column (1) estimates difference-in-differences estimates of trading technology availability growth following Reg NMS in counties with different pre-existing IT infrastructure. To benchmark the estimated effect, columns (2) and (3) run the same regression with overall finance patents and non-finance sector patents in the same as the outcomes. Standard errors are clustered at the county level.

Table 1.4: Did Automation Increase Industry Trading Profits?

	Aggregate Profits		% Profits Top Brokers
	No Missing Filings	All	
	(1)	(2)	(3)
Post RegNMS $\times IT_c^{2004}$	1.009** (0.439)	1.445** (0.716)	0.044** (0.021)
$t = 2002\text{--}2014$	<i>X</i>	<i>X</i>	<i>X</i>
Year FE	<i>X</i>		<i>X</i>
County FE	<i>X</i>	<i>X</i>	<i>X</i>
Observations	2,859	3,510	6,940
# Counties	288	340	288

Notes: Column (1) estimates a difference-in-differences model with county level aggregate net income from broker-dealers. This sample only includes firms with complete available profit information for all years the firm was operating (it allows for firm entry and exit). Column (2) uses all reporting firms, considering brokers that occasionally appear missing because their net income information was not available. Column (3) has as outcome the share of total local profits at the top broker-dealers (above the median total assets in the county). Standard errors are clustered at the county level.

Table 1.5: Did Automation Increase Firm Trading Profits?

	Broker-Dealer Profit							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post RegNMS $\times IT_c^{2004}$	0.098*** (0.029)	0.100*** (0.029)	0.053* (0.028)	0.058** (0.029)				
Post RegNMS $\times IT_c^{2004}$ $\times$ Total Assets _{ct}					0.066*** (0.014)	0.066*** (0.014)	0.011*** (0.002)	0.011*** (0.002)
Year FE		X		X		X		X
Broker-dealer FE			X	X			X	X
$t = 2002\text{--}2014$	X	X	X	X	X	X	X	X
County FE	X	X	X	X	X	X	X	X
Observations	25,862	23,510	25,000	25,000	25,862	23,510	25,000	25,000
# Counties	340	340	333	333	340	340	333	333

Notes: Column (1) regresses firm-level net income on the Reg NMS indicator interacted with the pre-policy local IT level. Column (2) adds year fixed effects to the model. Columns (3) and (4) include broker-dealer fixed effects. Columns (5) through (8) implement the same sequence of difference-in-differences regressions, interacting the variable of interest with the firm's total assets. Using lagged values of total assets or fixed pre-policy shifters changes the magnitude of estimated coefficients, but not the significance. Standard errors are clustered at the county level.

Table 1.6: Impact of Automation on Broker-Dealer Entry & Survival

# Traders at Firm	5-year Survival Rate						Broker-Dealer Entry
	All Firms	≤ 50	50-100	100-250	250-1000	1000+	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Post RegNMS $\times IT_c^{2004}$	0.016*** (0.004)	0.019*** (0.004)	0.01 (0.012)	0.018 (0.014)	0.034** (0.014)	0.081*** (0.025)	0.248** (0.119)
Average # Traders	66	8	71	157	454	3,219	
# Counties	311	300	77	64	65	31	
# Firms	3,688	3,202	186	146	106	48	3,452

Notes: This table compares, from columns (1) through (7), how firm survival rates of different-sized broker dealers changed as a consequence to trading automation. All regressions are of the form $y_{fct} = \alpha_0 + \delta (IT_c^{2004} \times \text{Post RegNMS}_t) + \gamma X_{fct} + \lambda_t + \mu_c + \varepsilon_{fct}$, where the outcome is a dummy for whether in a given point in time, the firm is still in business in the next 5 years. $\text{Post RegNMS}_t \equiv \mathbb{1}\{t \geq 2007\}$ is an indicator for the post-Reg NMS period, interacted with the pre-policy IT exposure variable, IT_c^{2004} . Column (1) runs the specification on all broker dealers and counties. Columns (2)-(6) gradually subsample firms based on employment level, from the smallest to the largest brokers. Column (7) has the outcome the probability that a firm the investment industry in a given year. Standard errors are clustered at the county level.

Table 1.7: Employment & Wage of Securities Brokers

	$Employed_{ict}$	$Employed_{ict}$	$Employed_{ict}$	$\ln(Wage_{ict})$
	(1)	(2)	(3)	(4)
Post RegNMS $\times \widetilde{IT_c^{2004}}$	-0.322*** (0.014)			
Linear trend $\times IT_c^{2004}$		-0.210*** (0.011)		
Post RegNMS $\times IT_c^{2004}$			-0.401*** (0.207)	
$IT_c^{2004} \times \text{Post RegNMS}_t \times \text{Skilled}_i$				0.091** (0.034)
Year fixed effects	X		X	X
County fixed effects	X	X	X	X
Linear trend		X		
$t \times \mu_c$			X	

Notes: Column (1) uses a high-speed internet broadband availability measure as the pre-policy exposure variable. (2) Interacts the IT intensity variable with a linear trend. (3) replicates the baseline model and interacts county fixed effects with a linear trend. (4) runs a triple-DiD specification of wages with differential effects for high and low skilled traders. Standard errors are clustered at the county level.

Table 1.8: Placebo and Robustness Tests

	$Employed_{ict}$	$Employed_{ict}$	$Employed_{ict}$	$\Delta \ln \left(Emp_c^{2004-2006} \right)$	$\Delta \ln \left(Emp_c^{2004-2009} \right)$
	(1)	(2)	(3)	(4)	(5)
$\mathbb{1} \{t \geq 2004\} \times IT_c^{2004}$	-0.022 (0.025)	-0.031 (0.025)	-0.041 (0.033)		
ΔIT_c^{2004}				-0.0003 (0.0006)	-0.001 (0.001)
$t = 2001-2004$	X				
$t = 2001-2005$		X			
$t = 2001-2006$			X		
Year FE	X	X	X		
County FE	X	X	X		

Notes: Columns (1)-(3) replicate the baseline model for stock trader employment with a placebo policy intervention and by gradually changing the estimating sample in the pre-Reg NMS years. Columns (4) and (5) test whether the growth rate in local IT infrastructure prior to the 2004 level I use to assign treatment intensity is predictive of future local area stock trader employment. Standard errors are clustered at the county level.

Table 1.9: Great Recession & Online Brokerage Services

	$Employed_{ict}^{otherfin}$	$Employed_{ict}$			
	(1)	(2)	(3)	(4)	(5)
Post RegNMS $\times IT_c^{2004}$	0.371*** (0.016)	-0.202*** (0.011)	-0.038*** (0.017)	-0.209*** (0.015)	-0.317*** (0.015)
$\Delta \ln(\text{House Prices}_{ct})$				-0.046 (0.307)	
E*Trade _{ct}					0.002 (0.004)
Omit Big 5		X			
Omit New York			X		
Year fixed effects	X	X	X	X	X
County fixed effects	X	X	X	X	X
Full set of controls	X	X	X	X	X
Individual-year observations	11,904	18,496	1,617,318	22,987	9,304

Notes: This table shows estimates of different potential margins the Great Recession and online brokerage services may drive the baseline results for employment. Column (1) compares stock traders to all other finance workers (including outside the investment industry) in the same area. Column (2) drops the Big 5 (Goldman Sachs, Merrill Lynch, Morgan Stanley, JP Morgan, and Lehman Brothers) investment banks. Column (3) excludes workers and firms in New York from the sample. Column (4) controls for local changes in home prices. Column (5) adds as control imputed local number of E*Trade accounts.

Chapter 2

The Bias of Simple Bid-Ask Spread Estimators

2.1 Introduction

Following Roll (1984), the use of bid-ask spread estimators has become widely popular in empirical finance. Often times, such estimators are the simplest, or possibly the only way, to measure liquidity costs in emerging and voice-based markets (Li et al. (2018)), during historic periods (Patton and Weller (2020), Bernstein et al. (2019)), or when researchers predominantly use low-frequency data (e.g., McLean and Pontiff (2016)), as is common in the asset pricing literature. Despite the usefulness and popularity of low-frequency spread estimators, results of performance evaluation in different markets reveal puzzling results.

Consider for instance one of the best-known spread measures, the high-low estimator from Corwin and Schultz (2012). While it correlates well with effective spread levels in US stocks (Abdi and Rinaldo (2017)) and bonds (Chakravarty and Sarkar (2003)), it only marginally captures spread changes in FX markets (Karnaukh et al. (2015)) and ETFs (Marshall et al. (2018)), and has no correlation with commodity trading costs (Marshall et al. (2011)). Even when the measure remains superior to alternatives in global stock markets (Fong et al. (2017)), its proxy quality varies substantially across countries. Other simple spread estimators show the same variability in performance.

A second set of issues arises regardless of the application context. Estimators often yield spreads that are negative or implausible. Researchers are required to implement a number of *ad hoc* steps or arbitrary imputations, such as replacing negative spreads with zeros. Discrepancies in performance and problematic estimates are usually attributed to differences in microstructure across markets, which at first seems a reasonable explanation. Model assumptions may hold in certain settings and be violated in others. However, the same estimation shortcomings persist even in simulated trading data with ideal characteristics or in very large samples.¹

¹An additional hypothesis that may account for performance issues is the changing nature of liquidity provision in modern financial markets. For example, Barardehi et al. (2018) argue that temporal dependence of orders submitted by low-latency traders make calendar-fixed liquidity proxies fail to capture relevant trading costs in algorithmic markets. Even traditional bid-ask spread estimators may capture broader liquidity dimensions than just the spread, including price pressure. Despite that, Easley et al. (2020) find that traditional microstructure measures, including price-based liquidity proxies, remain useful in a wide number of markets.

What explains differences in performance and long-standing estimation issues of bid-ask spread measures and can these issues be addressed empirically? This paper derives a theoretical framework to determine the causes and implications of misbehavior in spread estimators, how to empirically measure estimation bias, and how to increase the degree of confidence in out-of-sample use of spread benchmarks. We focus our analysis on one of the most widely used spread measures, the high-low estimator of Corwin and Schultz (2012) and show how our conclusions apply to other spread estimators based on daily price data.

We begin our analysis by revisiting the empirical setting used in the literature to evaluate the performance of bid-ask spread estimators. These proxies are tested in standard US stock data, where benchmark effective spreads are calculated from high-frequency trade and quote data and compared to estimated spreads. Performance is generally measured as how well an estimator correlates *on average* with effective spreads in the cross-section and time-series. Considerable skewness in the distribution of liquidity in US stocks, however, plays a crucial role in inflating the measured performance of spread estimators.

We show that cross-sectional correlations between the high-low estimator and effective spreads in US stocks are heavily driven by a few very illiquid stocks whose removal from the cross-section decreases the average monthly correlation by 25% (from 0.62 to 0.48). These patterns are robust even in annual spread averages, in spread levels or changes, with or without *ad hoc* adjustments, and in other estimators such as Roll (1984) and Abdi and Rinaldo (2017).²

While some previous results in papers deriving spread estimators have suggested the existence of some relationship between the underlying effective spread size and proxy performance, the extent to which reported performance depends on the shape of the underlying distribution of trading costs in stocks is surprising. This persistent relationship motivates us to develop a framework that focuses on common bias sources — irrespective of application context — as opposed to idiosyncratic market features difficult to pinpoint and unlikely to account for puzzling performance results across several distinct settings.

We first derive a closed-form solution for the estimation bias in the high-low spread estimator, which comprises two terms: a moment bias and a small sample bias. Spread proxies estimate the bid-ask spread from a combination of daily high, low, and close transaction prices, which contain microstructure noise. To separate the spread from latent price information and achieve a closed-form solution, estimators are generally derived under the validity of asymptotic relationships, resulting in nonlinear functional forms of population moments. Additional assumptions usually impose restrictions that are plausible only in short data intervals (e.g., stationarity or constant spreads), so that in practice population moments are replaced with sample analogues calculated with a small number of data points.

²This disproportionate impact of low liquidity stocks in driving cross-sectional results is similar to the well-known effect of microcaps (more likely to be illiquid) in validating return anomalies (Hou et al. (2018)).

Especially important, implementation of the high-low estimator requires replacing moments of the range with transaction-based price ranges over a two-day period. Crucially, our framework shows how this introduces a moment bias in the estimator due to its nonlinear functional form. We further establish that this bias is equivalent to the well-known Jensen’s inequality bias in the Roll measure (Roll (1984), Hasbrouck (2004)), which is also biased because of the use of the sample estimator of the autocovariance. The size of this moment bias in the high-low estimator can be expressed as a function of observable data and thus retrieved empirically. Nonetheless, even with a bias correction, the estimator still deviates from the underlying spread due to a small sample bias, which stems from an empirical convergence failure assumed in the estimator’s derivation.

The high-low spread proxy relies on the validity of the asymptotic distribution of the range, which is rarely observed in practice or even in large samples of simulated ideal prices. We show that, as a consequence, high-low estimates remain biased in sample sizes typical for financial data. This bias source is equivalent to the small sample bias in the Roll measure. In that case, the sample covariance estimator is biased in small samples (Harris (1990)), which introduces further bias in the proxy. These two common sources combine to drive a bias in the high-low measure of same nature as in the Roll measure, which relies on different transaction variables for identification and has a distinct set of assumptions.

Next, we explain the determinants of the two sources of bias. By writing the total bias as a sum of both moment and small sample terms, we prove that the high-low estimation bias is decreasing in the spread and increasing in volatility. These two latent variables in the data generating process affect the estimation bias in the high-low measure the same way they do with the Roll estimator, whose performance also worsens as the signal-to-noise ratio falls. This theoretical association rationalizes the previous patterns found between underlying effective spread sizes and estimation bias in the US stock sample. To eliminate potential idiosyncratic factors, we simulate ideal trading data that cleanly confirms the same relationship.

The fact that estimation bias depends on the underlying bid-ask spread magnitude has important implications for empirical applications using spread estimators to proxy for liquidity costs. For instance, in low-frequency cross-sectional regressions where researchers include spread estimates as independent variables, coefficient estimates will suffer from non-classical measurement error. In cases where the actual measured spread value matters, for example, to sort a portfolio on liquidity or to exploit changes in liquidity costs over time, researchers need to know how much trust to place on a spread estimate of 1% provided by a low-frequency estimator. Is this a good proxy of an underlying spread of 1% or is the actual effective spread 0.2% or 2%?

To equip researchers using the high-low estimator to address these issues, we develop an approach to obtain empirical bounds on the estimation bias. Exploiting the fact that the moment bias is always non-

negative and can be computed from daily data, we show how to sign the small sample bias, allowing us to bound the effective spread using only the high-low estimate and moment bias. These bounds are sufficiently tight so that they are informative in a variety of settings — to both quantify the bias and control for non-classical measurement error in linear regressions, in the cross-section and time-series. The bounds on the underlying spread can also be used to sort portfolios based on liquidity much more precisely than simply relying on the high-low measure. Moreover, the bias bounds are derived under the same assumptions as the high-low estimator and despite of potential idiosyncratic biases remaining, are empirically-consistent in the vast majority of cases, as we show in three different applications using US stock data.

It is important to emphasize that the causes and implications of misbehavior we discuss are not exclusive to the high-low measure. We focus on the proxy because of its popularity, perceived dominance relative to other estimators, and innovative identification strategy of the spread. But the lessons we draw about its shortcomings are more general. The same factors resulting in poor performance of the measure will also negatively affect other spread estimators, such as the Roll measure and the close-high-low estimator from Abdi and Ranaldo (2017), as we show in the Online Appendix C. Indeed, even with several modifications and overall superior performance compared to the Roll measure, the high-low estimator suffers from the same sources of misbehavior as the seminal estimator.

The main contribution we make is to establish the importance of structural factors common to simple bid-ask spread estimators in affecting performance and how to empirically quantify estimation bias in the high-low measure. Our estimation bias bounds should prove useful for researchers in market microstructure, asset pricing, or that require direct low-frequency measures of trading liquidity costs.

Recent work by Lou and Shu (2017) and Jahan-Parvar and Zikes (2019) challenges the pricing of liquidity measures in certain market circumstances. Our paper takes a more fundamental approach by analyzing the structure and source of bias in spread proxies, which informs researchers whether they are likely to face implementation issues in their own setting. Rather than suggesting the abandonment of simple spread estimators altogether, we recognize that these are crucial tools to measure liquidity in many circumstances, and provide researchers with tools to increase the credibility of their results.

Our results demonstrate that even though both volatility and the underlying spread size affect the estimator’s behavior, spread size is more important in generating and accounting for estimation bias. We show that volatility only begins to affect estimation bias when its magnitude is at least four times that of the underlying spread. Although this ratio is likely met in modern equity markets, which have a median effective spread of 0.3%, centering the analytical response of the bias to the spread has a crucial advantage. By exploring the relationship between spread size (the quantity the researcher aims to estimate) and estimator bias, we are able to jointly assess the empirical size of bias and the underlying magnitude of the effective

spread.

Our initial set of findings microfound misbehavior in the popular high-low spread measure in the same fashion as the work in Harris (1990) for the Roll estimator. By focusing on common factors inducing estimation bias and negative or indeterminate estimates and how these factors react to changes in volatility and, in particular, the underlying true spread, we are able to establish a direct link between misbehavior in the Roll and high-low measures, and apply our framework to another simple bid-ask spread estimator, the close-high-low proxy. This contrasts our work from previous studies that relied on idiosyncratic factors, like Bleaney and Li (2015) and Nieto (2018), or in mechanical noise from truncating averages with zeros (Jahan-Parvar and Zikes (2019)).³

We argue that both the moment and small sample bias sources account for almost the entirety of empirical misbehavior in the high-low measure. To test that, we derive a counter-factual version of the high-low measure which completely corrects for all common bias sources. This corrected estimator version performs extremely well, correlating with the effective spread by almost 95% in the cross-section and 80% in the time-series. A back-of-the-envelope calculation implies that, in the cross-section, common factors contribute to a loss of over 30 percentage points in performance, while all other imperfections combined imply a loss of performance of only 5 percentage points.

We also draw several recommendations from our analysis for the development of new estimators and performance evaluation. To more credibly establish the quality and recommend a certain spread proxy, performance evaluation studies using tick data should include more detailed subsampling analysis, particularly by analyzing the relationship of correlation coefficients in the cross-section across different effective spread groups. Moreover, we also suggest in sensitivity analyses that performance is measured after dropping the most illiquid units from the sample as well as in random subsamples from the data when the cross-section is sufficiently large. Coefficient stability across these different tests indicate a greater degree of out-of-sample confidence.⁴

Going forward, newly developed spread estimators using transaction data that share the structure of the Roll framework should be subjected to rigorously designed simulation environments with true spreads in the data generating process well below 0.5%, the usual minimum spread used in the literature, and verify the role of price volatility and trading costs in potential misbehavior. More preferably, a formal analysis of the existence of the moment or small sample bias in new measures is also suggested, as these two sources can severely limit performance of simple bid-ask spread estimators. Finally, even if after rigorous testing

³We refer to these bid-ask spread estimators as “simple” following the designation from Roll (1984), Corwin and Schultz (2012), Abdi and Ranaldo (2017).

⁴In addition to a more robust and extensive empirical performance evaluation, the choice of the underlying spread benchmark is important, as effective bid-ask spreads might be inappropriate (Hagströmer (2021)).

new estimators perform better than existing proxies, they should still accompany practical tools to quantify empirical bias and reduce its importance in applications, as we do for the high-low measure.

This paper is organized as follows. Section 2 introduces the structure of simple bid-ask spread estimators, underscoring deriving assumptions and potential idiosyncratic factors important for empirical performance. Section 3 revisits the standard performance evaluation setting of spread estimators, analyzing the robustness of previous findings. Section 4 introduces our framework, where we study the determinants and implications of estimation bias in the high-low estimator. Section 5 implements our diagnostic test and shows how to empirically assess estimation bias and spread size. Section 6 concludes.

2.2 The Structure of Simple Spread Estimators

This section introduces the framework common to most simple bid-ask spread estimators, focusing on the derivations and assumptions of the high-low measure by Corwin and Schultz (2012). The building blocks and identification approach of the framework were introduced by Roll (1984), especially basing the estimation of the spread on daily transaction prices. Many estimators introduced later followed or extended the Roll measure (French and Roll (1986), Thompson and Waller (1987), Lesmond et al. (1999), Hasbrouck (2004), and Fong et al. (2017)). We underscore the two sources of bias in the Roll measure previously shown in the literature. Later in the paper, we will show that the same bias sources also affect the high-low proxy, and more generally, how simple bid-ask spread estimators employing different technologies have performance issues based on the same biases.

2.2.1 Setting

The derivation of simple bid-ask spread estimators begins with the usual assumption that log prices follow a geometric Brownian motion (GBM). Hence, continuous existence of prices during trading hours is implicitly assumed, though continuous observation thereof is not necessary. The relationship between observed (p_t) and true prices (P_t) is given by:

$$p_t = P_t + \frac{S}{2}Q_t \quad (2.1)$$

where S is the total effective spread (observed) and Q_t is the order flow indicator ($Q_t = 1$ for buyer-initiated orders and $Q_t = -1$ for seller-initiated orders). Given that the latent price evolves as a random walk, $\Delta P_t = \varepsilon_t$, with $\Delta P_t = P_t - P_{t-1}$, $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, trade direction is assumed to be independent of efficient prices. We refer to the true spread or the effective spread calculated from trading data interchangeably by

S .⁵

In spite of its simplicity, (2.1) represents the cornerstone expression shared by all spread proxies we characterize as *simple estimators*. Within this category, time is indexed by $n(t)$, $t = 1, \dots, T$, days, so that only end-of-day low frequency data is necessary to recover empirical estimates of the true spread S . Daily spread estimates may be averaged at the desired frequency, say months or years, in order to compute average spread estimates.

2.2.2 The Roll Estimator

Roll (1984) considers daily closing prices, $c_t = C_t + Q_t S/2$, as the relevant price realization for empirical use. By assuming that the distribution of price changes in (2.1) is plausibly stationary *in short time periods*, he arrives at the well-known Roll estimator formula:

$$S^{Roll} = 2\sqrt{-\text{Cov}(\Delta p_t, \Delta p_{t+1})} \quad (2.2)$$

The two bias sources in the Roll estimator

In practice, (2.2) is implemented by replacing the population autocovariance of price changes with its sample version, effectively leading to the estimator:

$$\hat{S}^{Roll} = 2\sqrt{-\widehat{\text{Cov}}(\Delta p_t, \Delta p_{t+1})}$$

While this replacement is seemingly innocuous, the Roll measure suffers from a Jensen's inequality problem, due to its nonlinear functional form. Average spread estimates are downward-biased with (absolute) bias size decreasing in the true spread level S , increasing in price volatility σ^2 , and increasing in the sample size $n(t)$: $\mathbb{E}[\hat{S}^{Roll}] - S = -S \left(\frac{\mu_4/\sigma^4 + 4}{8(n(t) - 1)} \right)$.

To complicate things further, Harris (1990) shows that another source of bias affects the behavior of the Roll measure. The sample covariance statistic used to calculate $\hat{S}^{Roll}$ is biased in small samples, so that even if the Roll measure were a linear function of the sample covariance, so that the Jensen's inequality bias would vanish, the estimator will remain biased in small samples. Unfortunately, "small sample" in this context covers the temporal dimension of financial data usually available, making the issue pervasive. Considering both bias sources in total, the Roll measure bias decreases in the spread and increases in volatility (overestimating S in liquid markets). These two bias sources are well-known, but strongly associated with

⁵The most common cost-based high-frequency liquidity benchmarks used empirically are the effective and quoted spreads. Effective spreads measure "immediate" execution, i.e., faced traders executing at observed prices.

the functional form in (2.2). We refer to them as a *moment bias* and a *small sample bias*. Finally, a separate issue with the Roll estimator is the presence of indeterminate daily spread estimates due to frequent positive price autocovariance.⁶

The structural link between spread size, bias, and frequency of indeterminate spreads in the Roll estimator has been relatively unexplored and often overlooked in the context of other spread estimators. Newer proxies have been shown to more accurately capture effective spreads in simulated and empirical environments relative to the Roll measure, with little attention being paid to estimation bias or other systematic issues.

2.2.3 The High-Low Estimator

Corwin and Schultz (2012) exploit the empirical regularity that the sign of the trade indicator Q_t is defined as positive in daily high prices and negative in daily low prices. Thus, under the assumption that high and low prices are buyer- and seller-initiated, respectively, (2.1) translates as the pair:

$$(h_t, l_t) = \left(H_t + \frac{S}{2}, L_t - \frac{S}{2} \right) \quad (2.3)$$

and more specifically, the daily log range $r_t \equiv h_t - l_t$ can be manipulated by using the moments derived in Parkinson (1980) and Garman and Klass (1980) for the true range R :

$$\mathbb{E} \left[T^{-1} \sum_{t=1}^T R_t \right] = k_1 \sigma \quad \text{and} \quad \mathbb{E} \left[T^{-1} \sum_{t=1}^T R_t^2 \right] = k_2 \sigma^2 \quad (2.4)$$

where $k_1 \equiv \sqrt{8/\pi}$ and $k_2 \equiv 4 \ln 2$, and σ^2 is an unbiased (under the assumption of no drift in the GBM) estimator of daily variance based on the range when $\mathbb{E} \left[T^{-1} \sum_{t=1}^T R_t \right]$ and $\mathbb{E} \left[T^{-1} \sum_{t=1}^T R_t^2 \right]$ are replaced with R_t and R_t^2 , respectively.

The appealing argument that using additional information contained in the daily range improves volatility estimation underscores the innovative path taken by Corwin and Schultz (2012). Since the range represents the boundaries of observed daily price oscillation, i.e., the volatility measured in σ^2 , it also includes the effective spread at those boundary points according to (2.3). The additional assumption that the spread is constant over every pair of consecutive days, $\bigcup_{k=1}^{n(t)-1} [t_k, t_{k+1}]$, while volatility remains proportional to time,

⁶Negative daily serial covariance is necessary to compute (2.1) or its implementable version. Harris (1990) argues that small sample bias alone explains the prevalence of positive covariance estimates. Nearly half of daily spread estimates turn out negative, requiring researchers to discard substantial information. Importantly, because the small sample bias in the sample covariance decreases in the true spread size, more liquid markets frequently induce positive autocovariance, ultimately generating indeterminate spread estimates more often.

allows for uniquely determining S from the following expressions:

$$\mathbb{E} \left[\sum_{t=1}^2 r_t^2 \right] = (8 \ln 2) \sigma^2 + \left(8 \sqrt{\frac{2}{\pi}} \right) \sigma \hat{S} + 2 \hat{S}^2 \quad (2.5)$$

$$\mathbb{E} \left[r_t^{*2} \right] = (8 \ln 2) \sigma^2 + \left(8 \sqrt{\frac{1}{\pi}} \right) \sigma \hat{S} + \hat{S}^2 \quad (2.6)$$

which relate the daily range (r_t) and the two-day range (r_t^*), where $r_t^* \equiv \max\{h_t, h_{t+1}\} - \min\{l_t, l_{t+1}\}$, to both the spread and daily volatility. The two-day range captures the overall volatility in each pair of days, and if time proportionality holds, it should correspond to twice the volatility of a single day.

After solving for σ^2 in (2.5) and (2.6), the high-low spread estimator is obtained as:

$$S^{HL} = 2 (L(\alpha) - 1) \quad (2.7)$$

for

$$\alpha \equiv \frac{\sqrt{2\beta} - \sqrt{\beta}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}}, \quad \beta \equiv \sum_{t=1}^2 \mathbb{E} [r_t^2], \quad \gamma \equiv r_t^{*2} \quad (2.8)$$

where $L(\cdot)$ is the logistic function. We transform the functional forms for S and α from Corwin and Schultz (2012) for ease of exposition using the steps in Appendix (B.2).

Specific model assumptions

Before moving on to the re-evaluation of the empirical performance of the high-low measure, we underscore important identifying assumptions of the estimator. First, daily high prices are buyer-initiated and low prices are seller-initiated. Second, the price process is continuous and that volatility is proportional to the trading horizon. Third, the asymptotic distribution of the true range holds. Fourth, the latent spread is constant over the trading horizon.

Abdi and Ranaldo (2017) check that the first assumption is verified in 95% of US stock-days, therefore it is a stylized fact in practice. The second assumption has devoted most of the attention from the analysis in Corwin and Schultz (2012) on model assumption failures, particularly time-varying volatility. We show later that, combined with time-varying spreads, these assumptions play a small role in affecting the empirical performance of the estimator when compared to the bias sources we highlight.

2.3 Reevaluating Simple Bid-Ask Spread Estimators

This section revisits the standard evaluation environment of bid-ask spread estimators by investigating the robustness of previous findings in the cross-section and time-series of US stocks. Our analysis will show that a skewness effect in US stocks — many stocks with small spreads and few that are extremely illiquid — produces cross-sectional correlations with a low degree of out-of-sample confidence. Removing few large effective spread stocks from the sample used in horse-race exercises decreases the measured performance of the high-low, close-high-low, and Roll estimators by almost 25%. This lack of robustness associated with the underlying level of effective spreads will motivate our theoretical framework developed later in the paper.

2.3.1 Benchmark Stock Data

Our baseline empirical setting uses the standard US stock sample employed in horse-race type papers like Goyenko et al. (2009). This is the sample used to benchmark the performance of simple bid-ask spread estimators (e.g., Corwin and Schultz (2012) and Abdi and Rinaldo (2017)). We use Daily TAQ (DTAQ) stock data from 2009 through 2013 with stocks listed at NYSE, AMEX, and NASDAQ, applying the Holden and Jacobsen (2014) procedure to compute daily effective spreads. We describe these steps in more detail in Appendix (B.1).⁷ Estimation of low-frequency proxies uses CRSP daily data, again closely tracking the setting in previous studies. We also apply the data cleaning steps in Corwin and Schultz (2012) for anomalous trading days, though the effects of carrying out these changes on the overall analysis are minimal.

Figure (2.1) plots the distribution of daily effective spreads across all US stocks. The median effective spread in the sample is 0.21% and more than 70% of stocks have effective spreads lower than 1%. The long right tail shows that only about 250 stocks have spreads higher than 3%, less than 10% of all stocks. This prominent skewness of effective spreads in US stocks, therefore, is a stylized fact in the data.

2.3.2 Reassessing the Performance of Simple Spread Estimators

Notation

The performance of low-frequency bid-ask spread estimators is usually evaluated by comparing monthly averages of effective spreads and the estimated proxy in the cross-section and time-series. Formally, the monthly average of high-low estimates is given by:

$$\hat{S}_n^{HL} \equiv \frac{\sum_{t(n)} \hat{S}_t^{HL} \cdot \mathbb{1}\{adjustment\}}{\sum t(n)} \quad (2.9)$$

⁷We thank Craig Holden and Stacey Jacobsen for making their code publicly available. We also thank Shane Corwin, Paul Schultz, Farshid Abdi, and Angelo Rinaldo for making their code publicly available for comparison.

where daily HL estimates are averaged over t trading days available in month n . We denote the high-low estimator used empirically as $\hat{S}^{HL}$.

Daily spread estimates may be discarded or imputed based on *ad hoc* adjustments used in the literature: i) a zero adjustment, which replaces negative daily estimates with zeros, ii) using only positive spreads, or iii) using all raw estimates without any correction. We refer to monthly average versions of the high-low measure as follows: *Raw* for using both negative and positive estimates (that is, no adjustment); *Zeros*, for the replacement of negative estimates with zero ($\mathbb{1}\{\hat{S}_t^{HL} < 0\} = 0$); and *Positives* when discarding negative or indeterminate estimates. We also calculate annual averages of spread estimates, which are then compared to annual effective spread averages, and replicate some of our analysis using *changes* in monthly and annual estimates. Computed sample averages of the Roll and close-high-low estimators follow the same approach in (2.9).

Analysis

With the previous cross-sectional effective spread patterns just illustrated, we now dissect the performance of simple bid-ask spread estimators. Figure (2.2) begins by calculating the average cross-sectional correlation between monthly high-low spread estimates and effective spreads. The HL measure performs relatively well, tracking the cross-section of over 3,300 stocks with an average correlation of 62%.⁸

Next, we break down the stock sample according to effective spread deciles and re-calculate the cross-sectional correlation coefficients within each liquidity level decile. Given the heterogeneity in effective spreads in US stocks, particularly the right skewness, this analysis verifies whether performance might be driven by a particular subset of stocks. We display results for each decile for three computed versions of high-low estimates: using the zero adjustment (which is the recommended *ad hoc* modification by Corwin and Schultz (2012), in blue in the figure), only positive daily estimates (in gray), and without any adjustments (raw, in brown). Although these other versions of the high-low measure are not recommended by Corwin and Schultz (2012), we include them in our analysis to show that the patterns we discuss next are not an artifact of the *ad hoc* zero adjustment.⁹

The distribution of performance across effective spread levels reveals important patterns masked by an evaluation of the full cross-section. First, the HL measure (zero adjustment or positive estimates) only

⁸The correlation point-estimate is identical to the value in Table 5 in Abdi and Ranaldo (2017), who use the same stock sample and similar period. Estimates for the Roll and close-high-low proxies are also identical.

⁹Note that imputing zeros or using only positive spread estimates introduces bias in sample averages of any parameter estimate. In the context of spread estimation, this mechanical (or sampling) bias from truncation or censoring trivially inflates the average estimated spread, especially relative to smaller true spreads and when the price variance is large. While the moment bias and the small sample bias are structural biases affecting *point* estimates of the Roll measure, zero imputation or other *ad hoc* empirical adjustments introduce an additional bias source. Within our framework in Section 2.4, the potential bias introduced from the zero adjustment is absorbed as part of an idiosyncratic bias.

performs modestly in the top effective spread decile, which has an average monthly spread of 4.3%, 20 times larger than the median effective spread for overall US stocks. Even in this most favorable case, both adjustments only yield average correlations below 50%. More importantly, dropping stocks with the largest underlying effective spread in the top decile yields a cross-sectional correlation across all remaining stocks (deciles 1 to 9) of 48%, 14 percentage points lower than the sample-wide performance. To put this magnitude of performance loss in perspective, 14 percentage points is larger than the measured improvement in performance from the close-high-low estimator over the high-low measure in Abdi and Ranaldo (2017).¹⁰

The same “skewness effect” impacts the Roll and close-high-low estimators, both for monthly and annual analyses (Online Appendix Figures (D.1) and (D.2)). The overall loss in performance from removing the top stock decile from the cross-section is of similar magnitude to the monthly high-low estimator, about 25%. Online Appendix Figures (D.3), (D.4), and (D.5) indicate that the cross-sectional correlation between changes of monthly and annual high-low, close-high-low, and Roll spread estimates and effective spreads are also inflated by a few illiquid stocks in the cross-section. We underscore this specific context as previous papers have found that the high-low measure may track well only changes in liquidity in certain markets, and not necessarily levels as in Figure (2.2).

Finally, a simple “placebo” test of the potential relationship between estimation performance and the underlying effective spread size is to drop the decile with the *most* liquid stocks. The idea here is to cleanly separate the effect of decreasing the sample size (in the cross-section) from dropping *certain* stocks from the sample (those few stocks with very large spreads). Across monthly and annual averages of levels and changes and for all estimators, removing the first stock decile leaves the performance in the remaining sample unchanged relative to the full cross-section.¹¹

Implications

The fact that the standard benchmark evaluation of simple bid-ask spread estimators overstates performance by almost 25% has important empirical and theoretical implications. On the empirical side, the widespread use of low-frequency spread estimators in markets outside US equities relies entirely on out-of-sample extrapolation — if an estimator “robustly” benchmarks effective spreads in US stocks, so it should in

¹⁰The same general pattern is observed in the performance of annual average spreads, also in Figure (2.2). Using longer time periods improves the relationship between the high-low measure and effective spreads in the cross-section, but a large contributor to performance is again the top decile of trading costs.

¹¹Apart from the inconsistent performance across different markets we previously mentioned and discuss in detail later, existing analysis suggests some degree of systematic variation in the high-low proxy quality even for US stock data. Abdi and Ranaldo (2017) show in their Table 5 a breakdown of correlation quality by effective spread quintiles, suggesting a much lower performance of considered spread proxies in more liquid stock groups. This holds for both the high-low estimator and their measure, the close-high-low spread proxy, as well as the Roll measure. A similar point is inferred from Table 6 in Corwin and Schultz (2012), where the cross-sectional correlation between the effective spread and the high-low measure decreases from 70% in low cap stocks to 18% in high market-capitalization firms, which are more liquid. While suggestive, there is no indication in previous papers that these subsample patterns are robust or even theoretically founded.

markets like futures, foreign exchange, bonds, and ETFs. While it is reasonable to expect that performance may not replicate due to idiosyncratic factors in each application context, the role of skewness in effective spreads in inflating performance raises concerns to applications in *any* market.¹²

On the theoretical side, the fact that three estimators using different technologies to recover the spread all have a much worse performance once few illiquid stocks are removed from the cross-section points to a common factor of misbehavior — possibly associated with the underlying latent spread. This conjecture motivates us to develop a framework focusing on structural forces determining the estimation bias of the high-low measure and how the true spread and volatility influence these biases.

2.4 Estimation Bias & Performance

This section provides a theoretical framework to evaluate the determinants of the structure and behavior of the high-low measure and its empirical consequences. We first show that replacing the population moment in the estimator’s parameter β with daily ranges introduces a moment bias, which makes the implementable version of the high-low estimator biased compared to its theoretical counterpart. This moment bias is structurally equivalent to the Jensen’s inequality bias in the Roll estimator.

More importantly, even if we remedy this moment bias, the estimator is still biased in practice because of its estimation horizon. The high-low measure relies on the asymptotic distribution of the range, which fails to hold for consecutive trading days. More generally, the relationship required for unbiased estimates fails to hold even over longer trading horizons, making the spread estimator biased irrespective of the moment bias. This small sample bias in the high-low measure corresponds to the small sample bias in the Roll estimator, where the sample covariance estimator introduces additional bias in the proxy. These two common biases almost completely accounts for performance loss in high-low estimator.

With a closed-form expression for the bias, we further show that the high-low bias increases in small spreads and high volatility, which explains our previous findings on poorer performance in more liquid stocks. We confirm those patterns in generated data that simulates an ideal setting for all other model assumptions. In the same simulated environment, we also find that the frequency of negative estimates increases when we shock the spread towards zero. Online Appendix C applies this framework to the close-high-low measure of Abdi and Ranaldo (2017), which yields by and large the same conclusions with respect to estimation bias size, latent spread and volatility, and negative estimates.

Finally, we conduct a counterfactual exercise to quantify the contribution of the two common bias sources — moment and small sample — to the overall performance of the high-low measure. This exercise

¹²Further analysis in the Online Appendix A demonstrates this point further by showing the high sensitivity of findings in horse-race type analysis in US stocks to changes in the sample size and other simple tests.

isolates the loss in performance attributed to the two common bias sources from *all* other idiosyncratic bias source, including US stock microstructure effects and model assumption failures (e.g., constant spreads). Counterfactually removing the two common biases improves empirical performance by over 30 percentage points (62% to 94%) in the cross-section and in the time-series (48% to 80%), jointly accounting for 90% of total performance loss in the high-low estimator.

2.4.1 Framework

Start by rewriting the high-low estimator formula in Equation (2.7) as

$$\tilde{S}^{HL} = \alpha$$

after linearizing the expression $2 \frac{\exp \alpha - 1}{\exp \alpha + 1}$ around $\alpha_o = 0$. This approximation generates negligible error for spread values between -100% and 100% . Going forward, we denote $\tilde{S}^{HL}$ as the theory-implied high-low estimator, which uses the moment-based parameter $\beta = \mathbb{E} \left[\sum_{j=0}^1 r_{t+j}^2 \right]$.

Next, the following equivalence allows us to solve for the parameter α in the difference of square roots involving β and γ :

$$\tilde{S}^{HL} = \underbrace{\frac{\sqrt{2\beta} - \sqrt{\beta}}{3 - 2\sqrt{2}}}_{\alpha} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}} \equiv (1 + \sqrt{2}) (\sqrt{\beta} - \sqrt{\gamma}) \quad (2.10)$$

We denote the expression above as the **theoretical high-low estimator**. The functional form in (2.10) provides a direct way to express how the high-low spread estimator is jointly — and fully — determined by the latent spread and price volatility

$$\tilde{S}^{HL} = S + \underbrace{(1 + \sqrt{2}) (\sqrt{2\mathbb{E}[R_t]} - R_t^*)}_{g(\sigma)} \quad (2.11)$$

The high-low estimator is therefore additively separable in (S, σ) , where the function $g(\sigma) \in \mathbb{R}$ drives the sign of the daily spread estimate. We return to the determinants of negative spreads in the Online Appendix B. Finally, though we show it formally below, it is clear from (2.11) that unless g is zero, the theoretical estimator is biased. In practice, $\tilde{S}^{HL}$ is estimated using a sample-based version of β , defined as $\hat{\beta} \equiv r_t^2 + r_{t+1}^2$. This replacement results in a different version of the functional form in (2.11). First, we

rewrite $\widehat{\beta}$ as¹³

$$\sqrt{\widehat{\beta}} = \sqrt{r_t^2 + r_{t+1}^2} = (R_t^{min} + S) \left(\sqrt{1 + \kappa^2} \right)$$

for $\kappa \equiv \frac{\max\{r_t, r_{t+1}\}}{\min\{r_t, r_{t+1}\}} \equiv \frac{r_t^{max}}{r_t^{min}}$. Then, the **implementable high-low estimator** is

$$\widehat{S}^{HL} = \omega S + (1 + \sqrt{2}) (\phi R_t^{min} - R_t^*) \quad (2.12)$$

where $\phi \equiv \sqrt{1 + \kappa^2}$ and $\omega \equiv (\phi + \sqrt{2}\phi - 1 - \sqrt{2})$. In practical terms, the implementable version of the estimator scales the underlying true spread by a factor influenced by the variation between daily ranges, which both determine ω and ϕ . The linear functions for the theoretical and implementable high-low spread estimators above make the computation of the high-low estimator bias straightforward.

2.4.2 Bias and Its Sources

The first important result we establish in this section leverages the convenience of the functional forms we derive in (2.11) and (2.12), to compute the estimation bias of the theory-based HL measure and from the actual bias from its empirical implementation. Proposition 1 shows the unbiasedness of the theoretical estimator and then demonstrates that the implementable high-low measure can only be unbiased under the unlikely empirical case of identical consecutive observed ranges. In all plausible application cases, it will necessarily suffer from estimation bias.

Proposition 1. The theoretical high-low spread estimator $\widetilde{S}^{HL}$ is an unbiased measure of the true spread S . The implementable estimator $\widehat{S}^{HL}$ is unbiased only if $r_t = r_{t+1}$.

Proof. Appendix (B.2). □

Unbiasedness requires identical daily price ranges

Proposition 1 establishes a crucial result: while the theoretical high-low spread estimator is an unbiased measure of the true spread, the implementable version of the estimator almost surely results in biased daily spread estimates, except for days with exactly identical observed ranges. Bringing this result to light is a direct consequence of our rewriting of the high-low estimator's functional form in (2.12). This additively separable form further shows that computing sample averages of $\widehat{\beta}$ and γ and then computing high-low

¹³This follows from the trigonometric relation $\sqrt{(\min\{r_t, r_{t+1}\})^2 + (\kappa \min\{r_t, r_{t+1}\})^2} = \min\{r_t, r_{t+1}\} \sqrt{1 + \kappa^2}$.

estimates from (2.10) does not eliminate the bias. Importantly, note that the bias in Proposition 1 is of the same nature as the Jensen’s inequality bias in the Roll measure. We refer to this source of bias in the context of both estimators as moment bias to stress the issue at hand: the use of a sample statistic or daily data points like in the high-low and close-high-low estimators, instead of the theory-derived moment.

Small sample validity of the asymptotic range distribution

To the extent that all other model assumptions hold, the size of the bias from the empirical use of the high-low estimator is completely determined by some function of the distance between r_t and r_{t+1} . The greater the variation between ranges over a two-day interval, the greater the moment bias. In practice, however, even if daily ranges *were equal*, resulting in an identity between the theoretical and implementable versions of the high-low estimator, daily spreads would still remain biased empirically. That is because the unbiasedness established in Proposition 1 relies on $T^{-1} \sum_{t=1}^T R_t^* = T^{-1} \sum_{t=1}^T \sqrt{2}R_t$, which rarely occurs during the estimation time interval of financial data. This small sample bias is again of the same nature as the small sample bias that the sample autocovariance suffers from in the Roll measure.

The computation of bid-ask spreads from short estimation windows, e.g., two consecutive days, is consistent with derivation assumptions only likely holding in short time-series. For instance, Roll (1984) requires the distribution of returns to be stationary, while Corwin and Schultz (2012) and Abdi and Rinaldo (2017) assume that the latent spread is constant.¹⁴ Defending these assumptions becomes increasingly difficult for longer data intervals. However, moment-based estimators are derived using relationships that may not hold in small samples. Particularly, finding a closed-form solution for the high-low estimator (and the close-high-low estimator) relies on the asymptotic distribution of the range from Feller (1951).

The unbiasedness of the theoretical high-low measure rests on the average two-day range being approximately equal to the average range, scaled by $\sqrt{2}$. The same must hold for the implementable HL estimator (in addition to the requirement of constant ranges). To verify how often this condition is met even in an ideal setting, in Figure (2.3) we simulate 10,000 months of trading data with a constant volatility value of $\sigma = 0.03$ and compare daily values and monthly averages of $\sqrt{2}R_t$ and R_t^* from the data. The sample distributions of the two ranges are distinct almost everywhere, as even monthly averages of each quantity seldomly fall on the 45-degree line (in blue in the figure).

Bias size

Next, we bring the two bias sources together and decompose the total bias in the implementable high-

¹⁴Of course, deriving each estimator requires a number of additional assumptions, such as time-constant volatility, or continuous observation of trades, as mentioned before.

low measure into closed-forms of the moment and small sample biases.

Proposition 2. The bias in the implementable high-low estimator ($\hat{S}^{HL}$) can be decomposed as

$$\mathbb{E} [\hat{S}^{HL}] - S = \underbrace{\left(1 + \sqrt{2}\right) \left[\underbrace{r_t^{min} (\phi - \sqrt{2})}_{\text{Moment Bias}} + \underbrace{(\sqrt{2}R_t^{min} - R_t^*)}_{\text{Small Sample Bias}} \right]}_{\text{Bias}}$$

Proof. Appendix (B.2). □

Proposition 2 separates out the bias in the high-low spread estimator bias two components: the moment bias, which depends on observed variables, thereby being quantifiable empirically, and small sample bias, which depends on two latent variables. Consistent with our previous discussion, the *Moment Bias* term vanishes when consecutive observed ranges are identical ($\phi = \sqrt{2}$) and the *Small Sample Bias* term goes to zero under the asymptotic validity of the range distribution in the estimating interval ($R_t^* \rightarrow \sqrt{2}R_t$).

2.4.3 Common Biases Account for 90% of Performance Loss

We now take the bias expression in Proposition 2 to the US stock data. Specifically, we investigate how much the small sample and moment biases account for the empirical performance of the high-low estimator. To answer that, we calculate a “best-performing” version of the high-low measure where we correct daily spread estimates $\hat{S}^{HL}$ as:

$$\widehat{S_*^{HL}} = \hat{S}^{HL} - \text{Moment Bias} - \widehat{\text{Small Sample Bias}} \quad (2.13)$$

in which we empirically retrieve the term *Small Sample Bias* by computing “true” ranges by subtracting daily effective spreads from daily observed ranges. *Moment Bias* is computed directly from the data. The bias-corrected version of the high-low estimator above nets out both theoretically-derived bias sources. The difference in performance between the measure and the standard high-low estimator (which tracks stocks by 63% in the cross-section and 48% in the time-series) therefore captures the performance loss in the implementable high-low spread attributed to the moment and small sample biases.¹⁵

¹⁵In the empirical applications using non-simulated data, we consider the effective spread — the underlying variable the simple bid-ask spread estimators we study attempt to recover — to correspond to the latent or “true” spread. We refer to the difference between a spread estimate and the effective spread in such contexts as empirical bias to maintain the language concise. Later in the paper, we refer to this difference interchangeably as measurement error in the linear regression framework to stress the econometric issues associated with the estimator’s bias in that case.

Figure (2.4) replicates Figure (2.2) using monthly averages of the bias-adjusted HL estimator and monthly effective spreads. Once again, we analyze cross-sectional correlations for the deciles of stocks sorted according to liquidity levels, plotting in the figure the within-subsample cross-sectional correlation between the measure and the effective spread as well as the average frequency of daily negative estimates. By removing common biases from high-low spread estimates, the adjusted proxy performs in the overall cross-section significantly better than the standard implementable estimator.

The overall cross-sectional correlation increases from about 63% to 94%. More strikingly, this substantial performance improvement occurs in all spread deciles whose correlation with effective spreads increases anywhere from two-fold to almost four-fold. This generalized improvement in performance significantly reduces the influence of few illiquid stocks in determining the cross-sectional correlation, which drops only by 5% after removing the top stock decile, instead of 25% as in the exercise in the previous section.

Since our derived bias expression assumes all high-low model assumptions hold, the residual noise in the cross-sectional correlation (6%) captures any estimation bias solely attributable to the microstructure of the US stock sample — idiosyncratic bias. That includes, for example, direct empirical violations from assuming constant spreads within two consecutive days or constant volatility. Counterfactually, this implies that if we were able to make the high-low estimator perfectly accurate, 90% of the performance improvement necessary would be attributed to eliminating the two biases that are also common to the Roll measure.¹⁶

Conducting a similar counterfactual exercise to evaluate the time-series performance of the high-low estimator corrected for common biases reveals similar results. The sample-wide time-series correlation improves from around 48% for the high-low estimator to 80% after adjusting for the moment and small sample biases. We select two stocks, each on opposite sides of the liquidity spectrum of US stocks, to contrast the change in performance in the time-series. Figure (2.5) plots estimated monthly averages of effective spreads, high-low estimates, and the best-performing HL measure for Tesla (\$TSLA) and ATAC Resources (\$ATC), a Canadian mining company. Tesla monthly effective spreads are on average 0.14% and ATAC’s over 6%. We use the zero adjustment for comparison as the other high-low *ad hoc* adjustments or using both negative and positive estimates make the measure’s baseline time-series correlation null or even negative.

High-low estimates are over ten-fold the size of Tesla trading costs, correlating only 16% in the time-series. After adjusting for the two common biases, the estimator tracks the underlying spread very closely, with a correlation coefficient of 80%. Among the stocks with largest effective spreads, ATAC offers a more favorable empirical setting for the high-low estimator, which tracks effective spread levels more closely and

¹⁶Note that the unexplained performance improvement of 10% would not only include underlying properties of stock prices, but also to mechanical biases introduced by any *ad hoc* adjustment, such as zero imputation, whenever those are implemented. This underscores the overwhelming importance of the moment and small sample biases in affecting the performance of the spread estimator.

correlate at about 50% — still somewhat poorly, nonetheless. Performance gains from correcting the biases in this case are relatively smaller, but the post-correction time-series correlation is again about 80%.

2.4.4 Determinants of Bias

Our framework thus far provides a decomposition of the bias in the high-low estimator into two components, which we show are structurally identical to the biases in the Roll measure. These common biases account almost entirely for the performance loss in the high-low measure. However, one puzzle remains — why is the performance of the high-low and other simple spread estimators much worse for relatively liquid stocks? The answer to this question is not only crucial to design better performance evaluation settings for new bid-ask spread estimators, but to increase the degree of confidence in out-of-sample use. If simple bid-ask spread estimators perform worse in liquid markets, without ex-ante knowledge on whether their specific setting is “liquid”, how confident can researchers be in the use of low-frequency spread proxies?

Answering why spread estimators perform worse in liquid assets empirically is challenging because stocks like \$TSLA and \$ATC might systematically differ from each other along other dimensions than just effective spread levels. Any of these unobservable factors could drive differences in estimation performance. To cleanly observe whether there exists a relationship between bias size and the latent spread, we use a simulation environment with “ideal” financial data where once at a time, we allow for only the true spread or price variance to change, while holding constant all other pricing features. This is equivalent to imposing that \$TSLA and \$ATC are identical in all dimensions other than the true spread and volatility. The simulated environment has the additional benefit of controlling for potential model violations (idiosyncratic biases), as we set the data generating process of stock prices to closely match derivation assumptions.

Simulation evidence: bias, spread size, and frequency of negative estimates

To isolate potential model assumption violations — what we term idiosyncratic bias — in driving the observed relationship in US stock data between bias, negative estimates, and the effective spread size in the stock data, we use a simulation environment with financial price series data providing an “ideal” data generating process. This exercise confirms the importance of the underlying spread and volatility magnitude in determining the bias, which we then formally prove. It also indicates a similar relationship with respect to the frequency of indeterminate estimates, which we explore in the the Online Appendix B.

We generate 10,000 twenty-one-day months of efficient price data. Prices are created for 390 minutes m on each day, following $\mathcal{P}_m = \mathcal{P}_{m-1} + x\sigma_m$, where $x \sim N(0, 1)$, and $\sigma_m = \sigma/\sqrt{390}$ is the standard deviation per minute. The daily standard deviation of efficient price returns is set constant to $\sigma = 3\%$. Observed

prices are then calculated after compounding or discounting true prices by half the spread S , as in equation (2.1). For every disjoint interval of 390 minutes, we compute daily high, low and closing observed prices. At the beginning of each month, we normalize the true price as \$ln 100. This data generating process is identical to the simulation environment in Corwin and Schultz (2012) and reflects an ideal “empirical” setting since both volatility and spread levels are held constant, and prices are always observed.

Our main goal is to analyze the behavior of estimation bias with respect to the underlying latent spread, particularly for spread sizes below 0.5%, as these correspond to the vast majority of US stock liquidity costs. Spreads below 0.5% are generally omitted from similar simulation exercises in the literature. The first panel in Figure (2.6) compares average monthly high-low estimates (with the zero adjustment) with the underlying, true spread level in the data generating process from 10,000 months of trading data replicated with true spread increments of 0.1%. Two salient features emerge. Estimated spreads are almost always different from the underlying true spread in levels and smaller trading costs are associated with overstated estimates, which then become downward-biased for larger spread levels. This is confirmed in the second panel, where an effective spread of 0.1% generates a bias ($\hat{S}^{HL} - S$) of 1.2%. As we increase the latent spread size, the bias size gradually declines, until eventually becoming negative for large effective spreads.

Effect of spread and volatility on the bias size

The relationships unveiled in the simulated environment suggest a strong relationship between the high-low estimator bias size and the latent spread. To properly gauge such relationships as structural, however, we must again turn to theory. Proposition 3 exploits the convenience of the bias formula derived in Proposition 2 to establish how the bias magnitude behaves as we shock the underlying true spread size.

Proposition 3. The bias in the implementable high-low spread estimator is decreasing in the true spread. Given sufficiently high values of volatility relative to the the latent spread (about 4 times), the bias increases in price volatility.

Proof. Appendix (B.2). □

Proposition 3 shows that increasing the latent spread size decreases the *Moment Bias* term. Intuitively, the term is zero when $r_t = r_{t+1}$, which implies ϕ must equal $\sqrt{2}$. The parameter ϕ is in turn given by $\sqrt{1 + \frac{r_{t+1}^2}{r_t^2}}$, where the observed range is an additively separable linear function of S . Since the spread over the two-day trading interval is assumed constant, very large spreads move the ratio closer to one, regardless of

the values of each day’s true range. That is, large spreads make distinct observed ranges closer in magnitude, therefore decreasing the estimation bias size.

The role of volatility

Proposition 3 further establishes that the high-low estimation bias increases in volatility when σ is sufficiently large relative to the underlying spread ($\sigma/S \approx 4$). The current effective spread levels of US stocks likely attain this noise-to-signal ratio, making volatility an additional driver of empirical bias in the high-low measure. Nonetheless, because the estimation bias always decreases in the spread and volatility only matters when spreads are sufficiently low, we focus our main analysis on the relationship between bias and frequency of negative spreads with the latent spread, while keeping the volatility of the DGP constant. For completeness, we show the effect of σ on negative estimates for various levels of S in the Online Appendix B.

Lessons from model predictions to other simple spread estimators

Note that the effect of underlying spread and volatility levels on the bias we derived in the high-low estimator generates the same predictions as in the Roll estimator. Given our earlier result that the HL proxy suffers from the same common biases as the Roll estimator, it is perhaps not surprising that both measures become less accurate when the underlying spread decreases and volatility increases. Albeit we do not derive a formal bias expression for the close-high-low estimator, we show in Online Appendix Figures (C.1) and (C.2) that by and large the same relationships hold in the simulated data, as they also do in the US stock sample.

2.5 Measuring Bias Empirically

The high-low estimator bias is comprised of a moment bias term and a small sample bias term. The moment bias can be directly calculated from the data. The small sample bias, however, is a function of true daily ranges which are unobservable. Even if we wanted to estimate it using $\sqrt{2}\mathbb{E}[R_t^{min}] - \mathbb{E}[R_t^*]$, this would require a complicated approach to compute these moments in small samples, as the *Small Sample Bias* should be zero in expectation. Because of that, even though empirically correcting the high-low estimator for both biases would make its performance nearly perfect, such approach is unfortunately not feasible.

Without being able to point-identify the estimator’s empirical bias, in this section we pursue the next best option: deriving theory-consistent bounds on the bias. These bounds are easily computed from the same data used to estimate the high-low measure. They are also sufficiently tight so that they are informative in

a variety of settings — to both quantify the bias and control for non-classical measurement error in linear regressions, in the cross-section and time-series. The bias bounds are derived under the same assumptions as the high-low estimator and despite potentially idiosyncratic biases remaining, are empirically consistent in the vast majority of cases.

2.5.1 Deriving Empirical Bias Bounds

Begin by rewriting the closed-form expression for the high-low bias in Proposition 2 as $S = \hat{S}^{HL} - \text{Moment Bias} - \text{Small Sample Bias}$. By construction, the moment bias is always non-negative, since $\phi \geq \sqrt{2}$. Note that when $\text{Small Sample Bias} > 0$, the following expression is always valid: $S < \hat{S}^{HL} - \text{Moment Bias}$, so that the true spread S is bounded from above by correcting the high-low estimate for the moment bias.

To make progress toward signing the small sample bias term, equation (2.14) decomposes the term $\sqrt{2}R_t^{min} - R_t^*$. This identifies the variation that generates the failure of asymptotic distribution of the range to hold in financial data. Simply put, the small sample bias is non-zero when consecutive two-day price ranges differ significantly (the price variation on one day is much higher than the other day) or when one price range shifts vertically with respect to the previous range (one day experiences a significant and sustained price jump or drop):

$$\sqrt{2}R_t^{min} - R_t^* = \frac{1}{2} \underbrace{(r_t^{min} - r_t^{max})}_{\Delta \text{ range}} - \underbrace{(\eta_t^{max} - \eta_t^{min})}_{\text{Shift in range}} + (\sqrt{2} - 1) R_t^{min} \quad (2.14)$$

By replacing the last term in (2.14) with the *observed* minimum range, $(\sqrt{2} - 1) r_t^{min}$, we calculate a modified version of the small sample bias, $\widetilde{\text{Small Sample Bias}}$, to verify its sign empirically. Note that this substitution implies necessarily that the true small sample bias term is overestimated, so that the procedure will fail to identify some number of negative small sample bias terms by mistakenly attributing them as positive.

Figure (2.7) plots 40,000 randomly chosen stock-day estimates of the “true” small sample bias, calculated using effective spreads, and $\widetilde{\text{Small Sample Bias}}$ (patterns for the full stock sample are identical). Blue markers indicate a sign correspondence between the two biases. For 90.5% of stock-day data points, using the computed version of the small sample bias that only uses observed data correctly predicts the sign of the small sample bias (the accuracy is identical using all stock-days). As we show next, this high accuracy rate is more than sufficient to robustly sign the small sample bias empirically. False negatives or false positives are displayed in gray, comprising the remaining 10.5% of observations. From these erroneous predictions,

the vast majority represent false positives, which aligns with the fact that $(\sqrt{2} - 1) r_t^{min} > (\sqrt{2} - 1) R_t^{min}$.

To derive a lower bound on the high-low estimation bias, or equivalently, the minimum empirical bias, we compute

$$\widehat{Bias_n}^{min} = \widehat{S_n}^{HL} - f_{n(t)} \left(\left(\widehat{S_t}^{HL} - Moment\ Bias_t \right) \cdot \mathbb{1} \left\{ \widetilde{Small\ Sample\ Bias_t} > 0 \right\} \right)$$

where the minimum empirical estimation bias $\widehat{Bias_n}^{min}$ at frequency n (e.g., monthly, annual) is obtained from discounting monthly averages of high-low spreads with a measure of central tendency — either the monthly median or average — of upper bounds on the latent spread. Using median or mean values of the bound produces very similar empirical results, although the median usually gives slightly tighter bias bounds because it reduces the contribution of false negatives or positives when calculating *Small Sample Bias*.

To obtain an upper bound on the estimation bias, note that because the small sample bias is a function of price volatility and the moment bias is determined by both volatility and the latent spread, the two biases are highly correlated: above 91% in the US stock cross-section. When empirically computing the small sample bias with *Small Sample Bias*, the relationship between the two common biases becomes even stronger, as the term further introduces the latent spread into the small sample bias computation. This has an important empirical consequence: positive values of *Small Sample Bias* imply corresponding moment bias terms greater than the true small sample bias in over 96% of US stock-days, such that $S > \widehat{S}^{HL} - 2 \times Moment\ Bias$, since $Moment\ Bias > \widetilde{Small\ Sample\ Bias}$.

By determining positive values of *Small Sample Bias*, we can then derive an upper bound on the high-low estimation bias — the maximum empirical bias — as:

$$\widehat{Bias_n}^{max} = \widehat{S_n}^{HL} - f_{n(t)} \left(\left(\widehat{S_t}^{HL} - 2 \times Moment\ Bias_t \right) \cdot \mathbb{1} \left\{ \widetilde{Small\ Sample\ Bias_t} > 0 \right\} \right)$$

The interval $\left[\widehat{Bias_n}^{min}, \widehat{Bias_n}^{max} \right]$ provides researchers a theory-consistent approach to bound the empirical estimation bias from using the high-low spread estimator. Computing the bounds requires the same daily data used to originally obtain the high-low measure — high and low daily prices — and imposes minimum computational burden. The method also involves no *ad hoc* modifications or adjustments.

2.5.2 Empirical Implementation

We now assess whether the derived bias bounds are empirically useful in a variety of dimensions. We state the implementation of the empirical bounds in general terms, but remain using the US stock sample throughout to keep it consistent with all of our previous results.

Assume a researcher has computed daily high-low estimates from a certain dataset with i units (e.g., stocks, futures contracts, bonds) and $n(t)$ months of trading data with t days. The researcher then calculates monthly high-low spread averages for each unit in the cross-section. The researcher may be interested in asset pricing regressions, where returns are regressed on liquidity risk — proxied with the high-low spreads — in the style of Acharya and Pedersen (2005) and Pontiff and Singla (2019) or to control for liquidity as in McLean and Pontiff (2016). Our previous results show that simply using low-frequency spread estimates in these cases is severely problematic. Because estimation bias is a function of the underlying latent spread, linear regressions using estimated spreads will suffer from non-classical measurement error.¹⁷

Alternatively, the researcher may be interested in the actual measured spread values, perhaps to sort a portfolio on liquidity or to exploit changes in liquidity costs as a result of an intervention or structural changes in markets over time. In any application where the level of spread estimates matters, the use of simple spread estimators would lead to misleading results, as units with lower liquidity costs would look like illiquid assets, since estimation bias decreases in the spread size. Researchers need to have tools to assess whether a simple bid-ask spread estimate of 1% is a good proxy of an underlying spread of about 1% or whether the actual effective spread is instead 0.2% or 2%.

Empirical Bias in the Cross-Section

Our first empirical exercise evaluates the bias bounds in the cross-section of stocks. After computing daily high-low estimates and their monthly averages, we compute daily values of the moment bias, $(1 + \sqrt{2}) r_t^{min} (\phi - \sqrt{2})$, and daily estimates of the small sample bias, *Small Sample Bias*. Note that only days with positive high-low estimates can be used to compute estimation bias, as negative estimates are not theoretically-consistent. To obtain monthly empirical bias bounds, we apply the formulas for $\widehat{Bias}_t^{min}$ and $\widehat{Bias}_t^{max}$ using days with positive small sample bias estimates. We recommend choosing the median for $f_{n(t)}$ instead of the mean as this reduces the potential influence of idiosyncratic biases in making the signs of *Small Sample Bias* and the true small sample bias fail to coincide.¹⁸

Figure (2.8) plots cross-sectional averages of three variables: the actual high-low empirical estimation bias ($Bias_n$), which measures the average percentage-point deviation of the estimator from the effective spread, and the minimum and maximum empirical estimation bias bounds. We present monthly averages over our time period in three panels — all US stocks, US stocks below the median effective spread — about

¹⁷More specifically, because measurement error in the context of simple bid-ask spread estimators is negatively correlated with the effective spread — that is, mean reverting — the attenuation bias decreases, which may even flip the sign of the estimated effect.

¹⁸Although the choice between the mean or median in our US stock sample leads to minimal empirical differences, false positives are associated with very large bounds on the true spread, so that using the median in applications where idiosyncratic biases may be important is akin to trimming abnormal values.

0.21% — and above the median. Across all panels, the estimated biases successfully track and bound the actual high-low estimation bias. The correlation between each bound and $Bias_n$ is remarkable — over 98%. Therefore, including the bias bounds or its midpoint in linear regressions would be highly effective in controlling for non-classical measurement error when using daily or monthly spread estimates.

Recovering Effective Spreads

The bounds not only provide the researcher with a precise estimate of how much bias on average the high-low estimator suffers from in her particular context, but also allow her to obtain theoretically-consistent bounds on the effective spread, significantly improving the degree of confidence in the use of the measure. To see how, consider that the researcher has calculated the average monthly high-low estimate for all US stocks (top panel of the figure), which is 1.80%. Next, she obtains the average estimated lower bound on the bias — 0.54% — and the estimated maximum empirical bias of 1.58% (for reference, the actual estimation bias in the US stock sample is 0.95%, which the researcher cannot observe in practice but that is successfully bounded by our bias estimates).

Discounting the high-low estimate with each bound, that is, $1.80\% - 1.58\% = 0.22\%$ and $1.80\% - 0.54\% = 1.26\%$, yields back implied effective spread bounds, which in this case results in the prediction that the actual effective spread is between 0.22% and 1.26%, or should be approximately equal to 0.74% at the midpoint. The average monthly effective spread in the sample is 0.85%. This means that the implied effective spread midpoint generated by our empirical bias bounds is only 15% above the effective spread, while the high-low measure is 112% higher.

Replicating the exercise for stocks below the effective spread median gives a similar high degree of precision, with an important distinction. Recall that when we derive the maximum empirical bound on the spread, the relationship $S > \hat{S}^{HL} - 2 \times \text{Moment Bias}$ is implied by a positive value of the estimated small sample bias. Note that when the true spread S is very small, the moment bias term approaches the small sample bias, so that the inequality becomes close to binding.¹⁹ This can be seen in the second panel of Figure (2.8), where the actual estimation bias falls very close to the upper bias bound.

As a result, if the researcher has reason to believe that her application context involves liquid assets, using the upper bound to obtain the implied effective spread will more accurately approximate the effective spread *level* than the implied spread midpoint — 0.03% for an effective spread of 0.09% and a high-low estimate of 1.28%. To control for estimation bias, however, any of the bounds correlates similarly well, above 98%. As an additional example, we revisit the example of Tesla stocks we used in Figure (2.5).

¹⁹Intuitively, because the moment bias depends both on volatility and the spread and the small sample bias only on volatility, a very small spread makes both variables have a perfect correlation. Formally, the moment bias is equal to the small sample bias when $R_t^{max}/R_t^{min} \approx 1.24$, given $\lim_{S \rightarrow 0} \left\{ \tau_t^{min} (\phi - \sqrt{2}) - \sqrt{2}R_t^{min} + R_t^* \right\}$.

Applying the upper bias bound in the \$TSLA time-series recovers an average implied effective spread of 0.26% (for an effective spread of 0.14% and a high-low estimate of 1.78%).

An inverted relationship exists in illiquid assets. The actual estimation bias will approximate more closely the lower bound on the estimation bias, as very large spreads make the inequality $S < \hat{S}^{HL} - \text{Moment Bias}$ closer to binding. This appears in the bottom panel of Figure (2.8), where the lower bound tracks the estimation bias very closely. In cases where the researcher has reason to believe that spreads are very large, using the lower bound will give more accurate implied effective spread levels than the midpoint. For the subsample of stocks above the median effective spread, this approach recovers a spread estimate of 1.64% for an effective spread of 1.46% (and a high-low estimate of 2.21%). Revisiting the time-series example of the highly illiquid stock \$ATC from Figure (2.5) and using the lower bias bound yields an almost perfect point-estimate of the effective spread: 6.26% for an effective spread of 6.26% and a high-low of 5.72%).

Empirical Bias in the Time-Series

We now use the bias bounds to estimate *stock-level* high-low biases. Figure (2.9) shows again the actual high-low empirical estimation bias and the minimum and maximum bias bounds. This time, we average monthly estimates for each stock over the entire time period, so that each stock has a single value for each of the three bias variables. The top panel shows the distribution of the empirical bias and its bounds for all US stocks. We sort observations along the x -axis from the largest estimation bias (in percentage-points) to the smallest (that is, when the high-low measure understates the effective spread). To make the interpretation easier, we report the x -axis as the cumulative share of stocks from the left (0% of the total sample) to the right (100% of stocks).

Consistent with the previous results in the cross-section, the performance of the empirical bias bounds in the time-series is also robust. For the vast majority of stocks, the true estimation bias lies inside the theory-consistent bounds, which behave properly in almost all stocks that generate upward-biased high-low estimates. This includes around 80% of the stocks in the cross-section. Stocks that generate downward-biased high-low spreads — primarily illiquid stocks — are the only cases where the actual empirical bias lies outside the bounds $\widehat{Bias}^{min}$ and $\widehat{Bias}^{max}$.

To gain further insight into the relationship between the estimated bounds and the actual bias, the second panel in the figure shows results only for stocks with spreads below the sample median of about 0.21%. In this case, almost all stocks have an actual high-low bias inside the estimated bounds, which almost perfectly track bias levels from 4 to about 1 percentage point. When normalized by the underlying effective spread, these biases amount anywhere from 5-fold to 20-fold the actual spread level. Similarly, the last panel in the figure shows the performance of the bias bounds for stocks above the median effective spread.

This subsample includes the illiquid stocks for which the high-low estimator produces severely understated spreads, which lead to the estimation bias to lie outside the bounds. Nonetheless, for actual biases from 12 to 0.5 percentage point — about 75% of stocks in this subsample — the empirical bounds once again work consistently well.

Sorting Stocks Based on Liquidity

The analysis above illustrates the time-series usefulness and reliability of the empirical bias bounds. Researchers can credibly estimate the high-low bias to measure it directly or report it to establish credibility of spread estimates. To provide another example of how researchers are able to benefit from the approach, we analyze the gains in predictive power when sorting stocks on liquidity by using the bias bounds. Even coarsely sorting assets may result in severely distorted rankings and portfolios if the levels of the sorting variable fail to pick up the latent factor. For example, a researcher interested in dividing stocks in portfolios based on liquidity levels who would use the high-low proxy would assign several highly liquid assets to an illiquid portfolio, since these spread estimates will be more upward biased than estimates of actually illiquid assets.

Our minimum bias bounds can be used to “recover” theory-consistent bounds on the effective spread, lending credibility to the use of the high-low estimator in settings where estimate levels across the cross-section matter. Table (2.1) performs a series of exercises confirming this point. The first question we answer there is: how well does the ranking of spread estimates predict the actual effective spread ranking? In columns (1) through (4), we regress a ranking variable at the stock level — which orders observations from the lowest effective spread to the largest effective spread — on the spread ranking generated by the high-low measure without *ad hoc* adjustments and with the zero adjustment, and by correcting the high-low spread with the bias bound.

The first column shows that a researcher using only the positive spread estimates obtained from the high-low measure would predict the correct ranking of effective spreads by about 72%. Using the zero adjustment increases the predictive power to 82%, since the *ad hoc* adjustment introduces mechanical attenuation bias that coincidentally offsets some of the upward bias in liquid stocks generated by the small sample and moment biases. Sorting based on the bias bound correction significantly improves accuracy — a move of 10 positions in the liquidity rank is associated with a move of 9 positions implied by our measure. The spread implied by the bias bounds absorbs 80% of the variation in the effective spread ranking, several percentage points more than the standard high-low measures. This last point becomes even clearer when we regress the effective spread rank on both the high-low and the bound-corrected spread, where almost no additional variation is absorbed vis-à-vis only using the bound-corrected spread.

Columns (5) through (8) perform a similar exercise, now using a more coarse sorting approach. We divide stocks in 10 portfolios based on effective spread deciles and regress the decile rank on the similar measure constructed with the spread estimates. The overall direction and differences in magnitude of results are identical to the previous columns, with the spread implied by the bias bounds performing significantly better in allocating stocks to the correct portfolio based on the effective spread.

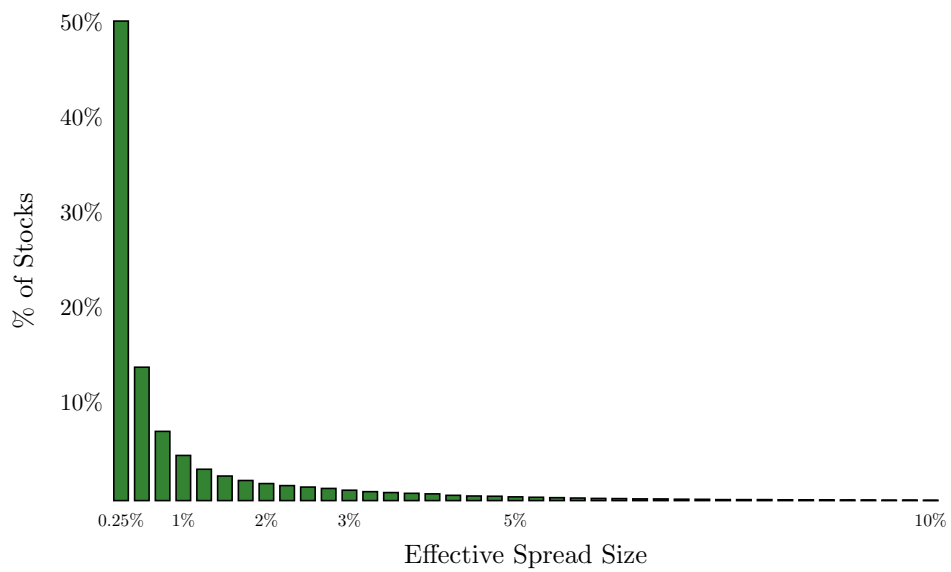
2.6 Conclusion

This paper studies why simple bid-ask spread estimators present conflicting performance results in different markets, often yield negative or indeterminate estimates, and how to empirically assess and address these issues. Using the popular high-low estimator as our main setting, we show that the estimator’s misbehavior is almost entirely driven by two common biases: a moment and a small sample bias. These are the same sources of bias found in the seminal Roll estimator, with idiosyncratic factors dominating relative to specific market settings, or conditions, and model assumptions failure. We show how to empirically measure the minimum and maximum estimation bias from using the high-low measure, equipping researchers to remedy issues arising with the use of low-frequency spread estimators, including non-classical measurement error in regression analysis, portfolio sorting on liquidity, and tracking spread levels in the cross-section and time-series.

We also draw several recommendations from our analysis, from the development of new estimators to performance evaluation. To more credibly establish the quality and recommend a certain spread proxy, performance evaluation studies using tick data should include more detailed subsampling analysis, particularly by analyzing the relationship of correlation coefficients in the cross-section across different effective spread groups. Moreover, we also suggest in sensitivity analyses that performance is measured after dropping the most illiquid units from the sample as well as in random subsamples from the data when the cross-section is sufficiently large. Coefficient stability across these different tests indicate a greater degree of out-of-sample confidence.

Going forward, newly developed spread estimators using transaction data that share the structure of the Roll framework should be subjected to rigorously designed simulation environments with true spreads in the data generating process well below 0.5%, the usual minimum spread used in the literature, and verify the role of price volatility and trading costs in potential misbehavior. More preferably, a formal analysis of the existence of the moment or small sample bias in new measures is also suggested, as these two sources severely limit performance of simple bid-ask spread estimators.

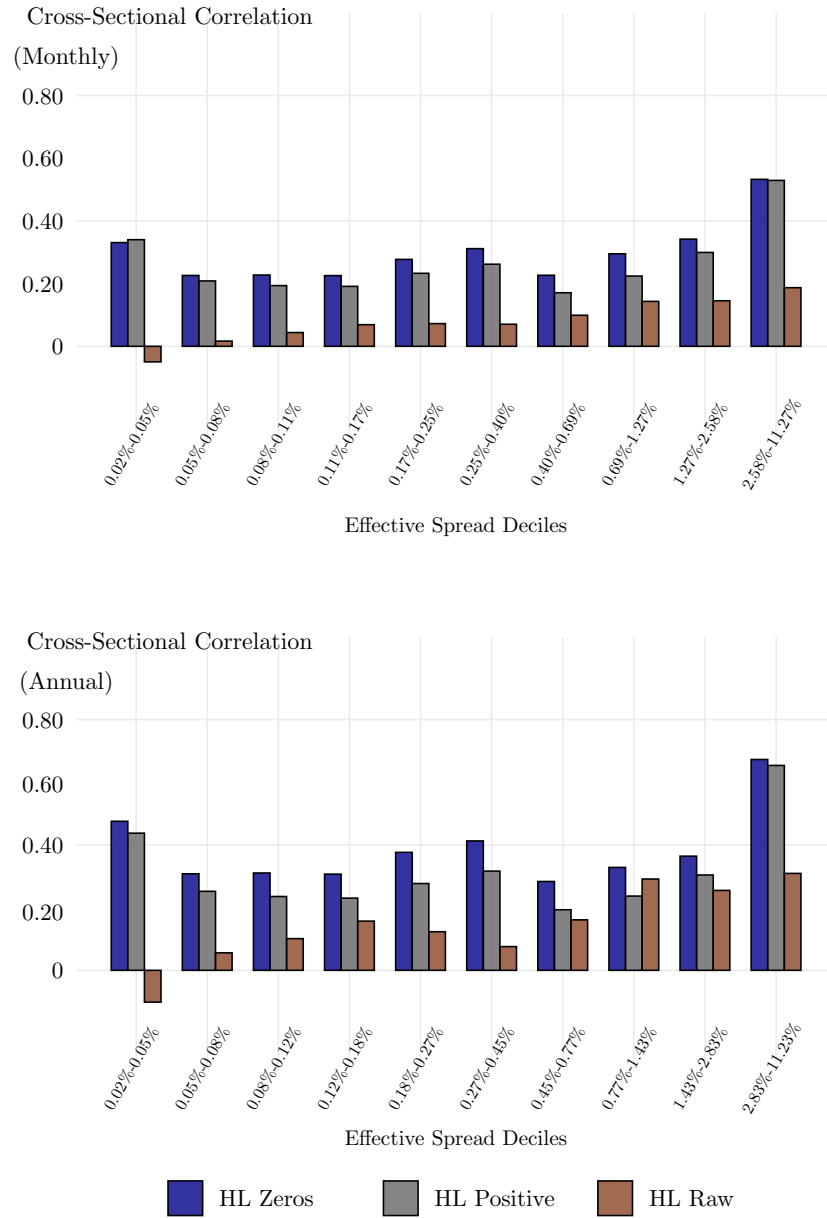
2.7 Figures and Table



Over 70% of US stocks have effective spreads lower than 1%. There are very few stocks with spreads greater than 5%.

Notes: This figure shows percentiles of monthly effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

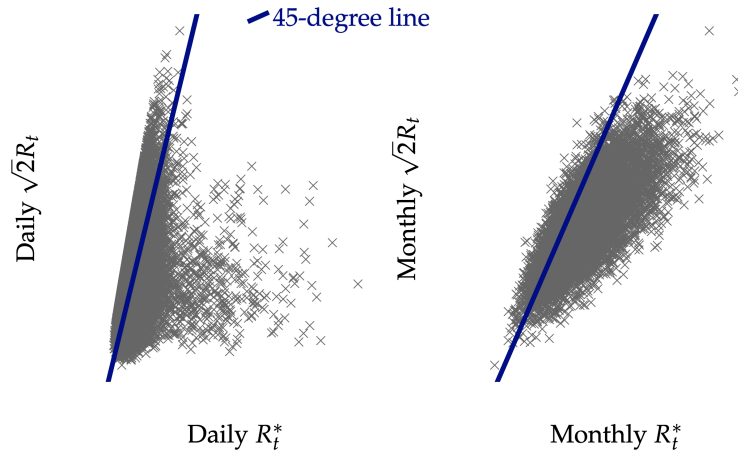
Figure 2.1: Distribution of Effective Bid-Ask Spreads in US Stocks



In all effective spread size deciles, monthly averages of high-low estimates without ad hoc adjustments correlate poorly with monthly average effective spreads. Correlations from the estimates with the positives and zeros adjustments behave similarly, with the best performance in the least liquid stocks (above 40% correlation with effective spreads).

Notes: This figure shows cross-sectional correlations between three versions of the high-low spread estimator and effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

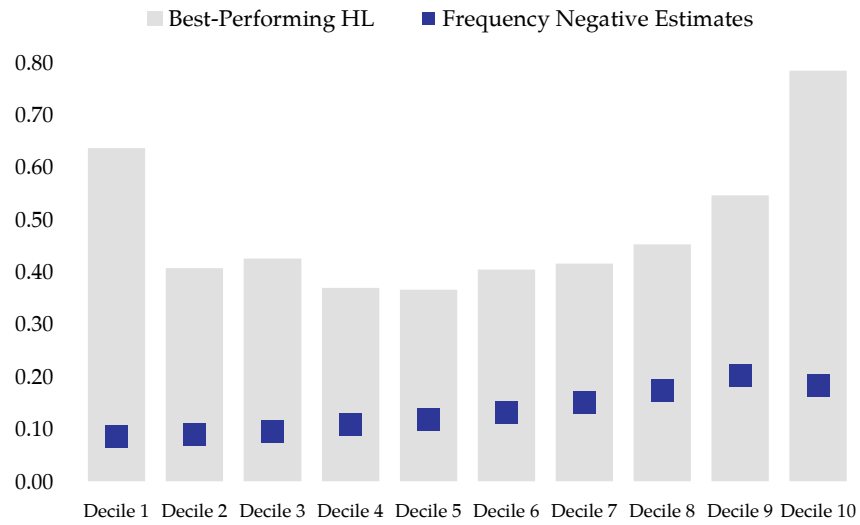
Figure 2.2: Cross-Sectional Correlation: HL Estimates and Effective Spread By Effective Spread Decile Size



The identity required for the small sample bias in the high-low estimator to be zero fails to hold in almost all of 210,000 trading days and 10,000 monthly averages.

Notes: This figure compares 210,000 daily values of $\sqrt{2}R$ and R^* (left) and 10,000 monthly averages of each range in trading data simulated with ideal conditions. The main text contains more information on the data generating process.

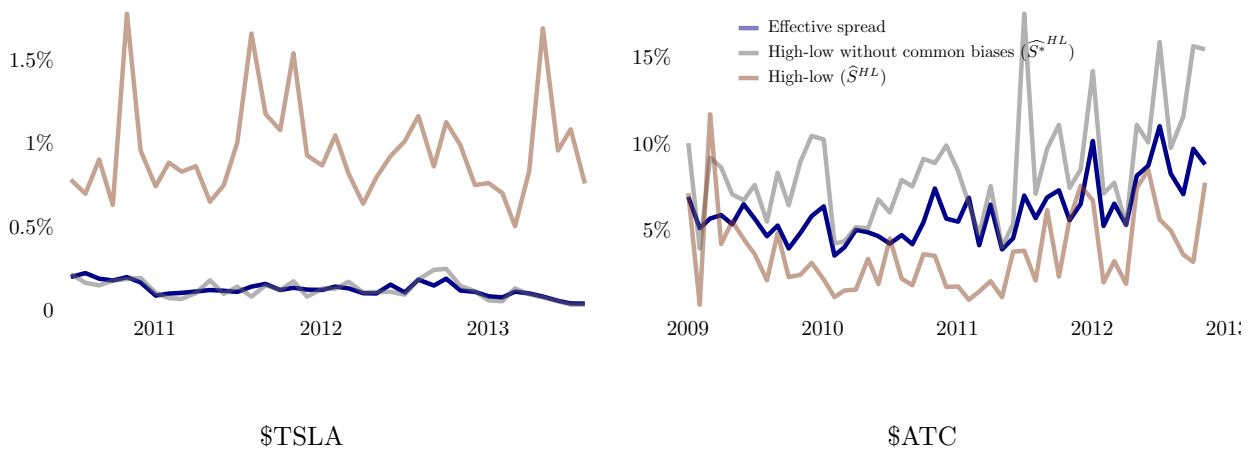
Figure 2.3: Relationship Between $\sqrt{2}R$ and R^* Under Ideal Conditions



In all effective spread deciles, monthly averages of the best-performing high-low estimator correlate twice as well with monthly average effective spreads as high-low estimates.

Notes: This figure shows cross-sectional correlations between our calculated best-performing high-low spread estimator and effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1). The *Best-Performing HL* is a bias-free version of the high-low measure, obtained as $\hat{S}^{HL} - \text{Moment Bias} - \text{Small Sample Bias}$.

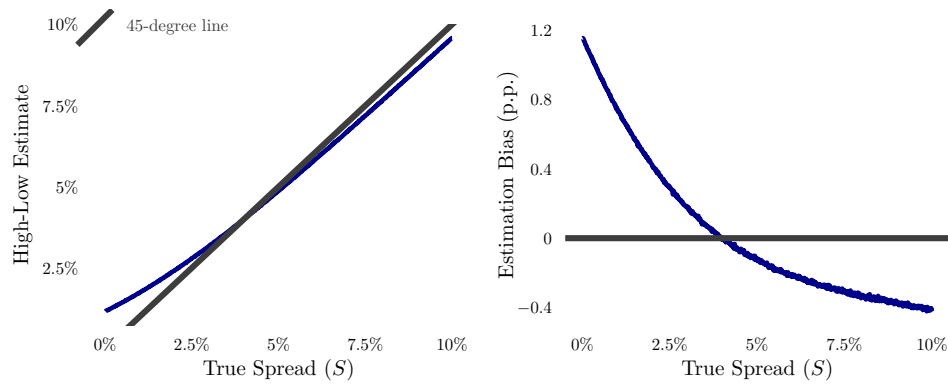
Figure 2.4: Cross-Sectional Correlation: Best-Performing HL & Effective Spread by Effective Spread Decile Size



The best-performing high-low measure tracks the effective spread almost perfectly in a very liquid stock (\$TSLA). When the stock is relatively illiquid (\$ATC), the measure still improves considerably on the standard high-low estimator.

Notes: This figure shows monthly averages of effective spreads and two versions of the high-low estimator for Tesla and ATAC Resources stocks. The *HL Zeros Adjustment* represents monthly averages of the high-low estimator with zeros imputed for negative daily estimates. The *Best-Performing HL* is a bias-free version of the high-low measure, obtained as $\hat{S}^{HL} - \text{Moment Bias} - \text{Small Sample Bias}$.

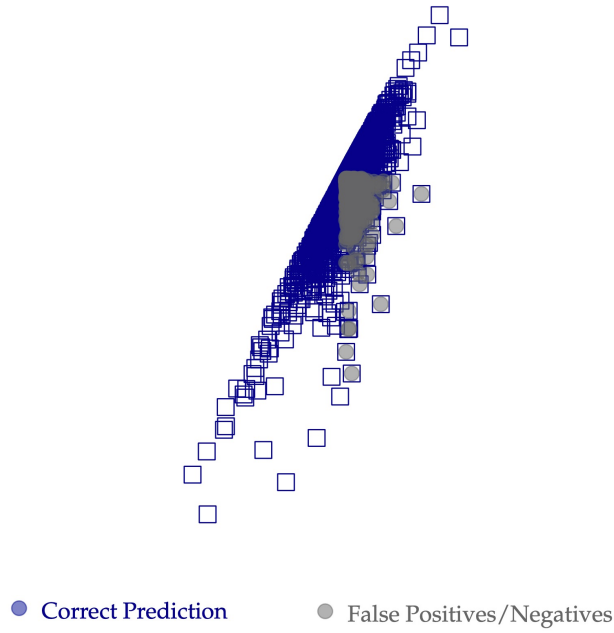
Figure 2.5: Two Examples of Performance: Best-Performing HL Estimator



For small spreads (below 1%), the high-low estimator suffers from an upward bias. As the spread becomes larger, the bias decreases. For reference, the median effective spread in US stocks during 2003-2015 is 0.3%.

Notes: This figure computes 10,000 monthly averages of the high-low spread estimator (with zeros adjustment) for each 0.1-percentage-point increment in the spread between 0.1% and 10% from simulated data. The data generating process is described in detail in the text and is designed to maintain all model assumptions, including constant true spreads in each sample. Bias is defined as the difference between the estimated spread and the true spread.

Figure 2.6: Bias and True Spread Under Ideal Conditions

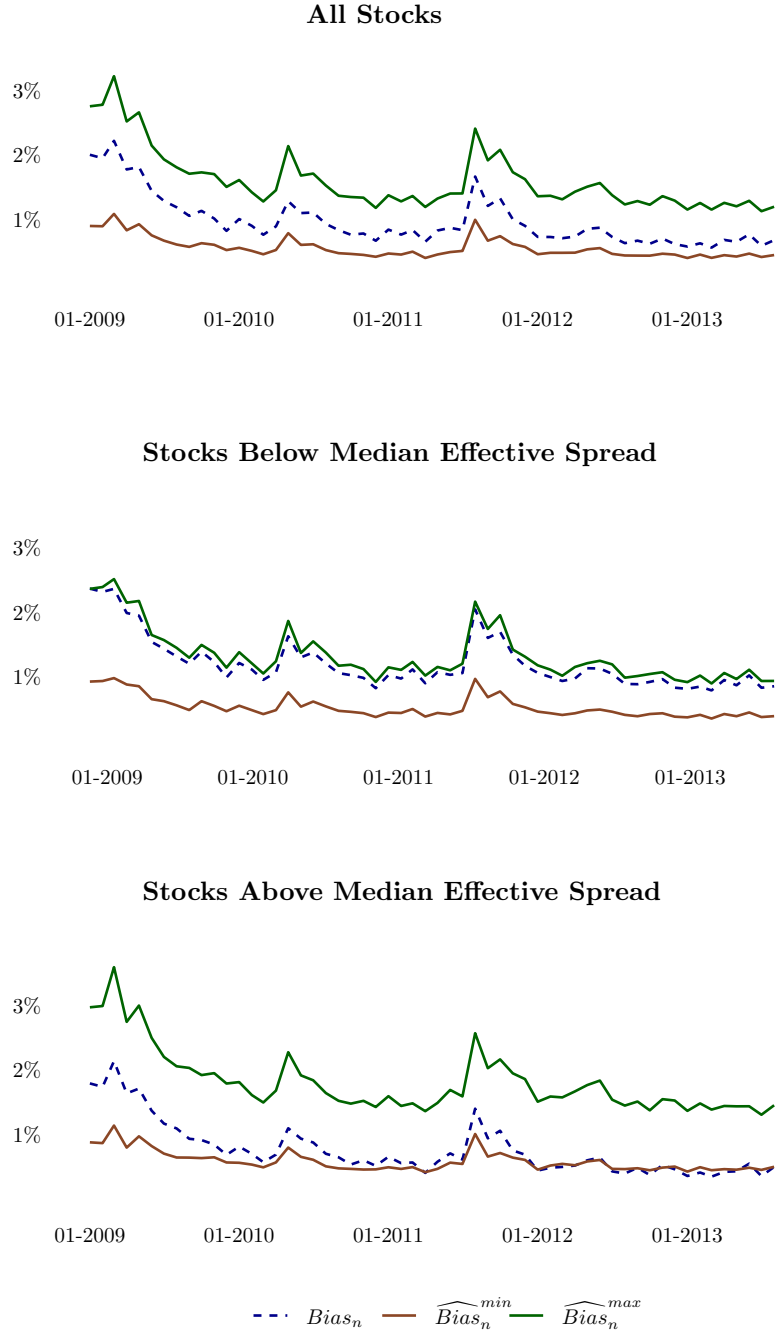


Small Sample Bias

The empirical approach to predict the sign of the small sample bias has over 90% accuracy.

Notes: This figure plots 40,000 randomly chosen stock-day estimates of an empirical approach to measure the small sample bias in the high-low measure. *Small Sample Bias* is calculated as $r_t^{min} + (1 + \sqrt{2})(0.5\Delta r_t - \Delta \eta_t)$. Blue markers indicate the correct correspondence between the respective derived small sample bias and its theory-implied value, thereby correctly predicting the sign of the small sample bias. Gray markets give cases where such correspondence fails, resulting in false positives or negatives.

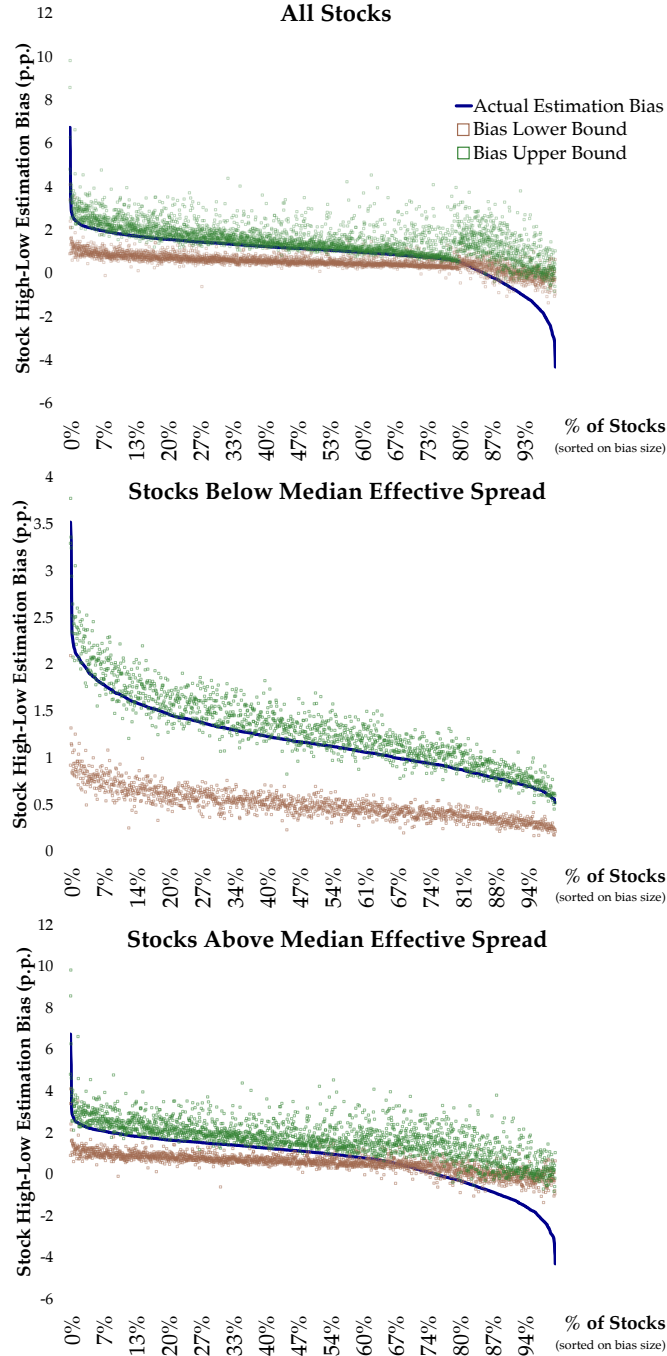
Figure 2.7: Testing the Accuracy of the Empirical Approach to Sign the Small Sample Bias



The empirical estimation bias bounds track the actual estimation bias almost perfectly.

Notes: This figure compares the average cross-sectional high-low estimation bias with three measures: $Bias_n$, which is the actual empirical bias of the high-low spread estimator, the maximum empirical bias estimate $\widehat{Bias}_n^{max}$ and the minimum empirical bias estimate $\widehat{Bias}_n^{min}$. For a given stock $\times$ month, we calculate the empirical estimation bias of the high-low measure by discounting a stock's monthly effective spread with the monthly high-low estimate. We then compute monthly averages across all stocks. Empirical estimation bias bounds are obtained following the steps described in the text. The y -axis gives the empirical bias in percentage points.

Figure 2.8: Bounding the High-Low Empirical Bias in US Stocks



Estimated bias bounds are empirically consistent for over 80% of stocks.

Notes: This figure shows stock-level average actual estimation bias of the high-low measure, as well as estimates of average maximum and minimum empirical biases. Each panel orders stocks from the largest upward estimation bias in percentage points on the left to the largest downward estimation bias to the right. Therefore, gradually moving toward the right end of the x -axis completes the US stock cross-section sorted according to the high-low bias. The top panel shows all stocks, the middle panel subsets the stock sample and plots only stocks below the median effective spread during the period of 0.21%. The bottom panel plots stocks above the median spread.

Figure 2.9: Stock-level Estimation Bias Bounds

Table 2.1: Ranking Stocks Based on Liquidity

	Effective Spread Rank				Effective Spread Portfolios			
	(1, ..., I stocks)				(1, ..., 10 deciles)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
High-low (positive estimates)	0.724				0.713			
	[70.71]				[68.76]			
High-low (zero adjustment)		0.822		0.237		0.810		0.229
		[103.72]		[13.39]		[98.14]		[13.59]
Spread implied by bias bounds			0.902	0.701			0.890	0.697
			[133.03]	[39.33]			[131.34]	[41.26]
R-squared	0.52	0.67	0.80	0.82	0.51	0.65	0.79	0.80

Notes: This table compares how well each different estimation method to compute stock-level bid-ask spreads correlates with orderings based on actual effective spreads. Columns (1) through (4) regress the effective spread rank of each stock i (1 = smallest spread, I = largest spread) on a ranking variable obtained with high-low estimates (without *ad hoc* adjustments) in (1), high-low estimates with the zero adjustment (columns (2) and (4)), and using the estimation bias bounds derived in the main text in (3) and (4). A coefficient estimate of 1 indicates that the liquidity ranking constructed with the given spread proxy perfectly predicts the actual ordering of stocks based on effective spreads. In columns (5)–(8), we sort stocks in 10 portfolios depending on their effective spreads (effective spread deciles) and regress a stock's liquidity decile on the same measure constructed using the liquidity proxies. A coefficient estimate equal to 1 indicates that the spread estimator perfectly assigns stocks to their correct liquidity portfolio. Robust standard errors are omitted in the table and we report $[t\text{-stat}]$ values.

Chapter 3

New Wine into Old Wineskins: Manipulation in Financial Markets

“’Tis a trade founded in fraud, born of deceit, and nourished by trick, cheat, wheedle, forgeries, falsehoods, and all sorts of delusions”

Daniel Defoe — “Anatomy of Exchange Alley”, 1719

3.1 Introduction

Manipulation and fraud have been part of trading since trading has existed. Stockjobbers (akin to exchange floor specialists) at London’s Exchange Alley, the 17th century precursor of the London Stock Exchange, commonly made their brokers spread false negative stories while selling stocks to attract unaware sellers. As the selling pressure built, the stockjobber would have brokers promptly waiting to buy on the market at discounted prices. During the US railroad expansion in the early 1900s, several prominent cases of wash trades led to large volatility episodes, as when a Rock Island Company official ordered dozens of their brokers to buy shares of the company to lure in other buyers and inflate the company’s share prices (Ripley (1911)).

With the rise and adoption of electronic systems by many financial markets, where quoted prices are publicly available to all market participants, predatory practices became more sophisticated. High-frequency traders developed automated strategies exploiting low latency that either actively employed or gave the appearance of manipulative behavior. Terms like quote stuffing, spoofing, and layering became the focus of regulatory policy-making and financial enforcement, especially after the triggering event of the 2010 Flash Crash was attributed to a British trader’s spoofing. While most of these practices are banned and offer legal exposure or at minimum hefty fines by regulatory agencies, detecting defined manipulative behavior and meeting the legal burden of intent are challenging empirical tasks: The SEC only prosecuted 17 market manipulation cases in 1990 and 31 in 2021.

Self-trading — an illegal market manipulation form where a trader takes both sides of the same transaction, resulting in wash trades without ownership transfer — is one of the most difficult practices to detect and prove intent, even though it appears to be common.¹ Trading firms often claim that due to

¹A 2015 Joint Staff Report studying a liquidity dry-up episode on October 15, 2015 in the treasuries market fell short of

complex trading strategies and parallel management of multiple quoting and executing algorithms, eventual self-crossing orders are inevitable. As a remedy, firms and exchanges have developed tools to prevent orders of the same market participant from resulting in executed self-trades. These tools, known as self-trading prevention (STP) functionalities, are offered by every major exchange in the world for participants on a voluntary or mandatory basis, and many brokers have internal self-trading prevention mechanisms with the same purpose.

STP safeguards prevent volume transfer between orders of the same trader by deleting resting orders involved in self-matches. This insulates traders from executing a self-trade and attracting regularity scrutiny under wash sales prohibition rules. However, aggressing orders triggering STP functionalities are not penalized and remain alive in the limit order book, either filling other traders' orders or establishing new market quotes. In practice, self-trading prevention tools prevent the occurrence of potential wash trades, but are not themselves neutral in market impact.

To illustrate this point, consider a trader with a large passive order at the best ask along smaller orders entered by other participants at the same price level. The trader then sends a market buy order which first matches with her own order. At the match moment, the self-trading tool is triggered and cancels the trader's resting order to avoid a self-trade. The aggressor order remains intact and then moves to fill the other orders, successfully establishing a new trade price.

The traditional view of the benefits to self-trading would characterize the above example in the following way. Suppose the trader's goal is to drive prices down. By posting a large passive order at a new best ask price, the trader entices other traders who observe the price-improving quote as a signal of future selling pressure to join the ask price. Having attracted sufficient volume, the trader exits her position by submitting a market order that establishes a lower market price.

This paper shows that above and beyond the incentives in physical markets, current design rules of electronic markets provide profitable opportunities for self-trading that do not result in wash sales, making the practice pervasive in futures markets. We develop a non-forensic methodology to identify self-trades that relies solely on market-by-order data. The approach leverages certain data patterns generated by the use of automated self-trading prevention tools. These functionalities commonly offered by exchanges and trading platforms prevent volume transfer by deleting resting orders involved in self-matches. However, market orders triggering self-matches are not penalized and remain alive in the limit order book. We then measure the market impact of self-trades by analyzing how they direct order flow by luring other traders to post orders directionally. Finally, we show how certain sophisticated trading strategies employ self-matches connecting unusually high levels of wash trading to the liquidity dry-up because the "non-trivial" large volumes of self-trades (5% of total volume) during that day were also found during "normal" days.

to exploit allocation algorithms used to match orders and gain market advantage over other traders.

3.2 Related Literature and Contribution

We contribute to a growing literature on market manipulation detection and its effects. Studying manipulative behavior empirically is challenging because multiple phenomena can generate similar observable patterns, as with quote stuffing which may potentially be intentional (Ye et al. (2013)) or arise in equilibrium from high-frequency trading competition (Hasbrouck (2018)). Most of the literature has exploited publicly available filings of successfully charged cases of misconduct (e.g., Aggarwal and Wu (2006), Akey et al. (2020)), since forensic-type analysis as done by Lee et al. (2013) may also suffer from many confounding factors. We provide a direct assessment of manipulative behavior by studying a sequence of events linked to an order lifespan that identify self-matches from market data. This allows us to cleanly measure market effects from manipulative behavior.

Studying self-trading in electronic markets has two advantages over analyzing other types of market misconduct. First, self-matched orders are indirectly flagged by the exchange data feed, which objectively defines the event space and does not depend on subjective criteria to define a certain practice as predatory (e.g., how many cancellations one needs to observe to infer manipulative behavior). Second, similar to quote snipping (Aquilina et al. (ming)) and strategic runs (Hasbrouck and Saar (2013)), non-accidental self-matches are accompanied by a characteristic path of quoting behavior.

3.3 Data and Setting

Definitions. We begin by differentiating between **match** and **trade** events. A match occurs when a marketable order on one side of the book matches with one or more resting orders on the other side of the LOB. In general, a match results in a trade, where the exchange sends out a trade message confirming execution to the parties involved in the match. This is followed by inventory changes in each side’s brokerage account. The distinction between match and trade in our context is important because a match may fail to execute a trade, for example, when the only orders matching are from the same trader and the self-trading prevention functionality is active.

Although exchanges use the terms self-trading and self-matching prevention interchangeably, to keep consistency in the paper, we refer to these functionalities as self-trading prevention tools, since they are triggered by a match of the same trader’s orders.

3.3.1 Regulatory Background

Wash trades have been illegal in the US since the Commodity Exchange Act of 1936. The Law specifies that any person entering into or confirming the execution (i.e., brokerage firms) of a wash trade (also referred in its text as accommodation, fictitious, or cross trade) engages in unlawful trading behavior. In compliance with these rules, exchange regulations explicitly forbid wash trading. For example, CME’s trading rulebook

RULE 534 (WASH TRADES PROHIBITED) — *No person shall place or accept buy and sell orders in the same product and expiration month, and, for a put or call option, the same strike price, where the person knows or reasonably should know that the purpose of the orders is to avoid taking a bona fide market position exposed to market risk (transactions commonly known or referred to as wash sales). Buy and sell orders for different accounts with common beneficial ownership that are entered with the intent to negate market risk or price competition shall also be deemed to violate the prohibition on wash trades. Additionally, no person shall knowingly execute or accommodate the execution of such orders by direct or indirect means.*

clearly states that both individuals engaging in self-trading and brokerage firms enabling the conduct are subject to regulatory consequences.

Sidestepping how to define the legal threshold necessary to determine whether a market participant “*knows or reasonably should know*” that orders placed would cross-trade, the necessary condition for a wash trade is that a trade match event occurs. Therefore, even if orders from the same trader were to cross — a match — as long as a trade event is not triggered, that action would not qualify even as a potential wash trade. Building off of this principle, exchanges began offering in the early 2010s order management tools that would ensure a trade event would not occur whenever two orders from the same account crossed, essentially insulating market participants from technically registering a trade after a self-match event.

3.3.2 Self-Trade Prevention Functionalities

In June 2013, the CME exchange introduced an optional functionality in its electronic trading platform, Globex, to enable brokerage firms to avoid self-trading. CME’s Self-Match Prevention (SMP) tool automatically cancels a trader’s resting order in case of a match with an aggressing order from the same trader. The functionality allows users to change the default option, canceling the incoming (newest) order instead of the resting order (oldest). The functionality is offered for free and adds no latency penalty as it filters out the aggressing order in the trading engine.

Defining “same trader”. Brokerages choose at which operational level their orders, if self-matched, would imply a potential wash trade under CME’s definition of common beneficial ownership. This loosely translates as orders submitted for the same account or accounts that have common ownership (e.g., different desks at a proprietary high-frequency trader operating in a common market). In practice, a trading firm

requests an SMP ID tag with the CME, which tags all orders submitted by the firm within or across clearing firms as belonging to the “same trader”. Because the proper use of the SMP tool is to the advantage of the brokerage, firms have incentives to narrowly define common ownership.

Self-trading prevention in other exchanges. Virtually all centralized exchanges offer self-trade prevention functionalities. These functionalities largely follow the same engine as CME’s SMP. The ICE exchange has self-trading protection services since 2013 and makes the use of the tool mandatory for proprietary traders using algorithmic trading applications. In April 2021, Eurex also made the use of SMP mandatory to proprietary algorithmic firms. Self-trading protection tools at Nasdaq and NYSE are optional. Across these exchanges, traders have the choice of cancelling the resting or aggressing order, or both, in an event of a self-match.

3.3.3 Recent Enforcement Cases

Prosecution of wash trading is relatively rare. Enforcement ability stems primarily from establishing intent (“*knows or reasonably should know*”), which is empirically difficult to show beyond reasonable doubt. Self-trading prevention tools are designed to provide insurance to trading firms against potential wash sales by preventing self-matches from turning into executed trades. By deleting resting orders when a self-match happens, these functionalities have direct market impact and can be exploited strategically. While inappropriate use of STP functionalities always insulates traders from potential wash trading, manipulation of compliance tools is still a potential violation of trading rules.

CME’s RULE 575 (DISRUPTIVE PRACTICES PROHIBITED) bans orders not entered “*in good faith for legitimate purposes*”. The rule’s text directly addresses the possibility of manipulative use of the self-trading prevention tool: “*The use of self-match prevention functionality in a manner that causes a disruption to the market may constitute a violation of Rule 575. Further, if the resting order that was cancelled was non-bona fide ab initio [to begin with], it would be considered to have been entered in violation of Rule 575*”.²

To the extent that enforcement cases of wash sales are rare, cases involving disruptive trading practices are even more scant. Since 2013, only a couple of instances of abusive use of the STP functionality were punished. In 2015, the high-frequency trading firm Allston Trading was investigated by the CFTC following the complaint of another trading firm, HTG Capital, that the firm used CME’s STP tool to spoof treasury futures. In 2018, CME issued a notice of disciplinary action against a trader that violated RULE 575.

The notice describes that the trader “*entered orders on one side of the market at the best bid/offer*

²See it here: <https://www.cmegroup.com/rulebook/files/cme-group-Rule-575.pdf>. RULE 575 broadly targets spoofing practices.

without the intent to trade. After other market participants joined his resting orders, [the trader] entered an aggressive order on the other side of the market at the same book level. [The trader] used self-match prevention software, which caused his resting orders to be cancelled within the same millisecond of entry of the aggressive order. The aggressive order then immediately traded opposite the market participant who joined or bettered [the trader’s] resting bid/offer and turned the market.” The trader had to pay about \$100,000 in fines and had his trading account suspended for a month.

3.3.4 Trading Message Data

This paper uses trading data from multiple futures markets from the Chicago Mercantile Exchange (CME). The CME represents an ideal setting to study the impact of self-trading in modern financial markets because of two reasons: *i)* matching engine handling of self-trades and *ii)* variation in microstructure features both in the cross-section and time-series.

First, CME commercializes a version of feed data packaged as Market-by-Order (MBO), which provides all information necessary to identify self-trades, as well as native and synthetic iceberg orders, every update during an order’s lifetime, and detailed information on trade events, including partial fills.³ This level of detail on market messages combined with the default option of CME’s self-trading prevention functionality — which deletes the resting order involved in a self-trade — allow us to construct a simple algorithm that completely recovers the subset of self-trades that exert market impact.⁴

The second advantage from studying CME’s futures markets is the variety of order matching algorithms used across products. In pre-electronic markets, bids and offers were matched by traders at the exchange floor. Floor brokers relied on hand signals, shouting, and paper tickets to make their quotes public, find and negotiate with opposite-side traders. Electronic limit-order books centralize this sequence of continuous bilateral engagements by adopting a set of rules that dictates how, when, and which orders are matched. Exchanges’ matching engines are disciplined by an order priority rule, which may vary across markets and receive modifications at any time. Differences in the design of these rules and changes in their implementation generate cross-sectional and time-series variation to benefits born by self-trading.

Data sample. We obtain MBO message data from the CME spanning January 2019 to June 2020 for the following outright and spread futures markets: corn, soybean, and wheat.

³Because CME’s messages are disseminated under a FIX protocol instead of ITCH or similar systems, the exchange usually offers to the public data that only shows partial book depth (either top of the book, first five or ten best price levels), and tracks markets only by level (or price) updates. These data have no comprehensive order IDs or order-level status changes, which confounds several different market phenomena. For example, trade summaries — message packets sent out to market participants detailing all matches involved in a trade — from market depth data make it impossible to properly track iceberg or implicit orders, which may result in type I or II errors when attempting to identify self-trades. In contrast, with MBO data one can track orders and trade summaries to identify orders involved in self-matches.

⁴Deletion of the aggressing order instead of the resting order results in no direct market implications, except for using the exchange’s data feed and potentially queuing other incoming orders by imposing latency externalities.

Institutional details. Our data contains message packets disseminated by the CME in FIX/FAST protocols. Messages are timestamped in nanoseconds (with microsecond precision guaranteed), where we observe two timestamp types. Transaction timestamp — generated when an incoming message hits CME’s gateway — and sending timestamp — generated when an outbound message to the data feed leaves the limit order book software system.

The gateway is the outerwall of CME’s electronic market Globex and clocks the arrival time of orders (transaction timestamp) before messages are routed to the matching engine and subsequently displayed in the market data platform (the public limit order book). Transaction timestamping is crucial to track self-trading because a trade execution message and a delete message with same transaction timestamp were hardware-clocked at the same instant. Sending timestamps may be identical because of transmission latency from the exchange, not because those events were recorded at the same time. Therefore, using transaction timestamps allow us to determine the exact timing of changes to the order book by traders and the exchange.

Trading at CME happens almost without interruption, except for an afternoon period from 1:20 PM to 7:00 PM and a brief pause in the morning during 7:45–8:30 AM. Commodity markets operate without designated market makers, while treasury futures have a number of designated firms, including Goldman Sachs, Morgan Stanley, Nomura, and Allston Trading.

3.4 Anatomy of Self-Trading in Futures Markets

3.4.1 Algorithm for Detecting Self-Trading

We begin by illustrating our detection algorithm of self-trades with market-by-order data. These are strictly flagged as self-traded and handled by STP tools. In Figure 3.2, the best ask price of \$100 has six orders of varying lot sizes. As usual in US trading data, market participant IDs are completely masked, so that the traders behind these six separate orders are unknown. An incoming buy order at \$100 of size 700 then hits the market. Panel B in the figure shows the trade summary message for the event, which tracks the history of each order’s fulfillment against the aggressing order. Note that even though the best ask level was 635-lot deep, only 160 out of the 700-depth of the market buy was filled.

Inspecting the trade summary records more closely reveals the reason for the sub-fulfillment. At the same exact timestamp, 3 out of the 6 ask orders are cancelled with a filled quantity of zero, while the other 3 orders are completely filled. The only possibility for the three no-fill deleted orders is that they were removed from the book by the self-trading tool because they were sent by the same account as the crossing buy order.

This simple detection algorithm exploits the fact that CME’s STP feature prevents traders from aggressing into their own resting orders by cancelling the trader’s passive quotes. Crucially, the marketable order remains alive in the order book. These audit-type records allow us to sidestep forensic or indirect approaches to “identify” self-trades and therefore study market manipulation.

3.4.2 Aggregate Patterns

To gain a broad understanding of how common self-trading exists in futures markets, we begin by calculating the share of resting orders that are deleted by the self-trade prevention functionality. On average, 1% of posted liquidity — across all price levels and in both sides of the book — in corn futures is eliminated from the limit order book via self-trading. This fraction is about half in soybean and wheat, 0.55% and 0.45%, respectively. When considering the fraction of trades that involve self-matches, this magnitude jumps to 7% in corn. Figure 3.3 shows the intraday distribution of self-trading frequency and size. The figure plots all self-trade events from January 2019 to June 2020 and computes the total quantity deleted from the LOB in each self-trade.

The majority of self-trades happens during the “day” session, when trading volume is also larger relative to overnight hours. Self-trades in corn are as large as 800 contracts (about \$15 million in notional value). Liquidity taken in soybean and wheat is smaller on average, as marketable orders that self-trade are between 3 and 7 contracts in these markets.

Aggressing orders that trigger self-matches are on average twice the size of those in regular trades (without self-matches). Importantly, because the aggressing order in a self-trade takes more posted liquidity than its actual size, the order flow impact of a self-trade is larger than of a regular tradable order of same size. Table 3.2 shows the magnitude of additional liquidity taken from the limit order book due to automated cancellations by the STP functionality vis-à-vis self-trades.

On average, over 33% of liquidity taken when a trade involves at least one self-match is a STP-triggered cancellation. When smaller aggressing orders are executed, this fraction is as high as 50% in corn. Most self-trades are triggered by small marketable orders, between 1 and 5 contracts. As the size of marketable orders in self-trades increases, more liquidity is taken via fulfilment. Note that this result is not necessarily mechanical — if self-traders jointly scaled resting and marketable orders, quantity slippage could remain relatively constant across aggressing size. The negative relationship shown in the table is consistent with inventory control considerations.

3.5 Quantifying the Impact of Self-Trading

3.5.1 The Economics of STP Manipulation

In general terms, self-trading with intent of manipulating transaction prices involves two sequential types of market impact. From the moment the self-trader's limit order enters the book until it is cancelled by the STP functionality, self-trading acts much like spoofing. This sends the first market signal, generating order flow impact. Above and beyond spoofing, self-trading also involves trade execution, triggered by the trader's marketable order on the opposite side of the market of her limit order. This is the second market signal, generating trading pressure impact.

The first market impact component provides order flow signal to other traders across two potential dimensions: price and size. Traders only observe demand and supply signals at existing quotes. New quotes that narrow the bid-ask spread send credible price signals since the entering trader faces execution risk alone. Size also matters — larger orders increase depth at newer or existing price levels more than smaller orders. Price and size signals increase the probability of imbalanced order flow build-up in the short-term.

When a self-trader betters the outstanding top-of-book bid (offer) quote by adding liquidity, other traders receive an updated demand (supply) price signal. Even if the self-trader only adds liquidity to an existing price level, large enough orders still contribute to perceived imbalance of order flow. When successful, bona fide liquidity from other traders joins the self-trader's price level — or further narrows the spread. In the spoofing equilibrium of Williams and Skrzypacz (2021) bona fide order flow imbalance arises as a consequence of the market maker being unable to pinpoint whether an order cancellation *before* execution reflects a sincere change in liquidity needs or spoofing. In our context, a strategic trader has no need to manually cancel a posted quote — execution risk is at minimum partially offset by the ability to self-trade, which triggers the STP functionality to cancel an order during execution. Up until a self-match, other traders have no reason to consider whether order flow from a self-trader is any different from a sincere trader.

Self-trading not only potentially generates order flow impact, but also induces trading pressure impact. To see how, imagine that a self-trader wants the price of an asset to fall. The trader posts a price-improving offer, which is soon joined by offers from other traders, generating net selling order flow. The self-trader then enters a marketable buy order, lowering the trade price of the asset and triggering the STP tool to delete the resting offer. What signal does this execution send to other traders?

In the very short-term, say a couple of milliseconds, algorithmic traders identify a buyer-initiated order with potentially significant liquidity taken off the book (actual fills and STP-cancellations), which could trigger a marketable buy order. However, in light of the order book leading up to the execution, a

different set of traders, or the same algorithmic traders in a slightly longer time horizon, may see the buy execution by the self-trader as informed price pressure. That is because the self-trader appeared to be willing to buy even as net selling order flow built-up. Processing the execution signal of the self-trade in this way could trigger marketable sell orders.

3.5.2 Market Impact

Notation. To organize the ideas we develop below, we first set some notation. We define an **event** (e) as an action by a trader that alters the composition of the limit order book. A liquidity-providing event is a new bid or ask order, denoted as *new*. A liquidity-taking event removes liquidity from the LOB either via trading (*trade*) or when the trader cancels (*cancel*) a pre-existing limit order. Let t define a regular time interval following (similarly, with $-t$ preceding) the occurrence of an event so that $\{e_\tau\}_{\tau \in t}$ timestamps the complete sequence of events for all $\tau \in t$. Further define Δe_τ as the time elapsed between two consecutive events. In the various empirical applications that follow, we choose time intervals from $t = \{500 \text{ milliseconds, } 1 \text{ second, } 10 \text{ seconds, } 30 \text{ seconds, } 1 \text{ minute}\}$. Limit orders are entered at price levels $i = 1, 2, \dots$ where $i = 1$ is the best prevailing bid or offer. We denote price improving orders with *better*.

To fix ideas, consider only the subset of events that include new orders entered in the LOB starting at time τ_0 until $\tau_{0+1} \text{ minute}$. Following the order entered just before τ_0 , new_0 , one bid establishes the new best price in the LOB within 10 seconds, one offer enters the best level after 40 seconds, and another bid joins the newly-established top of the book after 58 seconds. This gives $\{\text{bid}_{10}^{\text{better}}, \text{offer}_{40}^1, \text{bid}_{58}^1\}$ and $\{\Delta e_\tau\}_{\tau \in t} = 10, 40, \text{ and } 48 \text{ seconds}$. The computation would be identical if this was the minute before e_0 (-1 minute). More generally, the set of events we track during a time interval before or after an initial event e_0 can include liquidity providing and taking orders.

We are interested in measuring a set of statistics around the occurrence of self-trade events, and comparing the limit order book configuration after the self-trade event with prior to it. A self-trade event may correspond to the moment a limit order is entered (which will later be part of a self-match with an aggressing order from the same trader). It may alternatively correspond to the self-match itself, which possibly takes bona fide liquidity (but always takes the trader's own resting liquidity) and possibly provides liquidity in case of incomplete fills. We begin by studying the impact on the order book configuration following the entry of limit orders part of a self-trading.

Order Flow Impact

For each market and over our sample period (Jan 2019-June 2020), we calculate: time from the moment

a limit order is submitted until self-trades, the number of orders and total quantity joining the same level as the self-trader's limit order, and the share of limit orders that self-trade that are price-improving when introduced in the order book. We supplement these measures of order flow impact with a measure that captures both the magnitude of quoting and posted liquidity activity and how fast liquidity is added to the order book.

Order flow intensity. We define the volume order flow intensity $OFInt^v$ as:

$$OFInt^v(new; t) \equiv \left(\frac{\#\{bid_\tau\}_{\tau \in t}}{\sum_{\tau \in t} \Delta bid_\tau} \times t \right) \sum_{\tau} v(bid_{\tau \in t}) - \left(\frac{\#\{offer_\tau\}_{\tau \in t}}{\sum_{\tau \in t} \Delta offer_\tau} \times t \right) \sum_{\tau} v(offer_{\tau \in t}) \quad (3.1)$$

where the difference between the volume of new bids and offers is inversely weighted by the average time between consecutive bids and offers, scaled by the total even time. For example, for a sequence of $\{bid_{10}, offer_{40}, bid_{58}\}$ with volumes of 100, 20, and 50 lots, respectively, and an event time of 1 minute, $OFInt^v$ is calculated as: $60(100/29 + 50/29 - 20/40) = 280.34$. Note that larger absolute values of net order flow intensity imply a quicker succession of new orders — therefore a cluster of market activity within t — and larger sized-orders. We also compute an alternative version of order flow intensity, $OFInt$, where we track the number of bids and offers added to the order book instead of quantities.

The first column in Table 3.3 shows the frequency of self-trades that are executed within 10 microseconds, 100 microseconds, 1 millisecond, 100 milliseconds, 500 milliseconds, 1 second, 10 seconds, 30 seconds, 1 minute, and after 1 minute from book entry. These are regular event times to make cross-market and time-series comparisons easier. A quarter of self-trades happen very fast — within 100 milliseconds, with a non-trivial share of 10% executing under 1 millisecond and about 1% under 10 microseconds. Execution times approaching sub-microsecond are impressive even for high-frequency trading firms. Cutting edge hardware including field-programmable gate arrays (FPGA) are necessary to provide latency low enough so that a self-trader can respond with a marketable order after sending a limit order. This implies that not only a considerable share of self-trading is an algorithmic activity, but performed by a subset of market participants with access to latency edge.

Limit orders entered by self-traders further provide price signal — on average 9% of orders in corn markets improve bid or offer quotes (on average by 8 basis points), and as much as 40% of self-trades executed within 1 millisecond in wheat futures narrowed the spread (on average by 12 basis points). Other traders react to the entry of these limit orders — and fast. On average, 2 unique orders join the price level within 1 millisecond in corn futures, with an average of 623 contracts added up to one minute. The average notional value of a corn contract in the period is about \$18,500, which implies that over \$11.5 million in liquidity is

supplied to the same price level of the self-trader’s limit order in just one minute. Until the execution of the self-trade, on average another \$15.5 million in liquidity is supplied. The average self-trading strategy execution time in corn futures is just over 2 minutes.

Responses in the wheat and soybean markets are broadly similar to corn futures, only smaller in absolute magnitude as these order books are less deep. Nonetheless, response time of other traders is similarly fast. A non-zero average quantity of orders respond to the self-trader’s limit order entry in less than 10 microseconds, and we identify several instances in the data of sub-microsecond reaction from other traders. For reference, the human blink lasts about 100 milliseconds (100,000 microseconds). In wheat markets, 10% of the total added liquidity following the entry of the self-trading limit order happens within 100 milliseconds, which is certainly provided by algorithms.

These results speak to the ability of the spoofing component of a self-trading strategy in being followed by order flow impact. One important component of the strategy is whether the price signal given by the limit order entry leads to excessive order flow activity relative to the other side of the book. To build up flow pressure, net imbalance is crucial. Table 3.4 presents results of our order flow intensity measure from the moment the self-trader enters the limit order until the STP tool is triggered by a marketable order.

We calculate net flow intensity following bids — therefore a demand build-up — and offers separately. Relative intensities are normalized so that $|100|$ represents the maximum flow intensity. Following the entry of bids, considerable net buying order flow intensity builds-up. Traders add bids faster and with larger sizes than offers. They also enter more quotes into the bid side of the book than on the ask side. Following a self-trader’s sell limit order, the order flow intensity results flip: large negative selling pressure builds-up across all markets.

Taken together, the results above imply that considerable order flow imbalance accumulates in the direction of a strategy where the self-trader posts bids (offers) with intent to increase (decrease) the asset’s trading price.

Trading Pressure Impact

Following Chordia and Subrahmanyam (2004), we define the high-frequency version of net buying pressure NBP as:

$$NBP(trade; t) \equiv \frac{\#\{buy_{\tau}\}_{\tau \in t} - \#\{sell_{\tau}\}_{\tau \in t}}{\#\{trade_{\tau}\}_{\tau \in t}} \quad (3.2)$$

where *buy* are buyer-initiated trades (i.e., triggered by aggressing bids), *sell* are seller-initiated transactions, and $\#$ denotes the cardinality of each set. Net buying pressure is often referred as signed order flow or order imbalance, but we leave flow designations to denote events tracking additions or cancellations of non-

marketable limit orders, in particular when studying order flow impact. An alternative formulation of net buying pressure tracks traded volume instead of number of trades:

$$NBP^v(trade; t) \equiv \frac{\sum_{\tau \in t} v(buy_{\tau}) - \sum_{\tau \in t} v(sell_{\tau})}{\sum_{\tau \in t} v(trade_{\tau})}$$

where $v(e_{\tau})$ gives the volume associated with an event.

Table 3.5 shows average net buying pressure for 5 time horizons following self-trades initiated by buys and sells. Tracking the normalized number of trades or quantity traded gives identical results, therefore we focus our discussion on the *NBP* measure. Across all markets, buyer-initiated self-trades are immediately followed by intense net buying for about a second. This is consistent with a response by execution algorithms tracking price signals from the aggressor side in trades, particularly because the bulk of the buying pressure occurs within 500 milliseconds following a self-trade.

As time passes after a self-trade, trading pressure reverses — net selling pressure peaks at about 10 seconds and falls by about half after 1 minute. This reversal in trading pressure is consistent with a market correction of a fast directional trade following the buyer-initiated execution, which is not particular to self-trading (note that absolute trading pressure is highest within 1 second). But pressure reversal is also consistent with traders considering price and size signals from the order flow build-up around the self-trade. In such case, larger order flow imbalances on the offer side as a result from the self-trader's limit orders could be predictive of larger trading pressure reversals or greater net selling pressure. We test formally below.

Results of net trading pressure are flipped for seller-initiated self-trades. Up until 1 second following execution, net selling pressure dominates both the number of trades and quantity traded. Pressure reverses to net buys beginning approximately at 10 seconds, decaying in absolute value over longer time horizons. The last two columns of the table subset self-trades by tracking trading pressure sign first when a buyer-initiated self-trade is followed by net buying pressure and then when a seller-initiated self-trade is followed by net selling pressure. This directly tests for the existence of pressure reversal.

3.6 Figures and Tables

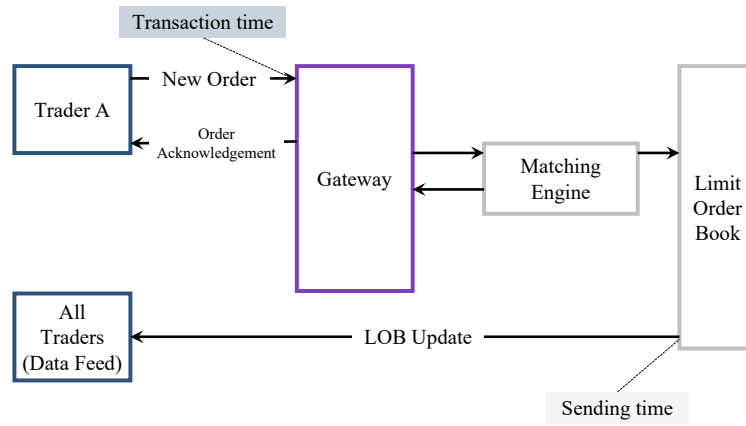


Figure 3.1: Stylized Routing System and Measurement Points of Timestamps

(a) Resting Orders at Best Ask and Offer									
Top of Book Level	\$100	Sell orders	Quantity	100	270	55	80	25	105
			Order ID	1	2	3	4	5	6
	\$98	Buy orders	Quantity	75	150	65	150	50	235
			Order ID	7	8	9	10	11	12

(b) Trade Summary After Aggressing Order							
Activity	Quantity	Orders Matched Against	Price	Filled Quantity	Order ID	Timestamp (microsecond)	
Trade	700	3	\$100	160	13	8:45:23.587663	Crossing Trade
Delete	0	1	\$100	0	1	8:45:23.587663	Self-Trade
Delete	0	1	\$100	0	2	8:45:23.587663	Self-Trade
Delete	0	1	\$100	55	3	8:45:23.587663	Filled
Delete	0	1	\$100	80	4	8:45:23.587663	Filled
Delete	0	1	\$100	25	5	8:45:23.587663	Filled
Delete	0	1	\$100	0	6	8:45:23.587663	Self-Trade

(c) Resting Orders at Best Ask and Offer After Trade									
Top of Book Level	\$100	Sell orders	Quantity						
			Order ID						
	\$100	Buy orders	Quantity						
			540						
			Order ID						
			13						
	\$98	Buy orders	Quantity						
			75	150	65	150	50	235	
			Order ID						
			7	8	9	10	11	12	

Figure 3.2: Identifying Self-Trades in Message Data

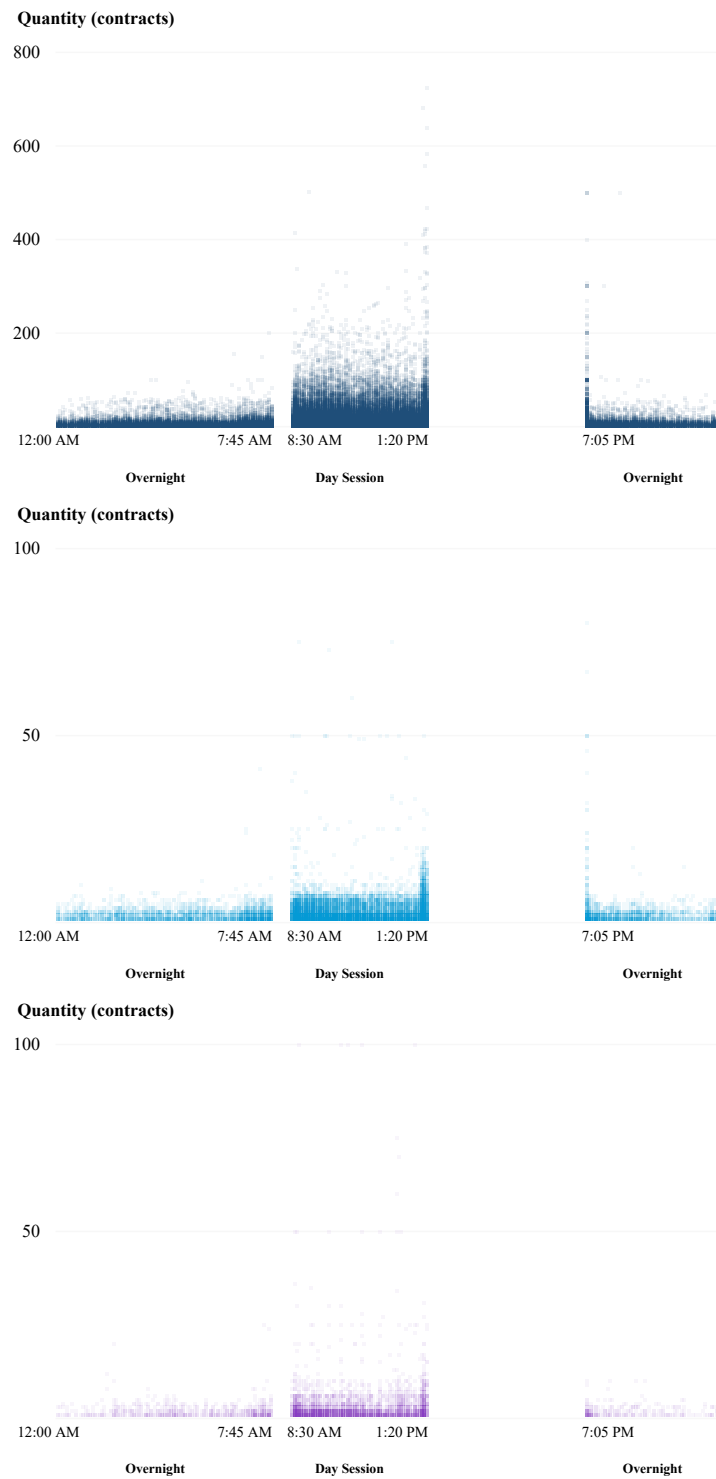


Figure 3.3: Intraday Distributon of Self-Trades

Table 3.1: Self-Trading in Futures Markets, Summary Statistics

	CORN	SOYBEAN	WHEAT
% Resting Orders $\rightarrow$ Self-Trades	1.02%	0.55%	0.45%
% Trades Including Self-Trades	7.15%	2.94%	1.51%
Trade Volume Self-Trade (contracts [\$])	16.84 [\$314,488]	7.73 [\$351,175]	3.07 [\$158,590]
Trade Volume Not Self-Trade (contracts [\$])	7.76 [\$144,916]	3.77 [\$171,254]	1.91 [\$90,231]

Table 3.2: Liquidity Taken By STP-Triggered Cancellations

AGGRESSING SIZE	% LIQUIDITY TAKEN $\left(\frac{\text{STP deleted}}{\text{STP deleted} + \text{filled}}\right)$	FREQUENCY
<i>A. Corn</i>		
All	37.93%	
Contracts (1 contract = \$289,000)		
1–5	49.78%	53.59%
5–10	35.01%	14.17%
10–20	26.72%	11.06%
20–50	18.86%	12.14%
50–100	13.09%	5.80%
100+	9.01%	3.24%
<i>B. Soybean</i>		
All	33.41%	
Contracts		
1–5	45.03%	59.00%
5–10	23.25%	16.61%
10–20	15.21%	13.29%
20–50	9.24%	9.58%
50–100	4.89%	1.43%
100+	1.87%	0.09%
<i>C. Wheat</i>		
All	37.22%	
Contracts		
1–5	44.34%	72.59%
5–10	22.01%	15.10%
10–20	17.09%	7.89%
20–50	8.94%	3.66%
50–100	4.92%	0.65%
100+	1.54%	0.11%

Table 3.3: Order Flow Impact

	EXECUTION FREQUENCY	# ORDERS JOINING PRICE LEVEL	QUANTITY JOINING PRICE LEVEL	% ORDERS WITH PRICE IMPROVEMENT
<i>A. Corn</i>				
Time From Order Entry				
10 microseconds	1.01%	0.24	3.82	11.11%
100 microseconds	2.73%	0.78	11.18	12.48%
1 millisecond	10.94%	2.12	29.54	14.16%
100 milliseconds	25.59%	6.15	81.91	12.14%
500 milliseconds	29.79%	7.81	118.75	11.67%
1 second	32.23%	9.11	144.58	11.30%
10 seconds	50.19%	18.37	331.57	9.40%
30 seconds	65.12%	26.67	496.70	8.32%
1 minute	76.94%	32.01	623.41	7.95%
> 1 minute	23.06%	25.15	843.20	6.32%
<i>B. Soybean</i>				
Time From Order Entry				
10 microseconds	0.99%	0.14	0.33	31.19%
100 microseconds	2.77%	0.43	1.15	30.36%
1 millisecond	12.99%	1.22	3.03	34.27%
100 milliseconds	27.19%	3.56	8.52	32.66%
500 milliseconds	30.13%	4.34	10.79	32.10%
1 second	32.42%	4.95	12.58	31.72%
10 seconds	50.36%	9.34	26.98	29.09%
30 seconds	61.84%	12.90	38.84	27.08%
1 minute	70.51%	15.50	47.13	26.49%
> 1 minute	29.49%	11.52	35.96	15.02%
<i>C. Wheat</i>				
Time From Order Entry				
10 microseconds	1.33%	0.14	0.32	37.78%
100 microseconds	2.99%	0.42	1.07	36.79%
1 millisecond	11.44%	1.14	2.97	40.15%
100 milliseconds	22.13%	2.65	7.05	39.66%
500 milliseconds	23.85%	3.23	8.77	38.60%
1 second	25.34%	3.68	10.03	37.76%
10 seconds	39.30%	6.92	20.39	34.46%
30 seconds	51.66%	9.80	29.82	31.64%
1 minute	60.77%	11.97	37.10	29.90%
> 1 minute	39.23%	10.44	34.66	16.06%

Table 3.4: Order Flow Intensity

	CORN	SOYBEAN	WHEAT
Bid entry $\rightarrow$ Self-Trade			
$OFInt$	0.052	3.182	3.344
$OFInt^v$	0.276	0.153	0.002
Offer entry $\rightarrow$ Self-Trade			
$OFInt$	-3.893	-0.008	-12.612
$OFInt^v$	-5.959	-14.827	-6.515

Table 3.5: Trading Pressure Following Self-Trades

Trade initiated by...	NBP^v		NBP^v		$NBP^v NBP^v(1 \text{ s.}) > 0$	$NBP^v NBP^v(1 \text{ s.}) < 0$
	Buyer	Seller	Buyer	Seller	Buyer	Seller
<i>A. Corn</i>						
0.5 second	0.545	-0.511	0.781	-0.804		
1 second	0.544	-0.505	0.775	-0.785		
10 seconds	-0.418	0.414	-0.478	0.465	-0.710	0.680
30 seconds	-0.272	0.278	-0.283	0.309	-0.550	0.549
1 minute	-0.198	0.171	-0.203	0.185	-0.455	0.404
<i>B. Soybean</i>						
0.5 second	0.499	-0.438	0.656	-0.720		
1 second	0.530	-0.513	0.692	-0.726		
10 seconds	-0.399	0.409	-0.396	0.547	-0.689	0.770
30 seconds	-0.262	0.305	-0.234	0.414	-0.455	0.650
1 minute	-0.175	0.216	-0.191	0.279	-0.401	0.505
<i>C. Wheat</i>						
0.5 second	0.505	-0.652	0.796	-0.822		
1 second	0.497	-0.678	0.757	-0.837		
10 seconds	-0.379	0.537	-0.478	0.595	-0.618	0.798
30 seconds	-0.296	0.346	-0.326	0.334	-0.482	0.564
1 minute	-0.257	0.241	-0.278	0.207	-0.432	0.426

References

- Abdi, F. and Ranaldo, A. (2017). A simple estimation of bid-ask spreads from daily close, high, and low prices. *The Review of Financial Studies*, 30(12):4437–4480.
- Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.
- Acemoglu, D. and Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30.
- Acemoglu, D. and Restrepo, P. (2020). Robots and jobs: Evidence from us labor markets. *Journal of Political Economy*, 128(6):2188–2244.
- Acharya, V. V. and Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, 77(2):375–410.
- Adrian, T., Etula, E., and Muir, T. (2014). Financial intermediaries and the cross-section of asset returns. *The Journal of Finance*, 69(6):2557–2596.
- Aggarwal, R. and Wu, G. (2006). Stock market manipulations. *The Journal of Business*, 79(4):1915–1953.
- Aghion, P., Antonin, C., Bunel, S., and Jaravel, X. (2020). What are the labor and product market effects of automation? new evidence from france. *CEPR Discussion Paper No. DP14443*.
- Akey, P., Gregoire, V., and Martineau, C. (2020). Price revelation from insider trading: evidence from hacked earnings news. *Available at SSRN 3365024*.
- Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, 57(3):1047–1091.
- Aquilina, M., Budish, E., and O’Neill, P. (Forthcoming). Quantifying the High-Frequency Trading “Arms Race”. *Quarterly Journal of Economics*.
- Aragon, G. O. and Strahan, P. E. (2012). Hedge funds as liquidity providers: Evidence from the lehman bankruptcy. *Journal of Financial Economics*, 103(3):570 – 587.
- Autor, D. H., Levy, F., and Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *Quarterly Journal of Economics*, 118(4):1279–1333.
- Barardehi, Y. H., Bernhardt, D., and Davies, R. J. (2018). Trade-Time Measures of Liquidity. *The Review of Financial Studies*, 32(1):126–179.
- Baron, M., Brogaard, J., and Kirilenko, A. (2012). The trading profits of high frequency traders. *Unpublished manuscript*.
- Barron, K., Kung, E., and Proserpio, D. (2018). The effect of home-sharing on house prices and rents: Evidence from airbnb. *SSRN Electronic Journal*.
- Beaudry, P. and Lewis, E. (2014). Do male-female wage differentials reflect differences in the return to skill? cross-city evidence from 1980-2000. *American Economic Journal: Applied Economics*, 6(2):178–94.

- Bernstein, A., Hughson, E., and Weidenmier, M. (2019). Counterparty risk and the establishment of the new york stock exchange clearinghouse. *Journal of Political Economy*, 127(2):689–729.
- Biais, B., Foucault, T., and Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, 116(2):292 – 313.
- Bleaney, M. and Li, Z. (2015). The performance of bid-ask spread estimators under less than ideal conditions. *Studies in Economics and Finance*, 32(1):98–127.
- Bloom, N., Sadun, R., and Van Reenen, J. (2012). Americans do it better: Us multinationals and the productivity miracle. *American Economic Review*, 102(1):167–201.
- Bond, P. and Glode, V. (2014). The Labor Market for Bankers and Regulators. *Review of Financial Studies*, 27(9):2539–2579.
- Budish, E., Cramton, P., and Shim, J. (2015). The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. *Quarterly Journal of Economics*, 130(4):1547–1621.
- Budish, E., Lee, R. S., and Shim, J. J. (2019). A theory of stock exchange competition and innovation: Will the market fix the market? Working Paper 25855, National Bureau of Economic Research.
- Card, D. (1992). Using regional variation in wages to measure the effects of the federal minimum wage. *Industrial and Labor Relations Review*, 46(1):22–37.
- C  lerier, C. and Vall  e, B. (2019). Returns to Talent and the Finance Wage Premium. *Review of Financial Studies*, 32(10):4005–4040.
- Chakravarty, S. and Sarkar, A. (2003). Trading costs in three u.s. bond markets. *The Journal of Fixed Income*, 13(1):39–48.
- Chao, Y., Yao, C., and Ye, M. (2018). Why Discrete Price Fragments U.S. Stock Exchanges and Disperses Their Fee Structures. *Review of Financial Studies*, 32(3):1068–1101.
- Chodorow-Reich, G., Nenov, P. T., and Simsek, A. (2021). Stock market wealth and the real economy: A local labor market approach. *American Economic Review*, 111(5):1613–57.
- Chordia, T. and Subrahmanyam, A. (2004). Order imbalance and individual stock returns: Theory and evidence. *Journal of Financial Economics*, 72(3):485–518.
- Colliard, J.-E. and Foucault, T. (2012). Trading Fees and Efficiency in Limit Order Markets. *Review of Financial Studies*, 25(11):3389–3421.
- Corwin, S. A. and Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *The Journal of Finance*, 67(2):719–760.
- Easley, D., L  pez de Prado, M., O’Hara, M., and Zhang, Z. (2020). Microstructure in the machine age. *The Review of Financial Studies*.
- Egan, M., Matvos, G., and Seru, A. (2019). The market for financial adviser misconduct. *Journal of Political Economy*, 127(1):233–295.
- Farboodi, M., Matray, A., Veldkamp, L., and Venkateswaran, V. (2020). Where has all the data gone? Working Paper 26927, National Bureau of Economic Research.
- Feller, W. (1951). The asymptotic distribution of the range of sums of independent random variables. *The Annals of Mathematical Statistics*, 22(3):427–432.
- Fong, K. Y. L., Holden, C. W., and Trzcinka, C. A. (2017). What are the best liquidity proxies for global research? *Review of Finance*, 21(1).
- French, K. R. and Roll, R. (1986). Stock return variances. *Journal of Financial Economics*, 17(1):5 – 26.

- Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *The Journal of Business*, 53(1):67–78.
- Glode, V., Green, R. C., and Lowery, R. (2012). Financial expertise as an arms race. *The Journal of Finance*, 67(5):1723–1759.
- Glode, V. and Lowery, R. (2016). Compensating financial experts. *The Journal of Finance*, 71(6):2781–2808.
- Goyenko, R. Y., Holden, C. W., and Trzcinka, C. A. (2009). Do liquidity measures measure liquidity? *Journal of Financial Economics*, 92(2):153 – 181.
- Hagströmer, B. (2021). Bias in the effective bid-ask spread. *Journal of Financial Economics*.
- Harris, L. (1990). Statistical properties of the roll serial covariance bid/ask spread estimator. *The Journal of Finance*, 45(2):579–590.
- Hasbrouck, J. (2004). Liquidity in the futures pits: Inferring market dynamics from incomplete data. *The Journal of Financial and Quantitative Analysis*, 39(2):305–326.
- 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. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16(4):646–679. High-Frequency Trading.
- Haslag, P. H. and Ringgenberg, M. (2016). The causal impact of market fragmentation on liquidity. In *Fourth Annual Conference on Financial Market Regulation*.
- Hendershott, T. and Moulton, P. C. (2011). Automation, speed, and stock market quality: The nyse’s hybrid. *Journal of Financial Markets*, 14(4):568 – 604.
- Hershbein, B. and Kahn, L. B. (2018). Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review*, 108(7):1737–72.
- Holden, C. W. and Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *The Journal of Finance*, 69(4):1747–1785.
- Hou, K., Xue, C., and Zhang, L. (2018). Replicating Anomalies. *The Review of Financial Studies*, 33(5):2019–2133.
- Ivashina, V. and Scharfstein, D. (2010). Bank lending during the financial crisis of 2008. *Journal of Financial Economics*, 97(3):319 – 338.
- Jahan-Parvar, M. R. and Zikes, F. (2019). When do low-frequency measures really measure transaction costs?
- Jain, P. (2005). Financial market design and the equity premium: Electronic versus floor trading. *Journal of Finance*, 60(6):2955–2985.
- Karnaukh, N., Ranaldo, A., and Söderlind, P. (2015). Understanding fx liquidity. *The Review of Financial Studies*, 28(11):3073–3108.
- Kwan, A., Masulis, R., and McNish, T. H. (2015). Trading rules, competition for order flow and market fragmentation. *Journal of Financial Economics*, 115(2):330–348.
- Lee, E. J., Eom, K. S., and Park, K. S. (2013). Microstructure-based manipulation: Strategic behavior and performance of spoofing traders. *Journal of Financial Markets*, 16(2):227–252.
- Lesmond, D. A., Ogden, J. P., and Trzcinka, C. A. (1999). A new estimate of transaction costs. *The Review of Financial Studies*, 12(5):1113–1141.

- Li, X., Subrahmanyam, A., and Yang, X. (2018). Can financial innovation succeed by catering to behavioral preferences? evidence from a callable options market. *Journal of Financial Economics*, 128(1):38 – 65.
- Longstaff, F. A. (2010). The subprime credit crisis and contagion in financial markets. *Journal of Financial Economics*, 97(3):436 – 450.
- Lou, X. and Shu, T. (2017). Price impact or trading volume: Why is the amihud (2002) measure priced? *The Review of Financial Studies*, 30(12):4481–4520.
- MacKenzie, D. (2021). *Trading at the Speed of Light: How Ultrafast Algorithms Are Transforming Financial Markets*. Princeton University Press.
- Maggio, M. D., Egan, M. L., and Franzoni, F. (2019). The value of intermediation in the stock market. Working Paper 26147, National Bureau of Economic Research.
- Marshall, B. R., Nguyen, N. H., and Visaltanachoti, N. (2011). Commodity liquidity measurement and transaction costs. *Review of Financial Studies*.
- Marshall, B. R., Nguyen, N. H., and Visaltanachoti, N. (2018). Do liquidity proxies measure liquidity accurately in etfs? *Journal of International Financial Markets, Institutions and Money*, 55:94–111.
- McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.
- Menkfeld, A. J. (2016). The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics*, 8(1):1–24.
- Mian, A., Rao, K., and Sufi, A. (2013). Household Balance Sheets, Consumption, and the Economic Slump. *Quarterly Journal of Economics*, 128(4):1687–1726.
- Mihet, R. (2020). Financial innovation and the inequality gap. *Working Paper*.
- Nieto, B. (2018). Bid-ask spread estimator from high and low daily prices: Practical implementation for corporate bonds. *Journal of Empirical Finance*, 48:36 – 57.
- O’Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, 116(2):257 – 270.
- O’Hara, M. and Ye, M. (2011). Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459 – 474.
- Pagnotta, E. S. and Philippon, T. (2018). Competing on speed. *Econometrica*, 86(3):1067–1115.
- Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, 53(1):61–65.
- Patton, A. J. and Weller, B. M. (2020). What you see is not what you get: The costs of trading market anomalies. *Journal of Financial Economics*.
- Philippon, T. (2015). Has the us finance industry become less efficient? on the theory and measurement of financial intermediation. *American Economic Review*, 105(4):1408–38.
- Pontiff, J. and Singla, R. (2019). Liquidity risk? *Critical Finance Review*, 8(1-2):257–276.
- Raman, V., Robe, M. A., and Yadav, P. K. (2017). The third dimension of financialization: Electronification, intraday institutional trading, and commodity market quality. *WBS Finance Group Research Paper*.
- Reshef, A. and Philippon, T. (2012). Wages and Human Capital in the U.S. Finance Industry: 1909–2006*. *Quarterly Journal of Economics*, 127(4):1551–1609.
- Ripley, W. Z. (1911). Railway speculation. *The Quarterly Journal of Economics*, 25(2):185–215.

- Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4):1127–1139.
- Shkilko, A. and Sokolov, K. (2020). Every cloud has a silver lining: Fast trading, microwave connectivity, and trading costs. *Journal of Finance*, 75(6):2899–2927.
- Spatt, C. (2018). The new realities of market structure and liquidity: Where have we been? where are we going? In *Public Policy & Financial Economics: Essays in Honor of Professor George G Kaufman for His Lifelong Contributions to the Profession*, pages 247–263. World Scientific.
- Stoll, H. R. (2006). Electronic trading in stock markets. *Journal of Economic Perspectives*, 20(1):153–174.
- Thanassoulis, J. (2012). The case for intervening in bankers’ pay. *Journal of Finance*, 67(3):849–895.
- Thompson, S. R. and Waller, M. L. (1987). The Execution Cost of Trading in Commodity Futures Markets. *Food Research Institute Studies*, 20(2):141. – 24.
- Tse, Y. and Zabolina, T. V. (2001). Transaction costs and market quality: Open outcry versus electronic trading. *Journal of Futures Markets*, 21(8):713–735.
- Williams, B. and Skrzypacz, A. (2021). Spoofing in equilibrium.
- Ye, M., Yao, C., and Gai, J. (2013). The externalities of high frequency trading. *WBS Finance Group Research Paper*, (180).

Appendix A

Appendix for Chapter 1

A.1 Broker-Dealer Financial Information

As mandated by Section 17 of the Securities Exchange Act of 1934, Rule 17a-10(a)(1), registered broker-dealers are required to submit X-17A-5 (or Financial and Operational Combined Uniform Single (FOCUS) Reports) forms every year. These documents have statements of financial and operation condition, among other information such as net capital requirements (in compliance with Rule 15c3-1). Contrary to other financial data readily available to the public, including from forms 10-K and 8-K, until 2017, these forms are not centrally index by the SEC.

Extraction steps. To access and extract broker-dealer financial information, I develop and implement the following general steps:

1. Obtain all entity names associated with a given FINRA CRD code
2. Obtain all Central Index Key (CIK) codes for each of the entity names
3. Find under which CIK forms 13F and X-17A-5 are filed
4. Find SEC Document Control Numbers (DCNs) of every filing
5. Retrieve paper document based on DCN

An example. To illustrate the multiple steps necessary to connect financial information across different SEC and FINRA databases in practice, consider Goldman Sachs. FINRA's database lists four different entity names under Goldman Sachs' CRD code 361: i) **Goldman Sachs & Co. LLC**, ii) **Goldman Sachs Asset Management**, iii) **Goldman Sachs & Co.**, and iv) **Marcus Invest Offering of Goldman Sachs & Co. LLC**.

Each of these entities may file under different Central Index Key (CIK) codes. Often times, the same nominal entity has several CIKs for filing specific forms. For example, under the **Goldman Sachs & Co. LLC** name, there are three CIKS: 42352, 734700, and 769993. Neither 734700 nor 769993 file X-17A-5

forms, but 769993 files 13F-NTs – a 13F notice required when the entity’s holdings are reported by another manager, usually the parent or holding company. These forms list **Goldman Sachs Group Inc** as the entity filing 13Fs including the assets of **Goldman Sachs & Co. LLC**. While I do not use holdings data in this paper, this procedure would match the broker-dealer entity **Goldman Sachs & Co. LLC** to its correct reporting 13F entity, **Goldman Sachs Group Inc**. Finally, note that the parent company may include holdings from other subsidiaries or their own.

Finally, **Goldman Sachs & Co. LLC** files X-17A-5 forms under CIK 42352. With the company’s correct CIK code, I access the its history of filings at the EDGAR Entity Landing Page, which provides filing statistics for all documents submitted beginning in 2001. The SEC associates a unique document ID to each filing received — known as film number or Document Control Number (DCN). The final step of retrieving forms X-17A-5 involves extracting each film number from Goldman’s records at the Entity Landing Page, then directly download the paper copy from SEC’s archive repository. I iterate the same steps for each registered broker-dealer in the US, resulting in over 80,000 unique X-17A-5 forms since 2001.

Parsing. X-17A-5 forms are PDF scans of the paper document. Prior to 2008, most files do not have digitized text, which first requires the use of an optical character recognition (OCR) engine to extract raw text. These are detection algorithms designed to extract text from images and can be fine-tuned to convert images of statement of cash flows tables, for example, into standard text strings.

A second issue is related to the variety of reporting standards adopted by broker-dealers. Although required balance sheet lines are generally uniform, different companies provide various disaggregation levels, details, or names to each accounting line. The same variation in reporting standards involves income statements, explanatory notes, and leverage calculation. To homogenize these information and extract reported values, I use the standard version of SEC’s entry system to report the information in X-17A-5 forms and allocate a broker-dealer’s balance sheet layout to this standardized mold.

Often times, broker-dealers report only part of the exhaustive list of balance sheet variables. Alternatively, some of standard lines may not apply to a particular firm — e.g., market value of options when the firm does not trade derivatives. To guarantee the accuracy of the extracted data, I manually check 750 randomly chosen forms and impose consistency constraints, such as total assets being equal to total liabilities and no asset component being larger than the total asset value.

A.2 Appendix Figures and Table

GOLDMAN SACHS & CO. LLC
(GOLDMAN SACHS & CO. LLC, GOLDMAN SACHS ASSET MANAGEMENT ("GSAM"), GOLDMAN, SACHS & CO., MARCUS INVEST OFFERING OF GOLDMAN SACHS & CO. LLC)

CRD#: 361/SEC#: 801-16048,8-129

B

Brokerage Firm *Regulated by FINRA*
(New York district office)

IA

Investment Adviser Firm ⓘ Visit SEC Site

MAIN ADDRESS
200 WEST STREET
NEW YORK, NY 10282-2198 UNITED STATES

Relationship Summary

Figure A.1: Goldman Sachs Listed Names at BrokerCheck.com

UNITED STATES SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549
FORM 13F
FORM 13F COVER PAGE

Report for the Calendar Year or Quarter Ended: 06-30-2021

Check here if Amendment

This Amendment (Check only one.):

Amendment Number:

is a restatement.

add new holdings entries.

Institutional Investment Manager Filing this Report:

Name:

Address:

Form 13F File Number:

GOLDMAN SACHS & CO. LLC

200 WEST STREET

NEW YORK, NY 10282

028-00687

The Institutional Investment manager filing this report and the person by whom it is signed hereby represent that the person signing the report is authorized to submit it, that all information contained herein is true, correct and cor

Person Signing this Report on Behalf of Reporting Manager:

Name:

Title:

Phone:

Terry Mosher

Vice President

972-368-6370

Signature, Place, and Date of Signing:

/s/Terry Mosher

[Signature]

Dallas, TX

[City, State]

06-13-2021

[Date]

Report Type (Check only one.):

☐ 13F HOLDINGS REPORT. (Check here if all holdings of this reporting manager are reported in this report.)

☒ 13F NOTICE. (Check here if no holdings reported are in this report, and all holdings are reported by other reporting manager(s).)

☐ 13F COMBINATION REPORT. (Check here if a portion of the holdings for this reporting manager are reported in this report and a portion are reported by other reporting manager(s).)

List of Other Managers Reporting for this Manager:

[If there are no entries in this list, omit this section.]

Form 13F File Number

Name

028-04981

GOLDMAN SACHS GROUP INC.

Figure A.2: Example of Goldman Sachs & Co. LLC 13-NT

159

CONDERA SECURITIES, LLC
STATEMENTS OF FINANCIAL CONDITION

	December 31,	
	2011	2010
ASSETS		
CURRENT ASSETS		
Cash and cash equivalents	\$ 188,300	\$ 175,610
Fees receivable, net	28,134	248,451
TOTAL CURRENT ASSETS	216,434	424,061
PROPERTY AND EQUIPMENT		
Furniture and fixtures	32,104	32,104
Equipment	1,856	1,856
Computers	26,950	26,950
Software	25,082	25,082
	85,992	85,992
Less: accumulated depreciation	78,150	71,456
NET PROPERTY AND EQUIPMENT	7,842	14,536
TOTAL ASSETS	\$ 224,276	\$ 438,597

FAR HILLS GROUP, LLC

STATEMENT OF FINANCIAL CONDITION

December 31, 2015	
ASSETS	
Cash and cash equivalents	\$ 1,711,889
Fees receivable	4,429,042
Property and equipment, net	293,053
Restricted cash, security deposit	207,048
Other assets	287,498
	\$ 6,928,530

Figure A.3: Broker-Dealer Balance Sheet Examples, Assets



NYSE Trading Floor, 1986



UBS Trading Floor (NJ), 2005



NYSE Trading Floor, 2007



UBS Trading Floor (NJ), 2016

Figure A.4: Two Sides of Trading Automation: Exchanges and Broker-Dealers

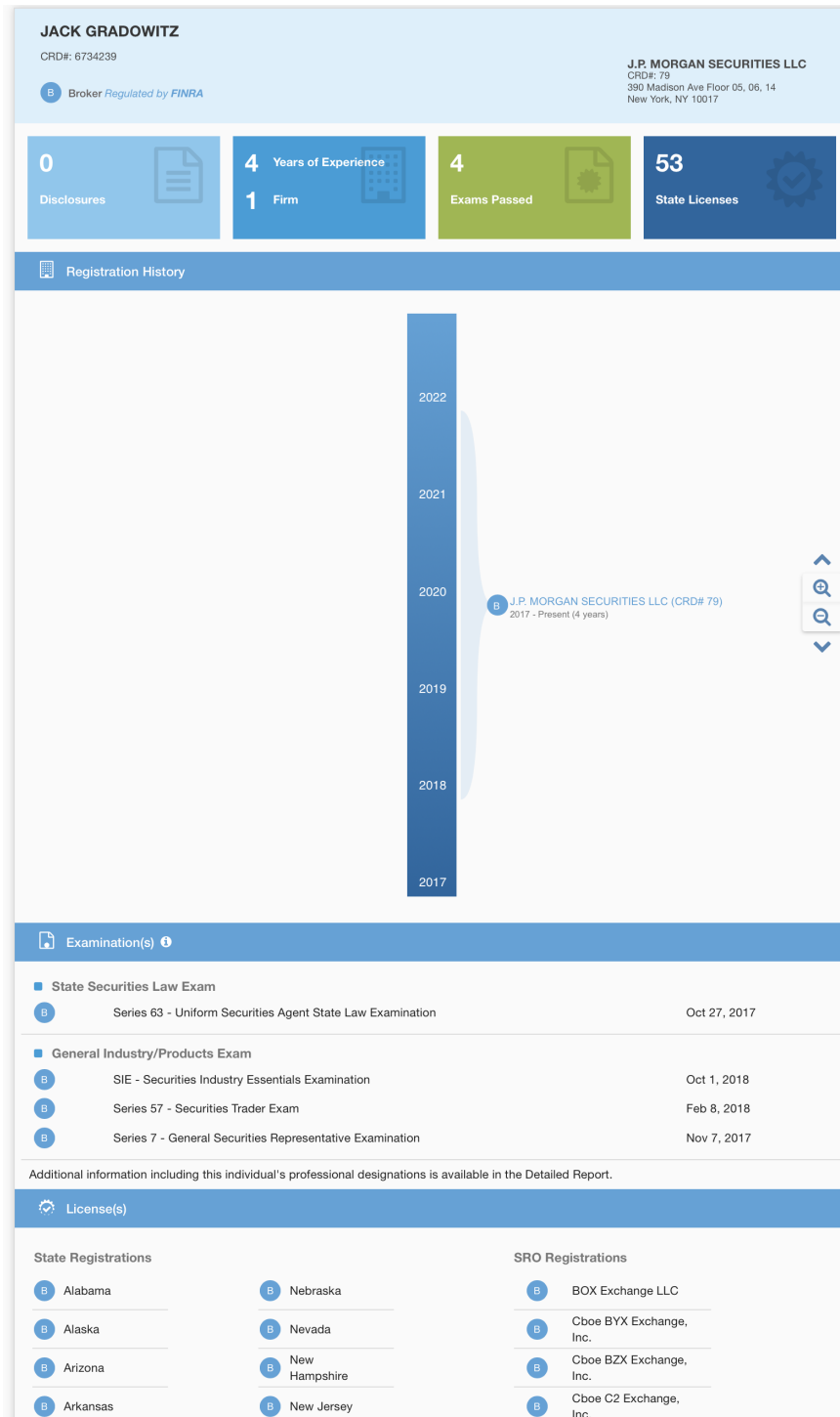


Figure A.5: Example of Stockbroker Registration Record from BrokerCheck.com

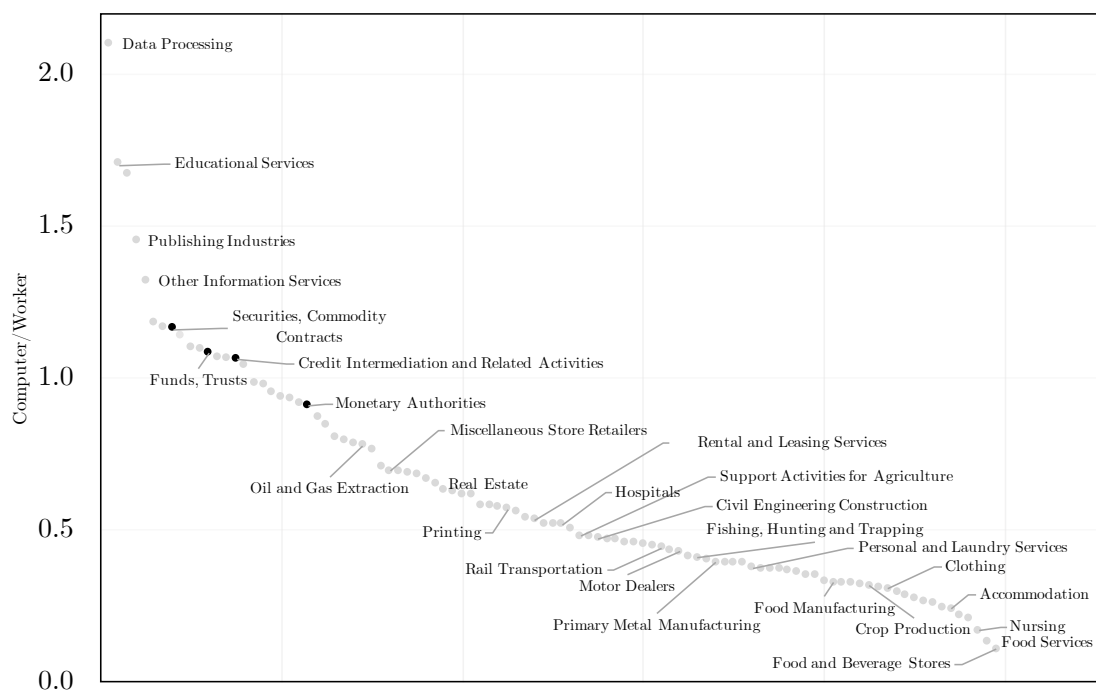


Figure A.6: IT Intensity Across Industries in 2004

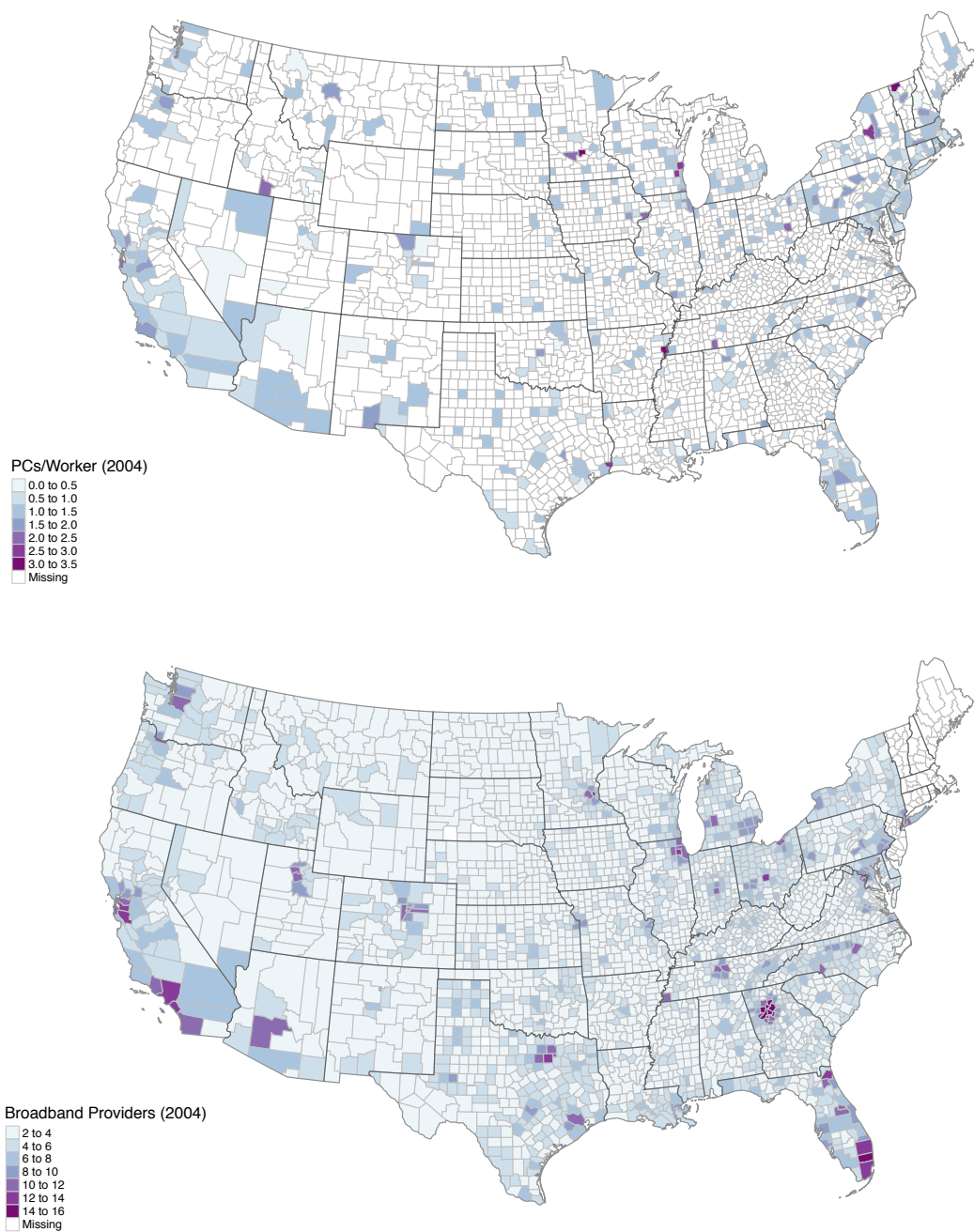


Figure A.7: Cross-Sectional Variation: IT Intensity and Broadband Providers

BB

SECURITIES COMMISSION
Washington, D.C. 20549

01 01 1018

ANNUAL AUDITED REPORT
FORM X-17A-5
PART III

FACING PAGE
Information Required of Brokers and Dealers Pursuant to Section 17 of the
Securities Exchange Act of 1934 and Rule 17a-5 Thereunder

OMB APPROVAL
OMB Number: 3235-0511
Expires: December 31, 1999
Estimated average burden
Hours per response: 2

RECEIVED
NOV 27 2001
354

SEC FILE NUMBER
8-32706

REPORT FOR THE PERIOD BEGINNING 10/01/00 AND ENDING 09/30/01
MM/DD/YY MM/DD/YY

A. REGISTRANT IDENTIFICATION

NAME OF BROKER-DEALER: NOBLE INTERNATIONAL INVESTMENTS, INC.
(Firm Name - Last, First, Middle Name)

ADDRESS OF PRINCIPAL PLACE OF BUSINESS: (Do not use P.O. Box No.)
6501 Congress Avenue, Suite 100
(No. and Street)

Boca Raton FL 33487
(City) (State) (Zip Code)

NAME AND TELEPHONE NUMBER OF PERSON TO CONTACT IN REGARD TO THIS REPORT
DENNIS J. ROSA (561) 994-1191
(Are Code - Telephone No.)

B. ACCOUNTANT IDENTIFICATION

INDEPENDENT PUBLIC ACCOUNTANT whose opinion is contained in this Report*
AHEARN, JASCO + COMPANY, P.A.
(Name - if individual, state last, first, middle name)

190 SE 19TH AVENUE POMPANO BEACH Florida 33060
(Address) (City) (State) (Zip Code)

CHECK ONE:
☒ Certified Public Accountant
☐ Public Accountant
☐ Accountant not resident in United States or any of its possessions.

FOR OFFICIAL USE ONLY

PROCESSED
MAY 20 2002
THOMSON FINANCIAL

*Claims for exemption from the requirement that the annual report be covered by the opinion of an independent public accountant must be supported by a statement of facts and circumstances relied on as the basis for the exemption. See Section 240.17a-5(e)(2).

Potential persons who are to respond to the collection of information contained in this form are not required to respond unless the form displays a currently valid OMB control number.

SEC 1410 (3-91)

Figure A.8: Face of Form X-17A-5 Submitted in Paper Format

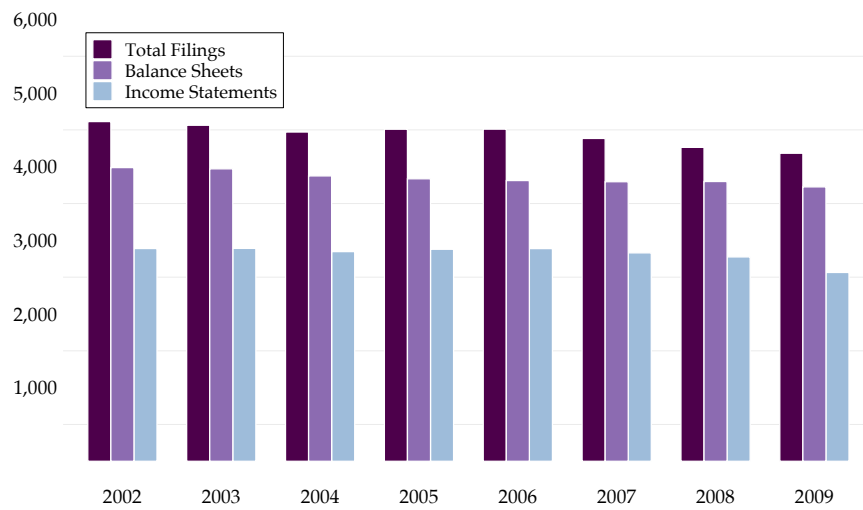
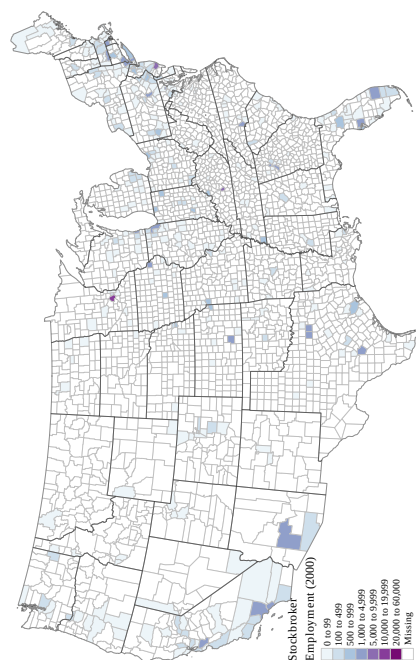
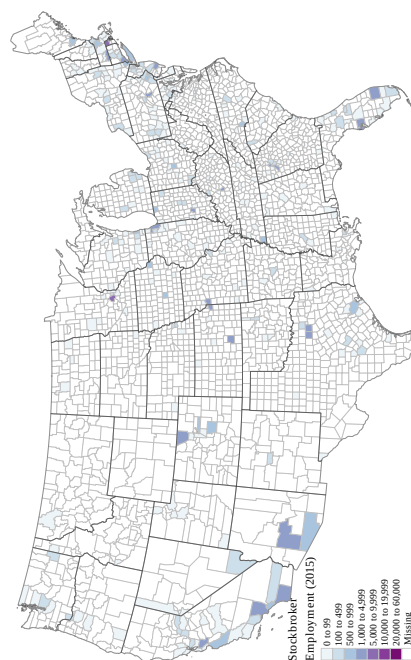


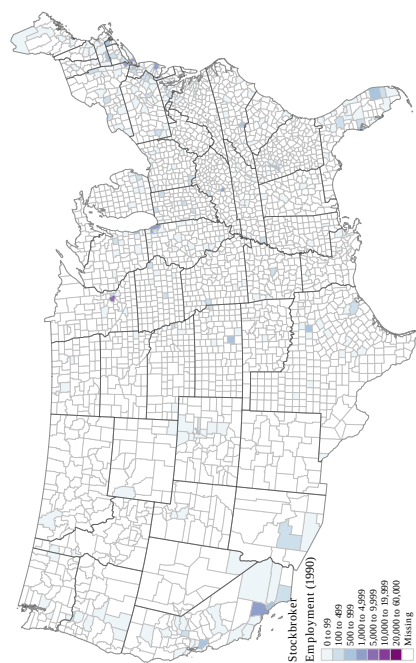
Figure A.9: Number of Form X-17A-5 Filed Every Year by Section Available



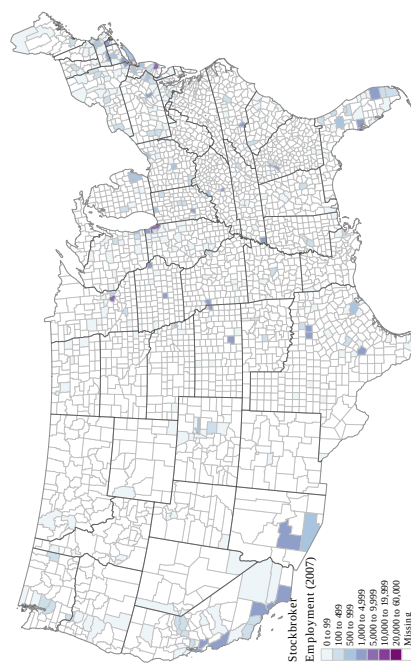
2000



2015



1990



2007

Figure A.10: Evolution of Broker-Dealer Employment: US Counties

Table A.1: Standardizing Broker-Dealer Balance Sheet, Assets

ASSETS
1. Cash
2. Cash segregated in compliance with federal and other regulations
3. Receivable from brokers or dealers and clearing organizations
3-A. Failed to deliver
3-B. Securities borrowed
3-C. Omnibus accounts
3-D. Clearing organizations
3-E Clearing organizations (deposits)
4. Receivables from customers
4-A. Securities accounts
4-B. Commodity accounts
4-C. Allowance for doubtful accounts
4-D Commissions receivable
5. Receivables from non-customers
5-A. Cash and fully secured accounts
5-B. Partly secured and unsecured accounts
5-C Registered representatives
6. Securities purchased under agreements to resell
7. Securities and spot commodities owned, at market value
7-A. Bankers acceptances, certificates of deposit and commercial paper
7-B. U.S. and Canadian government obligations
7-C. State and municipal government obligations
7-D. Corporate obligations
7-E. Stocks and warrants
7-F. Options
7-G. Arbitrage
7-H. Other securities
7-I. Sport commodities
8. Securities owned not readily marketable
9. Other investments not readily marketable
10. Securities borrowed, at market value
11. Secured demand notes
12. Memberships in exchanges
13. Investment in and receivables from affiliates, subsidiaries and associated partnerships
14. Property, furniture, equipment, leasehold improvements and rights under lease agreements
15. Other Assets
16. Prepaid Expenses
17. TOTAL ASSETS

Appendix B

Appendix for Chapter 2

B.1 Sample Handling & Construction

We briefly describe the data cleaning and computation of effective spreads from DTAQ US stock data. We closely follow the procedures in Holden and Jacobsen (2014), using Daily TAQ data. We only retain bids and asks generating bid-ask spreads within five dollars and 2.5 dollars above or below the previous midpoint. Intraday effective spreads are calculated as:

$$\frac{2|P_k - M_k|}{M_k}$$

where $M_k = (B_k + A_k) / 2$ is the midpoint of the national best bid-offer quotes of the k -th trade, given by P_k . We then compute dollar-weighted averages of the above to compute daily effective spreads used in the main paper. We then merge daily effective spreads to daily CRSP data, which includes end-of-day prices used to compute low-frequency spread estimators.

B.2 Proofs & Other Derivations

This section presents proofs and auxiliary derivations of the main theoretical results and propositions in the paper.

Useful Relationships

Simplified Form of α . Let α be a function of β and γ without explicitly considering each parameter's own arguments (observed high and low prices). The formula for α derived by Corwin and Schultz (2012) is:

$$\alpha(\beta, \gamma) = \frac{\sqrt{2\beta} - \sqrt{\beta}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}}.$$

Call $\varphi \equiv 3 - 2\sqrt{2}$, $\varphi > 0$. After rewriting the above, we have

$$\alpha = \frac{\sqrt{\varphi} [\sqrt{\beta} (\sqrt{2} - 1) - \sqrt{\gamma} \sqrt{\varphi}]}{\varphi \sqrt{\varphi}} = \frac{\sqrt{\beta} (\sqrt{2} - 1) - \sqrt{\gamma} (\sqrt{3 - 2\sqrt{2}})}{3 - 2\sqrt{2}}$$

which can be further simplified into

$$\alpha = \frac{(\sqrt{2} - 1) (\sqrt{\beta} - \sqrt{\gamma})}{3 - 2\sqrt{2}}$$

since $\sqrt{3 - 2\sqrt{2}} = \sqrt{2} - 1$. After employing a similar replacement for φ , we arrive at the simplified version of α used in the main text:

$$\alpha = \frac{(3 + 2\sqrt{2}) (\sqrt{2} - 1) (\sqrt{\beta} - \sqrt{\gamma})}{(3 + 2\sqrt{2}) (3 - 2\sqrt{2})} = (\sqrt{2} + 1) (\sqrt{\beta} - \sqrt{\gamma}).$$

Simplified Form of the High-Low Estimator. The original closed-form of the high-low spread estimator is

$$S^{HL} = 2 \left(\frac{e^\alpha - 1}{e^\alpha + 1} \right).$$

A well-known representation of the hyperbolic tangent function $\tanh(\cdot)$ is given by

$$\tanh x \equiv \frac{e^{2x} - 1}{e^{2x} + 1}$$

for which we can set $x \equiv \alpha/2$, and the simplified version of the HL estimator follows

$$S^{HL} = 2 \tanh \left(\frac{\alpha}{2} \right)$$

or, equivalently, the logistic representation below:

$$S^{HL} = 2 (L(\alpha) - 1).$$

Main Proofs

Proof of Proposition 1. Since $\mathbb{E}[R_t^*] = \sqrt{2}\mathbb{E}[R_t]$, unbiasedness of $\tilde{S}^{HL}$ immediately follows. For $\hat{S}^{HL}$ to be unbiased, it must be that $(1 + \sqrt{2})(\phi R_t^{min} - R_t^*) = (1 - \omega)S$. Expanding both sides yields

$$\mathbb{E}[\phi] \mathbb{E}[R_t] - \sqrt{2}\mathbb{E}[R_t] = \sqrt{2}S - \mathbb{E}[\phi]S$$

which holds if and only if $\phi = \sqrt{2}$, which in turn is implied by $r_t = r_{t+1}$, $\forall t$. Note that $\phi = \sqrt{2}$ results in $\omega = 1$. This completes the proof.

Proof of Proposition 2. Start with

$$\left[(\phi + \sqrt{2}\phi - 1 - \sqrt{2})S + (1 + \sqrt{2})(\phi R_t^{min} - R_t^*) \right] - S$$

which yields the following after straightforward algebra and by noting that $r_t = R_t + S$:

$$-2r_t^{min} + 2R_t^{min} - \sqrt{2}S + \phi r_t^{min} + \sqrt{2}\phi r_t^{min} - R_t^* (1 + \sqrt{2})$$

Adding and subtracting $\sqrt{2}r_t^{min}$ from the above and further algebra finally gives

$$(1 + \sqrt{2}) \left[r_t^{min} (\phi - \sqrt{2}) + (\sqrt{2}R_t^{min} - R_t^*) \right]. \quad (\text{B.1})$$

We know must show that the expression in (B.1), which measures the daily estimation error in the high-low proxy, is equivalent in expectation to the estimator's bias. First, consider the implementable high-low spread formula and its expected value:

$$\hat{S}^{HL} = (1 + \sqrt{2})(\phi r_t^{min} - r_t^*)$$

and then

$$\mathbb{E}[\hat{S}^{HL}] - S = (1 + \sqrt{2})(\mathbb{E}[\phi r_t^{min}] - \mathbb{E}[R_t^*]) - S(2 + \sqrt{2})$$

which gives the estimator's bias and is identical to the expectation of (B.1).

Proof of Proposition 3. For simplicity and without loss of generality, assume that $R_t < R_{t+1}$. Therefore,

$\partial\phi/\partial S < 0$ and thus the following holds for the bias in Proposition 2:

$$\begin{aligned} \frac{\partial (\hat{S}^{HL} - S)}{\partial S} &= (1 + \sqrt{2}) \left(R_t^{min} \frac{\partial\phi}{\partial S} + \phi + S \frac{\partial\phi}{\partial S} - \sqrt{2} \right) = \\ &= \frac{(1 + \sqrt{2}) \left(R_t^{min} \sqrt{\frac{2(R_t^{max} + S)^2}{(R_t^{min} + S)^2} + 2} + S \sqrt{\frac{2(R_t^{max} + S)^2}{(R_t^{min} + S)^2} + 2} - R_t^{max} - R_t^{min} - 2S \right)}{(R_t^{min} + S) \phi} < 0 \end{aligned}$$

since $R_t^{min} = R_t < R_{t+1} = R_t^{max}$ and $S > 0$. For the volatility result, first remember that $\phi r_t^{min} \equiv \sqrt{\hat{\beta}}$. Thus, the expected value of the bias expression is given by:

$$(1 + \sqrt{2}) \left(\mathbb{E} \left[\sqrt{(r_t^{max})^2 + (r_t^{min})^2} \right] - \sqrt{2} \mathbb{E} [r_t^{min}] \right) \leq (1 + \sqrt{2}) \left[\sqrt{8 \ln 2 \sigma^2 + 4S \sqrt{\frac{8}{\pi} \sigma} + 2S^2} - \sqrt{2} \sqrt{\frac{8}{\pi} \sigma} - \sqrt{2} S \right]$$

where the inequality holds from the Jensen's inequality. The derivative of the above with respect to σ is positive if and only if:

$$\frac{\sigma}{S} > 4.20 - \varepsilon \tag{B.2}$$

where $\varepsilon \geq 0$ is the error induced by the concavity of β . This completes the proof.

B.3 Additional Performance Evaluation in US Stocks

This section builds on Section 2.3 in the main paper. We conduct additional tests designed to assess the robustness and stability of the performance of the high-low spread estimator in US stocks. This sample is the standard setting used in horse-race type studies that evaluate the performance of multiple bid-ask spread estimators and recommend out-of-sample use — i.e., in other settings — based on in-sample performance.

Cross-section size and performance stability. We begin with a second factor that accounts for the performance of high-low spread estimates benchmarked in stock data. Our goal is to consider how the estimator's performance varies were a practitioner to observe random samples of an underlying asset class - i.e., the cross-section - with different sizes. In particular, we are interested in draws that are “small” compared to the full sample used in performance evaluation studies, which includes thousands of units (stocks). In most applications, a cross-section available to researchers includes no more than few dozen assets, as with US commodity or treasury futures. Thus, the stability of correlation estimates in samples of different sizes is

a desired feature, since inference over the performance of spread estimators in the evaluation sample implies out-of-sample extrapolation.

To do so, Figure (B.1) repeats the following experiment under three different settings. We draw 100 random samples of 10 stocks from the full DTAQ dataset, calculating the cross-section correlation between monthly HL estimates and effective spreads in each sample. We repeat the process for another 100 random draws with 100 and 1,000 stocks each time. The variability in performance in samples of 10 stocks is remarkable: the high-low estimator generates both exceptional cross-sectional correlations upwards of 80% and no correlation in other cases. Larger sample draws lead to more stable average cross-sectional correlation estimates, quickly converging to the full sample correlation of about 62%.

Liquidity heterogeneity in the sample. A robust finding for *all* moment-based bid-ask spread estimators that rely on a combination of closing, high, and low prices is a much poorer performance in the time-series compared to the cross-section. The larger the number of assets relative to the number of periods (usually months), the wider this gap. Although we are not interested in the reasons for this difference in performance, we exploit the fact that spread size heterogeneity is much higher between units than within to demonstrate the importance of heterogeneity in spread levels in the application context. This also purges the analysis from potential time-invariant unobserved factors within each stock.

First, from Figure (2.2) in the main paper, we learn that the high-low estimator performs better when effective spreads are larger. From Figure (B.1), we also know that larger cross-sections have higher and more stable correlations, suggesting that the proxy requires large variations in the spread *level* to be more accurately estimated. Mechanically, larger cross-sections tend to have more variation in spread levels than spread levels do when varying over time for the same stock, which might offer one explanation as to why performance is worse in the time series.

We test that below. Figure (B.3) plots within-stock average monthly time-series correlations between the high-low estimator and effective spreads against within-stock coefficients of variation of monthly effective spreads (standard deviation of spreads divided by the sample mean across months). The strong positive estimated slope confirms that the spread measure performs better in stocks with more dispersion in underlying effective spreads. Applying this rationale to the full cross-section generates an important result. Samples of assets with relatively homogeneous trading costs are likely to result in poor accuracy for the high low estimator.

Suggestions for performance validation. The exercises we perform in this section are simple and would greatly increase confidence in application of estimators across different markets and when benchmark-

ing performance of new spread measures. Some suggestions for implementation are the following. When the cross-section is sufficiently large, researchers should conduct a systematic analysis of the performance of the estimator in subsamples with different levels of liquidity.

Most preferably, newly-developed estimators should address how the measure may suffer from systemic bias with respect to the latent spread, and empirically analyze the extent to which such dependence may exist. They should also carefully report the stability of correlation coefficients in random subsamples from the full cross-section. This exercise confronts the fact that often times researchers in empirical finance may only have partial or sample data available. In such “small sample” applications, proxies with large variability of estimated spreads may yield somewhat spurious performance benchmarks, incorrectly concluding that the measure performs adequately in that context or market.

B.4 Negative Spread Estimates: Determinants & Empirical Use

Why do high-low estimates turn out negative? Negative high-low spread estimates are inconsistent with theory. Like many of other simple spread estimators, negative (or indeterminate) daily computed spreads with the high-low estimator are pervasive. The standard approach is to consider these point estimates as “meaningless”, either replacing them with zeros or discarding them. This implies that in many instances nearly half of the estimating sample is excluded from the analysis and with that, potential relevant information.

Negative spread estimates are not “random” or purely the result of model assumption failures. This is clear in Figure (B.2), which replicates the ideal simulated data we use in the main paper and computes the average frequency of negative high-low estimates by each effective spread bin. For small spreads the frequency of spread estimates that turn out negative is as high as 40%, even without any idiosyncratic source of bias. Based on that, we build on the following simple insight. If the intensity of negative estimates is a result of any aspect in the underlying data, we may use how often negative estimates arise to learn about the latent price process.

To fix ideas, we begin with mechanical relationships that account for generated negative spread estimates, but that offer no insight on mechanisms. We do that next. The high-low estimator depends upon both the first and second-day ranges, r_t and r_{t+1} , and the two-day range r_t^* . Small perturbations at price boundaries have small effects on daily range values, but may alter the domain of the parameter γ and lead to considerable swings in the HL point-estimate. Let $t \in \{1, 2\}$. If changes in the second-day high, for example, are sufficiently large so that h_2 becomes higher than the previous day high, the mapping of γ shifts from h_1

to h_2 , and the relationship between β and γ is updated. From the expression for α in the main paper or in Appendix (B.2), it is clear that the high-low estimator is positively defined only when $\beta > \gamma$.

Negative spread estimates do not arise only if the volatility-proportionality assumption in the high-low estimator is violated. Consider the following example. Suppose the first-day range is $r_1 = 0.01$, which is obtained when $(h_1, l_1) = (350, 345)$, and the second-day range is $r_2 = 0.27$, for $(h_2, l_2) = (401, 305)$. Clearly, the second day is much more volatile than the previous day, but the HL estimate is $0.09\% > 0$ over the two days. The simple explanation is that the square of the two-day range cannot be larger than itself added to the square of the first day range: $\gamma = r_2^2 < \beta = r_1^2 + r_2^2$. This happens because r_1 is enclosed in the second-day range. Note that it follows from this reasoning that the HL estimator can only be negative when the parameter γ combines boundaries from both consecutive days (i.e., either $h_1 - l_2$ or $h_2 - l_1$).

However, this is not sufficient for negativity. Consider the example on the left-hand side of Figure (B.4). With a partial overlap between r_t and r_{t+1} , the high-low estimator is not necessarily negative. Particularly, it is straightforward to show that some combination of high and low prices will produce a positive HL estimate when $l_{t+1} \in (l_t, h_t)$ and $h_{t+1} > h_t$. When we decrease the spread S by half (right-hand side panel), the new relative position of the second-day low, $l_{t+1} > h_t$, implies that the HL estimate is negative.

To organize these ideas, we derive two equivalent conditions that determine when the high-low spread is negative. While the first condition in Proposition 4 is very intuitive and arises directly from the functional form of the implementable high-low measure, the second condition will be useful for us to prove the driving factors of negative estimates, in similar fashion to our treatment of the bias.

Proposition 4. The high-low spread estimator is negative if and only if the following equivalent conditions hold:

$$\phi r_t^{min} < r_t^* \quad \text{or} \quad |\eta_{t+1} - \eta_t| > \sqrt{r_t^2 + r_{t+1}^2} - \left(\frac{r_t + r_{t+1}}{2} \right)$$

where η_t is the (log) mid-range, $\eta_t \equiv \frac{h_t + l_t}{2}$, and r_t is the log range on day t .

Proof. The range of the high-low spread estimator $S(\alpha)$ only contains negative values when S is evaluated at negative values of α . Hence, negativity of α defines non-positive spread estimates. The parameter α is expressed as $\alpha = (1 + \sqrt{2})(\sqrt{\beta} - \sqrt{\gamma})$, which implies that $\gamma > \beta$ determines when $\alpha < 0$. The parameter β always maps each day's range onto $\mathbb{R}_+$. Since γ includes extreme-valued functions, it may take the following values: $\gamma = r_t^2$, $\gamma = r_{t+1}^2$, $\gamma = (h_t - l_{t+1})^2$, or $\gamma = (h_{t+1} - l_t)^2$. In the first two cases, β is always greater

than γ . For $\gamma = (h_{t+1} - l_t)^2$, we can write $\gamma > \beta$ as:

$$\underbrace{h_{t+1} - \frac{r_{t+1}}{2}}_{\eta_{t+1}} - \underbrace{\left(\frac{r_t}{2} + l_t\right)}_{\eta_t} + \frac{r_t}{2} + \frac{r_{t+1}}{2} > \sqrt{\beta} \quad (\text{B.3})$$

and similarly, for $\gamma = (h_t - l_{t+1})^2$:

$$\underbrace{h_t - \frac{r_t}{2}}_{\eta_t} - \underbrace{\left(\frac{r_{t+1}}{2} + l_{t+1}\right)}_{\eta_{t+1}} + \frac{r_t}{2} + \frac{r_{t+1}}{2} > \sqrt{\beta} \quad (\text{B.4})$$

so that (B.3) and (B.4) combined can be stated as:

$$|\eta_{t+1} - \eta_t| > \sqrt{\beta} - \left(\frac{r_t + r_{t+1}}{2}\right) \quad (\text{B.5})$$

Note that the right-hand side of the negativity condition is always positive, since

$$(r_t^2 - r_{t+1}^2)^2 + 2(r_t^2 + r_{t+1}^2) > 0$$

which completes the proof of the proposition.

Generality. Throughout we assumed no consecutive-day price ties and well-defined daily ranges. Here, we relax this assumption and show that any price tie yields strictly positive spread values.

Identical consecutive-day low prices Let $h_t > h_{t+1}$ (w.l.o.g.) and $l_t = l_{t+1} \equiv a$. Given the definitions for β and γ , and the negativity condition slightly restated as $\gamma \geq \beta$, we have:

$$(h_t - a)^2 \geq (h_t - a)^2 + (h_{t+1} - a)^2 \quad (\text{B.6})$$

which cannot hold since $h_{t+1} > a$.

Identical consecutive-day high prices Consider when $h_t = h_{t+1} = b$ and $l_t < l_{t+1}$ (w.l.o.g.). The negativity conditions becomes:

$$(b - l_t)^2 \geq (b - l_t)^2 + (b - l_{t+1})^2 \quad (\text{B.7})$$

which is unfeasible since $b > l_{t+1}$.

Identical consecutive-day low and high prices Now, let $l_t = l_{t+1} = a$ and $h_t = h_{t+1} = b$. It is straightforward to see that $\gamma = 2\beta$, $\beta > 0$. Therefore, the spread is always strictly positive. $\square$

Using Proposition 4, we now show that the negativity condition is more easily attained when the underlying spread is smaller or volatility is higher.

Proposition 5. Let $f \equiv f(\eta_t, \eta_{t+1})$ and $g \equiv g(r_t, r_{t+1})$ be functions that define the left-hand and right-hand sides, respectively, of the negativity condition for the high-low spread estimator $\hat{S}^{HL}$:

$$f(\eta_t, \eta_{t+1}) > g(r_t, r_{t+1})$$

For *ex-ante* changes in the bid-ask spread level S and variance σ^2 , the following relationships hold:

$$\frac{\partial f}{\partial S} = 0, \quad \frac{\partial g}{\partial S} > 0, \quad \frac{\partial E[f]}{\partial \sigma^2} = (2 - 2 \ln 2), \quad \frac{\partial E[g]}{\partial \sigma^2} > 0, \quad \text{and} \quad \frac{\partial^2 E[g]}{\partial \sigma^2 \partial S} > 0$$

Proof. Because the efficient mid-range is identical to the observed mid-range, the left-hand side of the inequality does not depend on the spread level. That is, volatility estimated with squared returns of mid-prices is independent of the spread level. Thus $\partial f / \partial S = 0$ follows. On the other hand, the right-hand side of the inequality is a function of the observed range, $g(r_t, r_{t+1})$, which depends on the spread. If we rewrite observed ranges in terms of efficient ranges, the modified function g depends on S in the following way:

$$g \equiv \sqrt{(\mathcal{R}_t + S)^2 + (\mathcal{R}_{t+1} + S)^2} - \left(\frac{\mathcal{R}_t + \mathcal{R}_{t+1} + 2S}{2} \right)$$

and therefore

$$\frac{\partial g}{\partial S} = \frac{(\mathcal{R}_t + S) + (\mathcal{R}_{t+1} + S)}{\sqrt{(\mathcal{R}_t + S)^2 + (\mathcal{R}_{t+1} + S)^2}} - 1 > 0$$

for well-defined efficient ranges. That implies that the right-hand side in the negativity condition increases in the spread; hence as the spread widens, $\sqrt{\beta}$ relatively exceeds $(r_t + r_{t+1})/2$, so that the inequality becomes more difficult to be attained. For relatively greater spreads, the HL estimator is more likely to be positive.

The second part of Proposition B.4.2 is also simple. Denote $r_{t+1} = \kappa r_t$, $\kappa > 0$, but not necessarily bounded by 1. After squaring both sides of the negativity condition with the proper substitutions for r_{t+1} ,

we have

$$(\eta_{t+1} - \eta_t)^2 > \frac{1}{4} \left(\kappa - 2\sqrt{\kappa^2 + 1} + 1 \right)^2 r_t^2$$

for which we can replace the observed range with the efficient range. To simplify the calculations, assume $\bar{\kappa} = 1$. Under the maintained hypothesis of constant volatility, daily observed range values will be very close, and κ will differ from the unity when the spread is very small. The average value of κ in 210,000 days of simulated data when $\sigma = 3\%$ and with an infinitesimal spread is 1.09. The simplification is therefore empirically consistent with historical stock data. With $\bar{\kappa} = 1$, the term multiplying r_t^2 collapses to $(3 - 2\sqrt{2})$ and after taking expectations of both sides, we arrive at

$$E \left[(\eta_{t+1} - \eta_t)^2 \right] > (3 - 2\sqrt{2}) E [\mathcal{R}_t^2] + 2S(3 - 2\sqrt{2}) E [\mathcal{R}_t] + (3 - 2\sqrt{2}) S^2.$$

Direct substitutions for the moments above using (2.4) and (B.13) yield

$$\left(2 - \frac{k_2}{2} \right) \sigma^2 > (3 - 2\sqrt{2}) k_2 \sigma^2 + 2(3 - 2\sqrt{2}) k_1 S \sqrt{\sigma^2} + (3 - 2\sqrt{2}) S^2. \quad (\text{B.8})$$

Since the left-hand side of (B.8) is now $E[f]$, we have $\partial E[f] / \partial \sigma^2 > 0$. Similarly, the right-hand side is given by $E[g]$; hence $\partial E[g] / \partial \sigma^2 > 0$ follows. The growth rate of $E[g]$ in σ^2 may exceed $2 - k_2/2$, which largely depends on the magnitude of S . Since $\partial^2 E[g] / \partial \sigma^2 \partial S > 0$, smaller spreads will reduce $E[g]$ even while volatility increases, so that the net effect of volatility depends on the relative size between S and σ^2 . This completes the proof of the proposition.

Varying κ . Without setting $\kappa = 1$, we can replace $r_t = \mathcal{R}_t + S$ in $(\eta_{t+1} - \eta_t)^2 > \frac{1}{4} r_t^2 [(\kappa + 1) - 2\sqrt{1 + \kappa^2}]^2$ and manipulate it as:

$$(\eta_{t+1} - \eta_t)^2 > \mathcal{R}_t^2 \Psi(\kappa) + 2\mathcal{R}_t S \Psi(\kappa) + S^2 \Psi(\kappa) \quad (\text{B.9})$$

where $\Psi(\kappa) \equiv \frac{1}{4} (\kappa - 2\sqrt{\kappa^2 + 1} + 1)^2$. We can then take expectations of both sides and substitute $E[(\eta_{t+1} - \eta_t)^2] \approx 0.61\sigma^2$, $E[\mathcal{R}_t^2] = 4 \ln 2 \sigma^2$ and $E[\mathcal{R}_t] = (\sqrt{8/\pi})\sigma$:

$$0.61\sigma^2 > \sigma^2 \Psi(\kappa) 4 \ln 2 + 2S \Psi(\kappa) \left(\sqrt{\frac{8}{\pi}} \right) \sqrt{\sigma^2} + \Psi(\kappa) S^2. \quad (\text{B.10})$$

We assume $\bar{\kappa} = 1$ to make (B.10) more tractable. For $\kappa \in [0.8, 1.2]$, the function $\Psi(\kappa)$ varies from 0.15

to 0.21, with $\Psi(1) = 0.17$. After plugging 0.17 into the expression above we have

$$0.61\sigma^2 > 0.17 \left(2.77\sigma^2 + 3.19S\sqrt{\sigma^2} + S^2 \right) \quad (\text{B.11})$$

which clearly yields the first-order derivatives shown in the proposition. $\square$

We let two arbitrary functions f and g represent the inequality condition for the HL estimator to make the proposition more intuitive. The left-hand side f contains only mid-range prices, η , while the right-hand side g contains only ranges r . Mid-range prices are independent of the spread, since $\eta_t \equiv (h_t + l_t)/2 = (H_t + L_t)/2$. The greater f is compared to g , the easier the negativity condition is attained. In turn, decreasing the spread makes it easier for the negativity condition to hold and therefore generate a negative estimate.

While greater volatility also increases the frequency of negative estimates, even if volatility approached zero, a small spread would still generate a large number of non-positive spreads. That is because increasing volatility has an ambiguous effect on the proportion of negative estimates, as $E[g]$ may or may not exceed $E[f]$. Thus, S regulates the relative contribution of σ^2 in $E[g]$. This implies that ultimately the spread is more important than volatility to determine when the high-low estimator is negative.

To see this point formally, because S acts as a limiting factor to the growth of $E[g]$ with respect to σ^2 , we should expect a high proportion of negative spreads for a narrow spread, regardless of the volatility level. In the limit, as $S \rightarrow 0$, the right-hand side of (B.8) converges to $(3 - 2\sqrt{2})k_2\sigma^2$, which is the minimum of $E[g]$ with respect to S and smaller than $2 - k_2/2$. In the theoretical case, the expected frequency of negative spreads is 100%. In practice, either with simulated or actual data that can only be observed at a discrete set of times, the proportion of negative estimates is maximized when the underlying asset is very liquid. In Figure (B.5), when the true spread is 0.1%, increasing volatility from 1% to 5% produces almost no impact in the proportion of negative HL estimates.

Observable differences in the frequency of negative estimates resulting from varying σ^2 start to appear as the spread grows. Locally widening the spread enables volatility to contribute toward the growth rate of $E[g]$. The difference $E[f] - E[g]$ increases in σ^2 up to a spread level where $E[g]$ is greater than $E[f]$. At this point, the frequency of negative estimates is minimized, and different levels of volatility are again redundant to (B.8).

Figure (B.5) illustrates the proposition's prediction, where the frequency of negative estimates in the ideal simulated setting for each spread level is plotted using three different price volatility levels. The

monotonicity given in Proposition 5 clearly holds up. While the negative share with a daily volatility of 1% decays faster than shares at higher volatility levels, when spreads are around the median value of US stocks, the frequency of negatives is about the same, around 40%.

B.5 Supplementary Analysis: Close-High-Low Estimator

This section replicates most of our main results for the high-low and roll estimators using the close-high-low (CHL) measure from Abdi and Ranaldo (2017). We also consider this effective spread proxy because in the original paper, Abdi and Ranaldo (2017) show results, especially in the cross-section, that suggest a slight edge of the measure over the high-low estimator. Unfortunately, Figures (B.8) and (B.11) show that the estimator suffers the same significant decreases in performance in the cross-section of monthly, annual, and changes in effective spreads as the high-low proxy. Performance in the entire cross-section is significantly driven by few large-spread stocks, with correlation between the measure and the benchmark spread decreasing as the effective spread narrows. Indeed, the patterns are remarkably similar to the Roll and high-low estimators.

Although we do not offer formal proofs for the systematic biases we study in this paper - the moment and small sample bias - we show that the empirical and simulated results implied by our previous derivations apply almost identically to the close-high-low estimator. Furthermore, we present a complete analysis of the estimator's negativity determinants, which reveal the same driving forces behind degenerate estimates in this case as with the high-low proxy. Those are smaller spreads and higher volatility, with spreads contributing more significantly to generating negative spreads. Simulation results show that the bias behaves in the same fashion as with the high-low estimator. All in all, our results confirm the incidence of the same misbehavior patterns and driving forces for the CHL measure as in the HL and Roll estimators.

The estimator. The close-high-low spread estimator from Abdi and Ranaldo (2017) combines the use of high and low prices from the HL measure with daily closing prices from Roll (1984). In a sense, the measure uses the Parkinson-Garman-Klauss framework more broadly by integrating the range and daily returns to estimate bid-ask spreads along the lines of HL. The intuition for augmenting the information set is that closing prices are more “contaminated” by the bid-ask bounce than the range (Alizadeh et al. (2002)). Let the mid-range, or average price on day t , be defined as

$$\eta_t \equiv \frac{h_t + l_t}{2} = \frac{r_t}{2} + l_t \quad (\text{B.12})$$

which clearly coincides with the mid-range of daily efficient extreme prices, $\eta_t = (\mathcal{H}_t + \mathcal{L}_t)/2$. A crucial result in Abdi and Rinaldo (2017) establishes how the variance of the mid-range returns relates to the efficient price volatility σ^2 :

$$E \left[(\eta_{t+1} - \eta_t)^2 \right] = \left(2 - \frac{k_2}{2} \right) \sigma^2 \quad (\text{B.13})$$

where the variance may be replaced with $\hat{\sigma}_\eta^2$ for estimated squared returns of consecutive-day mid-range prices. Under the validity of (B.13), and because the average of consecutive mid-ranges,

$$S^{CHL} = 2\sqrt{(c_t - \eta_t)(c_t - \eta_{t+1})}. \quad (\text{B.14})$$

Negativity of the close-high-low estimator. We follow the same steps for the close-high-low estimator to prove the drivers of negative estimates as we did with the high-low proxy. Even though the CHL measure does not directly depend on the range, it does depend on the mid-range. Moreover, our simulation results show a behavior dynamic of CHL similar to HL with respect to negative spreads, true spread size, and bias Figure (B.7). This suggests a common channel through which the spread-to-volatility ratio affects both proxies. From the expression for CHL_t , the straightforward negativity condition of the close-high-low estimator is:

$$c_t \in (\eta_t, \eta_{t+1}) \quad (\text{B.15})$$

when the second-day mid-range is above the first-day mid-range, and $c_t \in (\eta_{t+1}, \eta_t)$ in the symmetric case. When the price variation across two days is large enough (measured by the difference in mid-prices), a wide range of values for c_t yields negative CHL spread estimates. We can rewrite (B.15) as (by analogy the symmetric case follows)

$$\eta_{t+1} - \eta_t > c_t - \eta_t \quad (\text{B.16})$$

and similarly to HL, define an inequality condition to investigate the effects of spread and volatility levels in generating negative spread estimates.

Proposition 6. Let $f \equiv f(\eta_t, \eta_{t+1})$ and $v \equiv v(c_t, \eta_t)$ be functions that define the left-hand and right-hand sides, respectively, of the negativity condition for the close-high-low spread estimator:

$$f(\eta_t, \eta_{t+1}) > v(c_t, \eta_t) \quad (\text{B.17})$$

For *ex-ante* changes in the bid-ask spread level S and variance σ^2 , the following relationships hold:

$$\frac{\partial f}{\partial S} = 0, \quad \frac{\partial E[v]}{\partial S} > 0, \quad \frac{\partial E[f]}{\partial \sigma^2} = k_3, \quad \frac{\partial E[v]}{\partial \sigma^2} = \frac{k_3}{2} \quad \text{and} \quad \frac{\partial^2 E[v]}{\partial \sigma^2 \partial S} = 0. \quad (\text{B.18})$$

Proof. The close-high-low spread estimator is negatively defined if the following holds:

$$\eta_{t+1} > c_t > \eta_t \quad (\text{B.19})$$

or, equivalently,

$$\eta_{t+1} - \eta_t > c_t - \eta_t > 0. \quad (\text{B.20})$$

We can replace observed mid-ranges and close prices with true values:

$$\eta_{t+1} - \eta_t > \mathcal{C}_t + q_t \frac{S}{2} - \left(\frac{\mathcal{H}_t + \mathcal{L}_t}{2} \right) \quad (\text{B.21})$$

and square both sides such that

$$(\eta_{t+1} - \eta_t)^2 > \left[\left(q_t \frac{S}{2} + \frac{\mathcal{C}_t}{2} - \frac{\mathcal{H}_t}{2} \right) + \left(\frac{\mathcal{C}_t}{2} - \frac{\mathcal{L}_t}{2} \right) \right]^2 \quad (\text{B.22})$$

and further simplify it as

$$\begin{aligned} (\eta_{t+1} - \eta_t)^2 &> \frac{1}{4} \left[q_t^2 S^2 + 2q_t S (\mathcal{C}_t - \mathcal{H}_t) + (\mathcal{C}_t - \mathcal{H}_t)^2 \right] + \\ &+ \frac{1}{2} \left[q_t S \mathcal{C}_t + \mathcal{C}_t^2 - q_t S \mathcal{L}_t + \mathcal{C}_t \mathcal{L}_t - \mathcal{C}_t \mathcal{H}_t + \mathcal{H}_t \mathcal{L}_t \right] + \frac{1}{4} (\mathcal{C}_t - \mathcal{L}_t)^2. \end{aligned}$$

After taking expectations of the above, we arrive at the intermediate expression

$$E[(\eta_{t+1} - \eta_t)^2] > \frac{S^2}{4} + \frac{1}{4} E[(\mathcal{C}_t - \mathcal{H}_t)^2] + \frac{1}{2} (E[\mathcal{C}_t \mathcal{L}_t] - E[\mathcal{C}_t \mathcal{H}_t] + E[\mathcal{H}_t \mathcal{L}_t]) + \frac{1}{4} E[(\mathcal{C}_t - \mathcal{L}_t)^2] \quad (\text{B.23})$$

which can be easily solved for with the moments given in Garman and Klass (1980). The final inequality is

$$(2 - 2 \ln 2) \sigma^2 > \frac{S^2}{4} + \frac{\sigma^2}{4} + \frac{1}{2}(1 - 2 \ln 2) \sigma^2 + \frac{\sigma^2}{4} \quad (\text{B.24})$$

and therefore:

$$(2 - 2 \ln 2) \sigma^2 > \frac{S^2}{4} + (1 - \ln 2) \sigma^2. \quad (\text{B.25})$$

□

Proposition 6 is intended to mimic the structure used in Proposition 5. The generic function f is intentionally identical to f in the HL estimator. The proof, however, is slightly different from the proposition in the high-low case. Since we do not assume the value of the trade indicator q_t in $c_t = C_t + q_t S/2$, we can only work with the expected value of v , without assuming a deterministic q_t . We also use moments derived for $E[C^a \mathcal{H}^b \mathcal{L}^c]$ in Garman and Klass (1980).

The interpretation of Proposition 6 is also very similar to Proposition 5, although simpler. Decreasing the spread size lowers the expected value of the right-hand side of the negativity condition, contributing to more negative estimates. When the spread is small, increases in σ^2 cannot induce $E[f] < E[v]$, which explains why in Figure (B.6) the proportion of negative CHL estimates is around 50% when $S = 0.1\%$, regardless of volatility size. Note that the figure shows the same monotonic patterns as with the high-low measure, revealing an even higher proportion of negative estimates in the close-high-low case for most spread sizes.

Moreover, for small spreads, any value of σ^2 increases $E[f]$ faster than $E[v]$ – a pattern that is eventually reversed when the spread is large enough so that the right-hand side becomes greater than $E[f]$. For completeness, we also provide the following alternative result that shows why smaller spreads induce negative CHL estimates more often.

Proposition 7. The probability that c_t falls in the interval (η_t, η_{t+1}) , and therefore the negativity condition for CHL is attained, increases as the spread size decreases.

Proof. The probability that the negativity condition of CHL is attained, $\Pr[\eta_t < c_t < \eta_{t+1}]$, is given by

$$\int_{\eta_t}^{\eta_{t+1}} f(c_t) dc_t = \int_{l_t}^{h_t} f(c_t) dc_t - \underbrace{\left(\int_{\eta_{t+1}}^{h_t} f(c_t) dc_t + \int_{l_t}^{\eta_t} f(c_t) dc_t \right)}_{\mathcal{K}}.$$

Let $h_t^* = h_t - \delta$ and $l_t^* = l_t + \delta$ represent modified daily high and low prices from a decrease of 2δ in the spread S . Let $\Pr [\eta_t^* < c_t < \eta_{t+1}^*]$ denote the probability that c_t falls within the negativity interval *ex post* the spread decrease. Then, it follows that

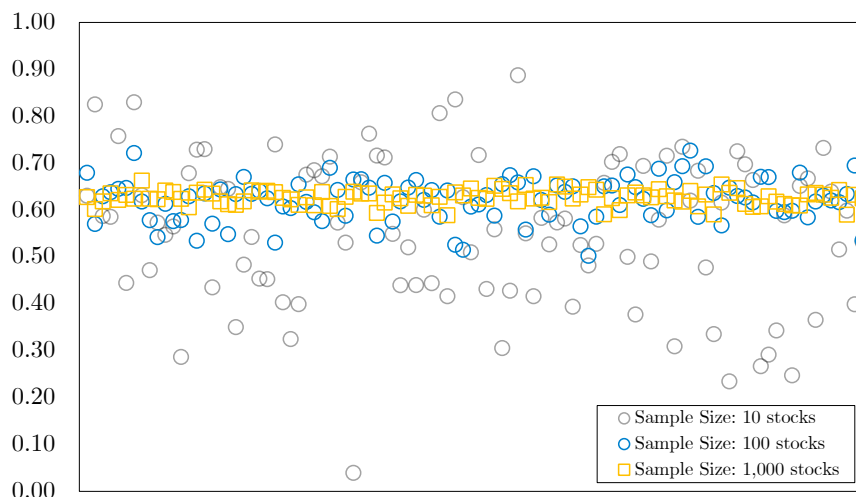
$$\left(\int_{\eta_{t+1}}^{h_t} f(c_t) dc_t + \int_{l_t}^{\eta_t} f(c_t) dc_t \right) \geq \left(\int_{\eta_{t+1}}^{h_t^*} f(c_t) dc_t + \int_{l_t^*}^{\eta_t} f(c_t) dc_t \right)$$

since

$$\begin{aligned} \left(\int_{\eta_{t+1}}^{h_t^*} f(c_t) dc_t + \int_{l_t^*}^{\eta_t} f(c_t) dc_t \right) &\leq \left(\int_{h_t - \delta}^{h_t} f(c_t) dc_t + \int_{\eta_{t+1}}^{h_t - \delta} f(c_t) dc_t + \right. \\ &\quad \left. \int_{l_t + \delta}^{\eta_t} f(c_t) dc_t + \int_{l_t}^{l_t + \delta} f(c_t) dc_t \right) \end{aligned}$$

and therefore $\int_{\eta_t}^{\eta_{t+1}} f(c_t) dc_t \leq \int_{\eta_t^*}^{\eta_{t+1}^*} f(c_t) dc_t$. The equality holds if and only if $\mathcal{K} = 0$. This can be the case only when $\eta_t = \eta_{t+1}$. Hence, $\Pr [\eta_t^* < c_t < \eta_{t+1}^*] > \Pr [\eta_t < c_t < \eta_{t+1}]$ and we conclude the proof. $\square$

B.6 Appendix Figures



A researcher with a random sample of 10 US stocks would obtain very distinct average cross-sectional correlations between monthly high-low estimates and monthly effective spreads every time a new sample would be drawn. Correlations become stable across randomly chosen samples when the stock data includes more than 100 units.

Notes: This figure shows cross-sectional correlations between the high-low spread estimator (using the zeros adjustment) and the effective spreads from 100 random samples of 10, 100, and 1,000 stocks from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

Figure B.1: Cross-Sectional Correlation: Stability Across Different Sample Sizes

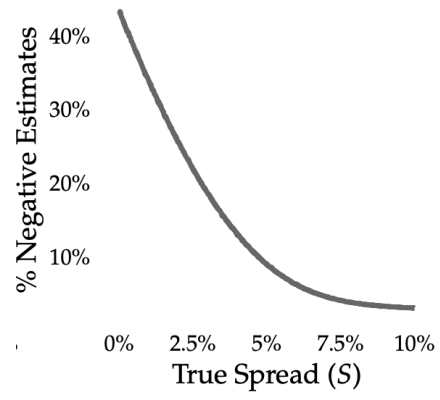
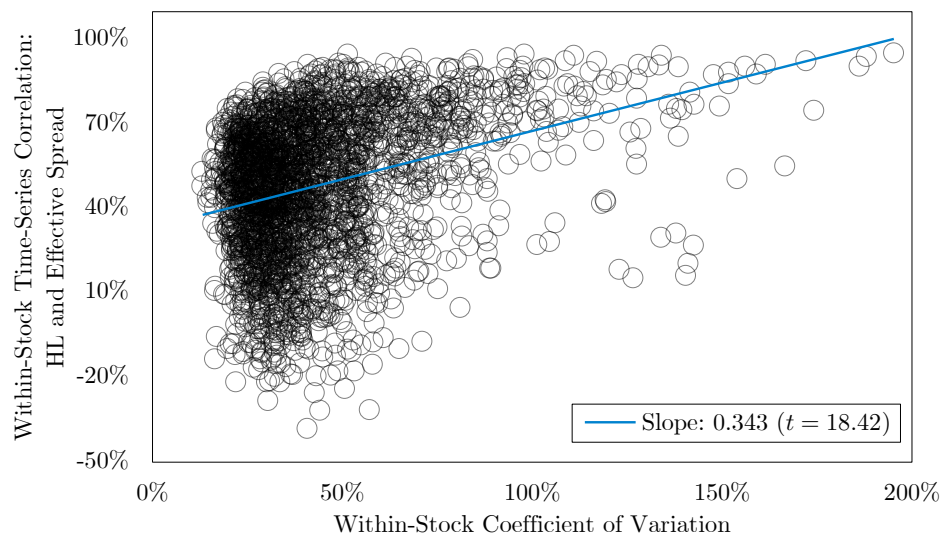


Figure B.2: High-Low Negative Estimates Under Ideal Conditions



Stocks with greater dispersion in effective spreads are associated with higher time series correlation between monthly high-low estimates and the spread.

Notes: This figure compares within-stock dispersion in the spread (sample standard deviation divided by mean) with time series correlations between the high-low spread estimator (using the zeros adjustment) and effective spreads. There are 3,398 stocks in total from DTAQ US stock data and sample construction details are described in Appendix (B.1).

Figure B.3: HL Performance in the Time Series and Dispersion of Effective Spread

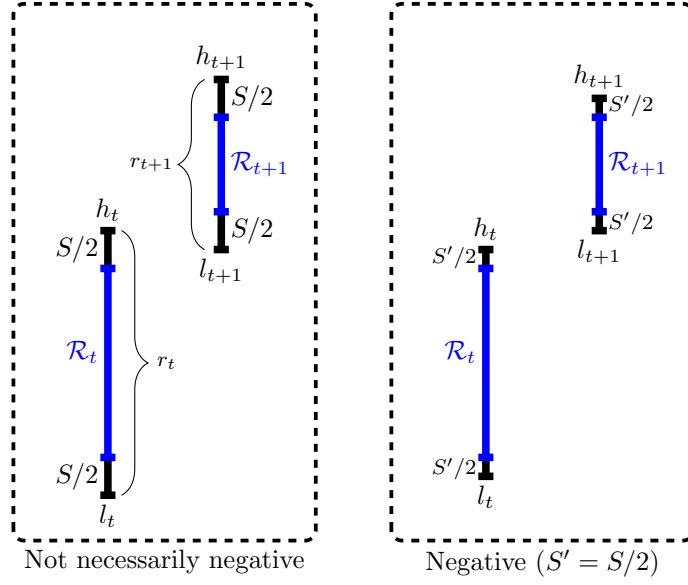
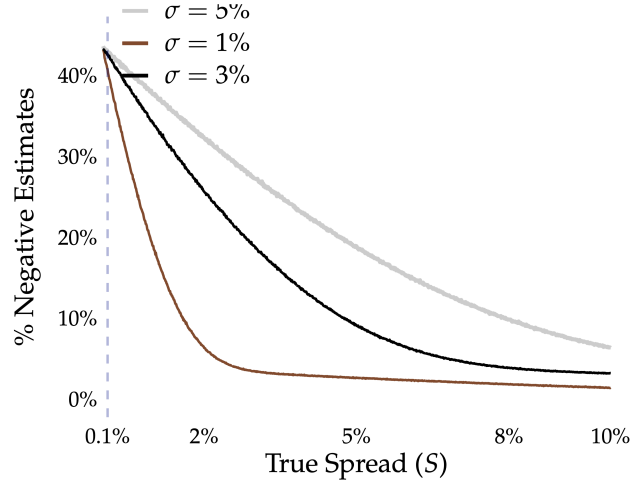


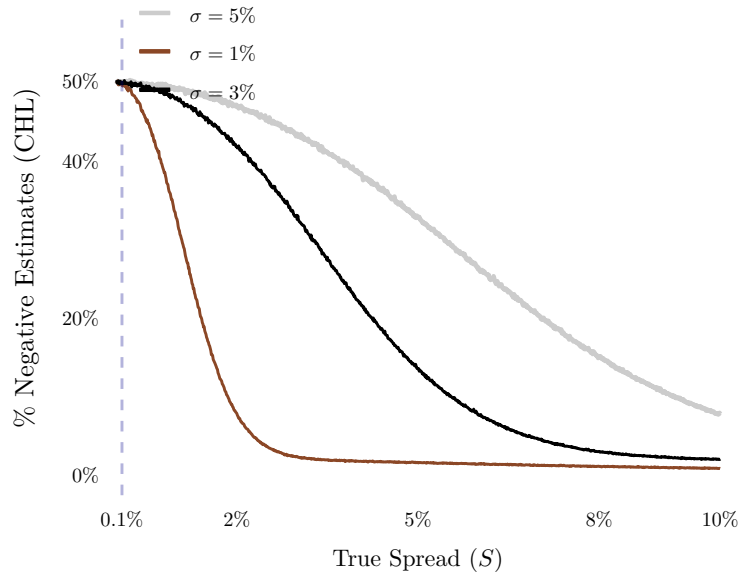
Figure B.4: Negativity Induced by the Spread Size in the HL Estimator



For very small spreads (below 0.1%), the frequency of negative spread estimates is always greater than 40%, regardless of price volatility. For reference, the median effective spread in US stocks during 2003-2015 is 0.3%.

Notes: This figure computes the average frequency of negative high-low estimates over 10,000 trading months for each 0.1-percentage-point increment in the spread between 0.1% and 10% from simulated data. The data generating process is described in detail in the text and is designed to maintain all model assumptions, including constant true spreads in each sample. We report the relationship with three different volatility values.

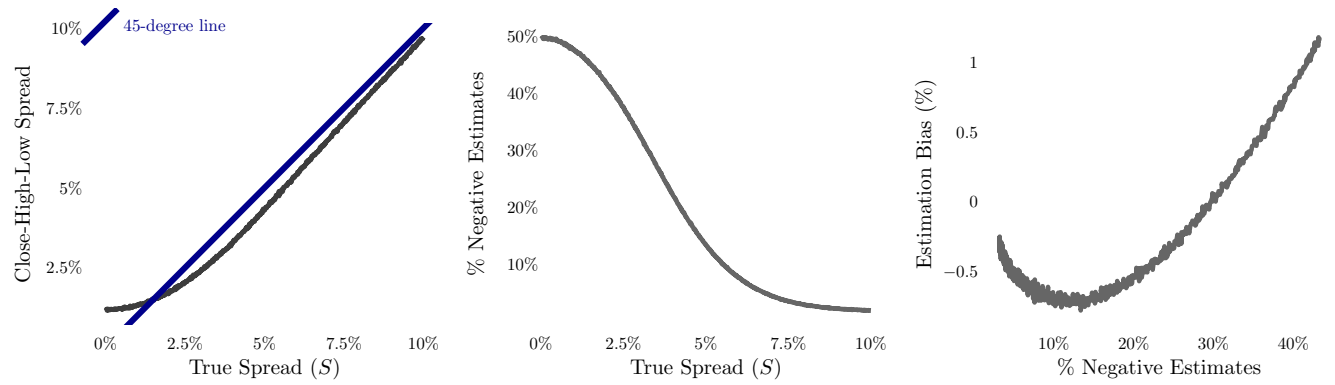
Figure B.5: Relationship Between Spread Size, % of Negative Estimates Under Ideal Conditions



For very small spreads (below 0.1%), the frequency of negative close-high-low spread estimates is always greater than 50%, regardless of price volatility. For reference, the median effective spread in US stocks during 2003-2015 is 0.3%.

Notes: This figure computes the average frequency of negative close-high-low estimates over 10,000 trading months for each 0.1-percentage-point increment in the spread between 0.1% and 10% from simulated data. The data generating process is described in detail in the text, being identical to the DGP used in the exercises for the high-low estimator. Moreover, it is also designed to maintain all model assumptions, including constant true spreads in each sample. We report the relationship with three different volatility values.

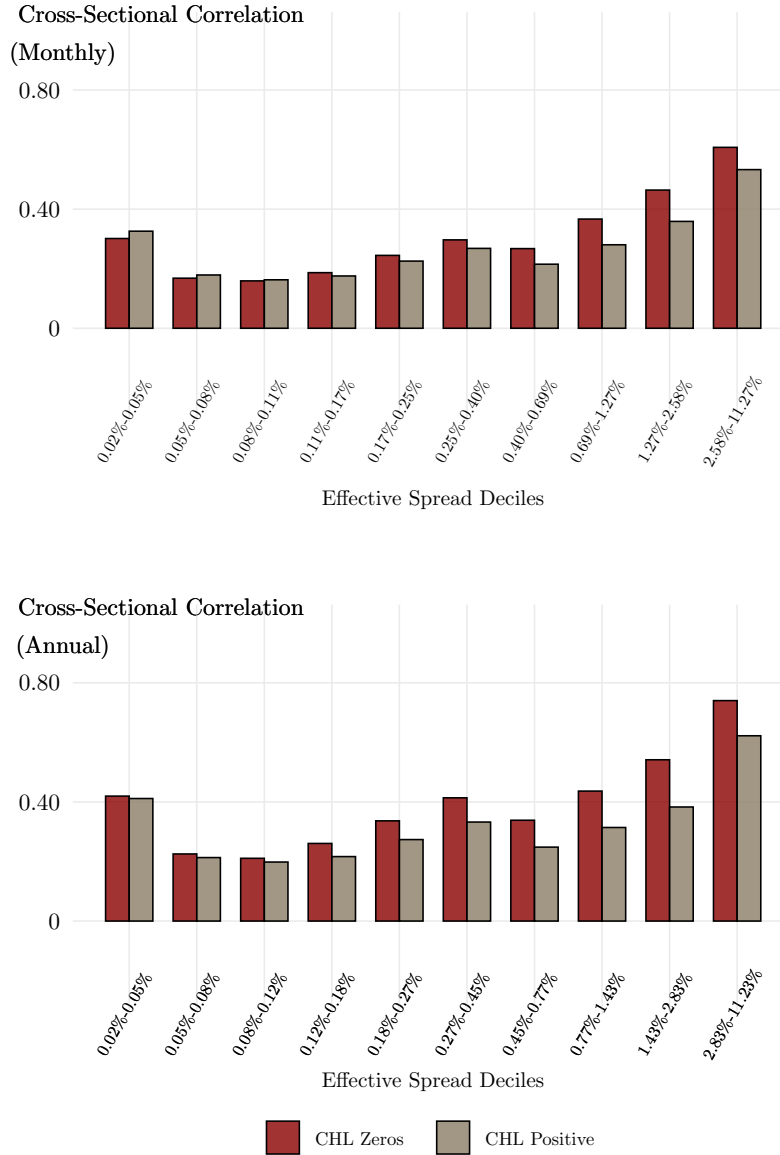
Figure B.6: Relationship Between Spread Size, % of Negative Estimates Under Ideal Conditions



For small spreads (below 1%), the close-high-low estimator suffers from an upward bias and a higher fraction of negative estimates. As the spread becomes larger, the bias decreases and fewer estimates turn out negative. For reference, the median effective spread in US stocks during 2003-2015 is 0.3%.

Notes: This figure computes 10,000 monthly averages of the close-high-low spread estimator (with zeros adjustment) and share of negative daily estimates for each 0.1-percentage-point increment in the spread between 0.1% and 10% from simulated data. The data generating process is described in detail in the text and is designed to maintain all model assumptions, including constant true spreads in each sample. Bias is defined by the difference between the estimated spread and the true spread.

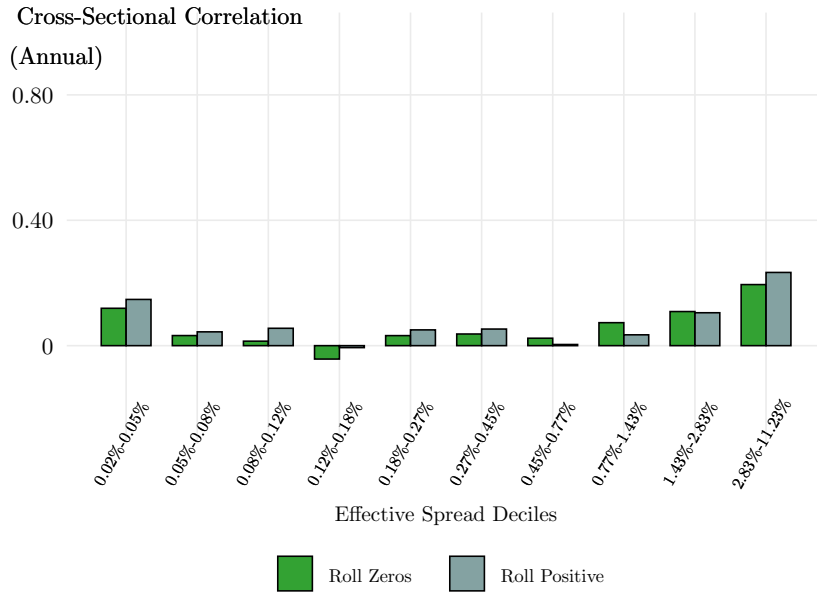
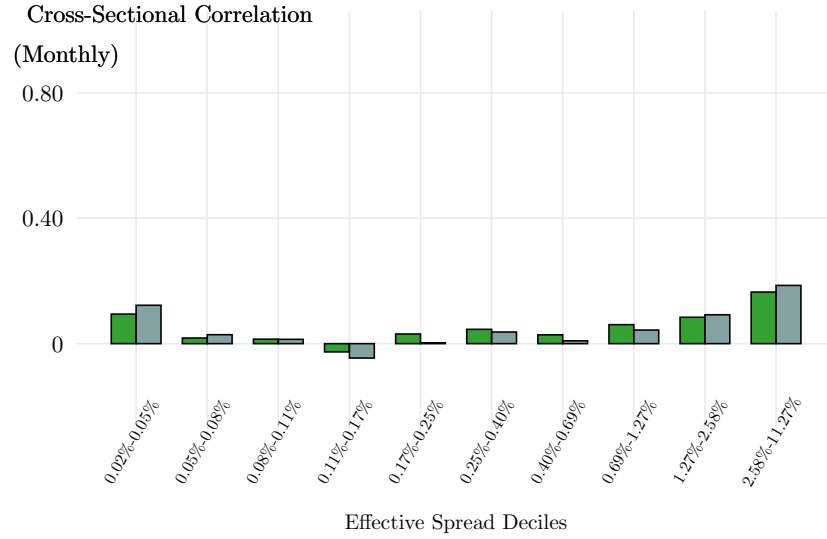
Figure B.7: Bias, Negative Estimates, and True Spread Under Ideal Conditions



In all effective spread size deciles, monthly averages of close-high-low estimates without ad hoc adjustments correlate poorly with monthly average effective spreads. Correlations from the estimates with the positives and zeros adjustments behave similarly, with the best performance in the least liquid stocks (above 50% correlation with effective spreads). The frequency of daily estimated spreads that turn out negative is above 50% in the most liquid stocks, decreasing as effective spread bands increase.

Notes: This figure shows cross-sectional correlations between three versions of the close-high-low spread estimator and effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

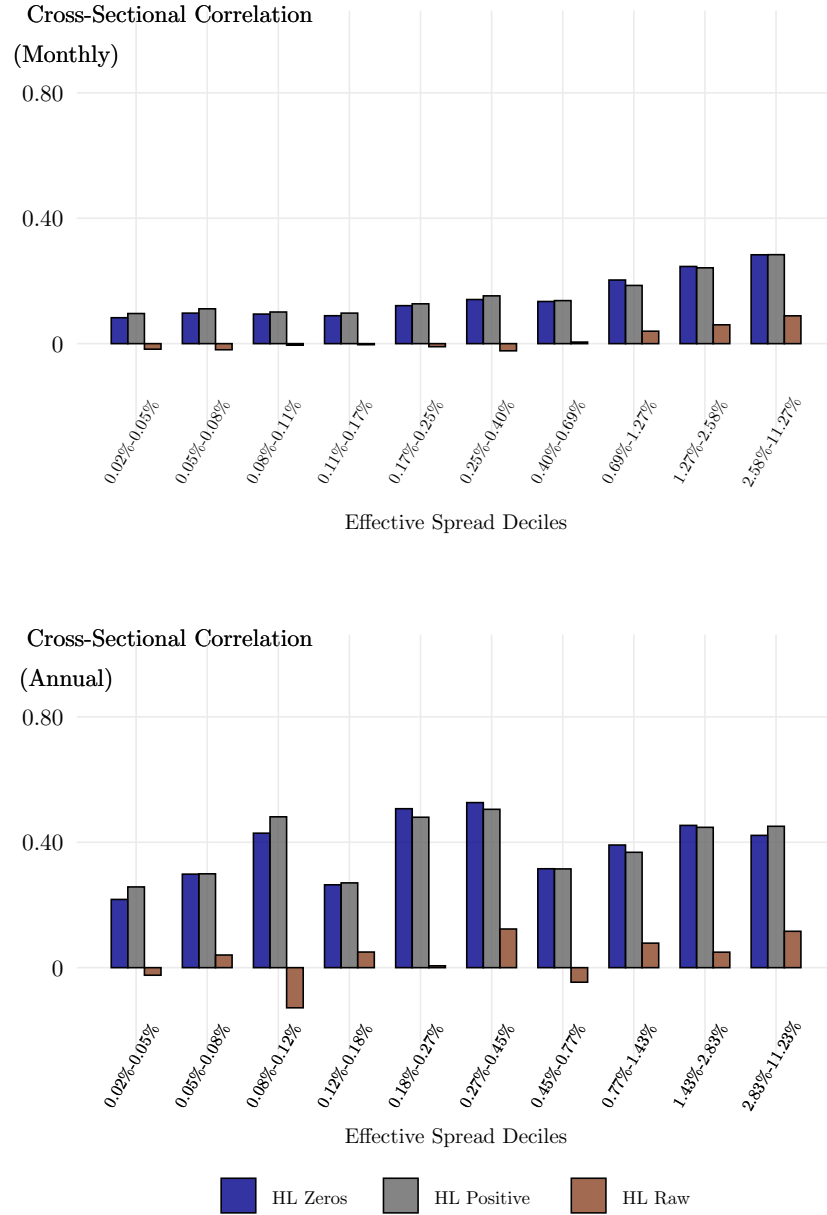
Figure B.8: Cross-Sectional Correlation: CHL Estimates and Effective Spread By Effective Spread Decile Size



In all effective spread size deciles, monthly averages of Roll estimates without ad hoc adjustments correlate poorly with monthly average effective spreads. Correlations from the estimates with the positives and zeros adjustments behave similarly, with the best performance in the least liquid stocks (above 15% correlation with effective spreads). The frequency of daily estimated spreads that turn out negative is above 40% in the most liquid stocks, decreasing as effective spread bands increase.

Notes: This figure shows cross-sectional correlations between three versions of the Roll spread estimator and effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

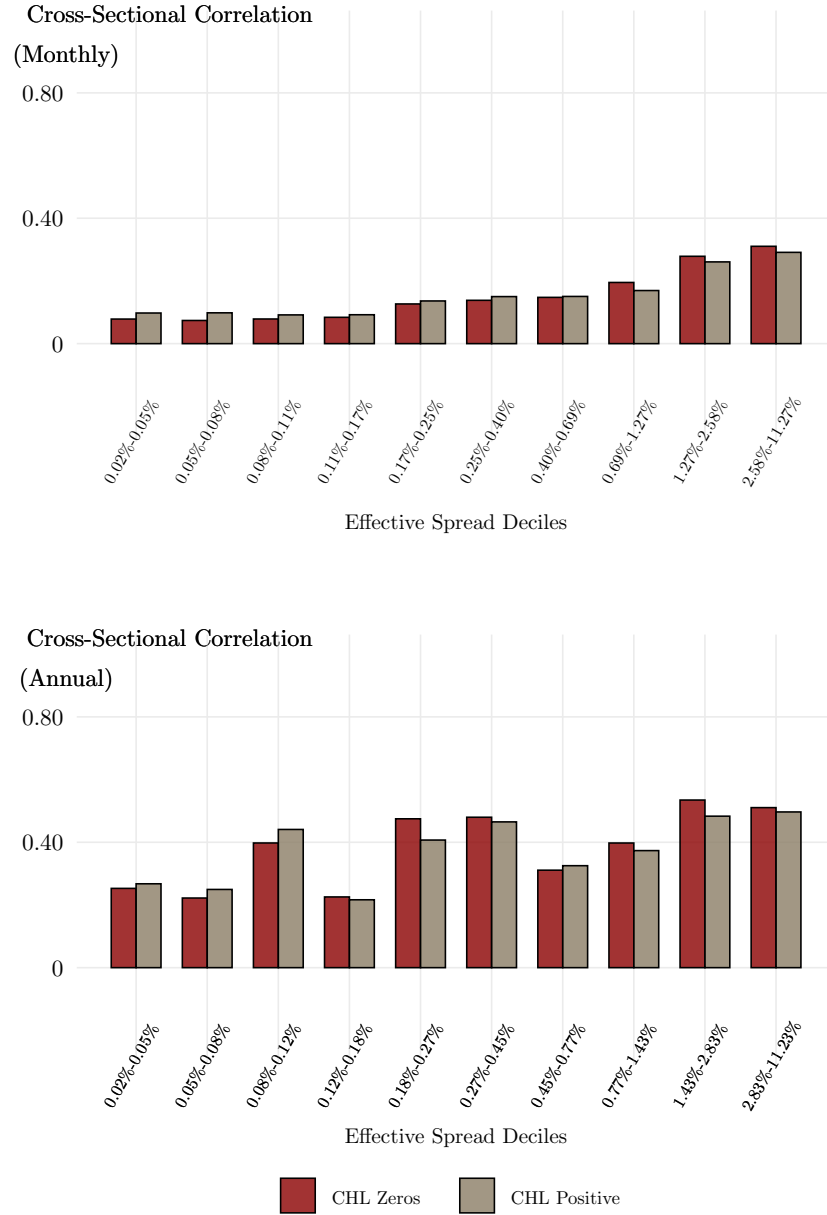
Figure B.9: Cross-Sectional Correlation: Roll Estimates and Effective Spread By Effective Spread Decile Size



In all effective spread size deciles, monthly averages of changes in high-low estimates without ad hoc adjustments correlate poorly with monthly average change in effective spreads. Correlations from the estimates with the positives and zeros adjustments behave similarly, with the best performance in the least liquid stocks (above 30% correlation with effective spreads). The frequency of daily estimated spreads that turn out negative is above 40% in the most liquid stocks, decreasing as effective spread bands increase.

Notes: This figure shows cross-sectional correlations between three versions of the close-high-low spread estimator and effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

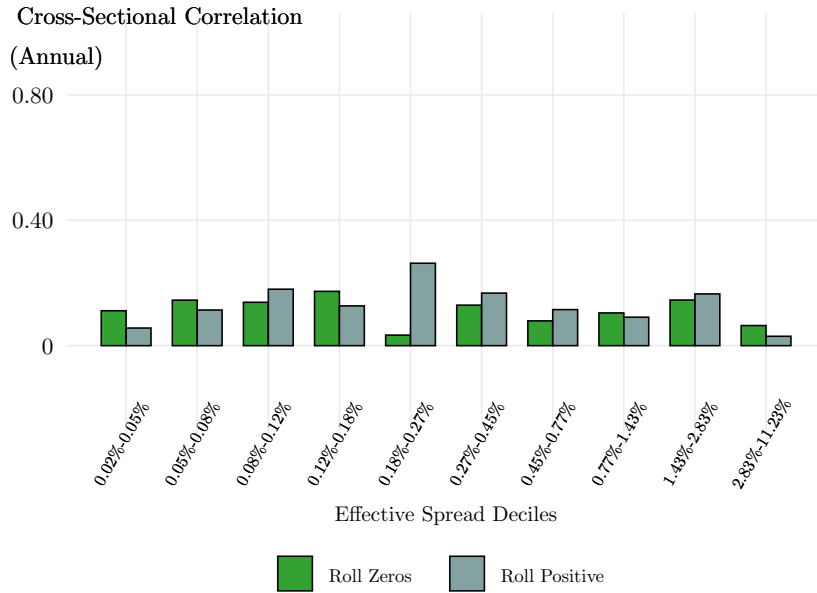
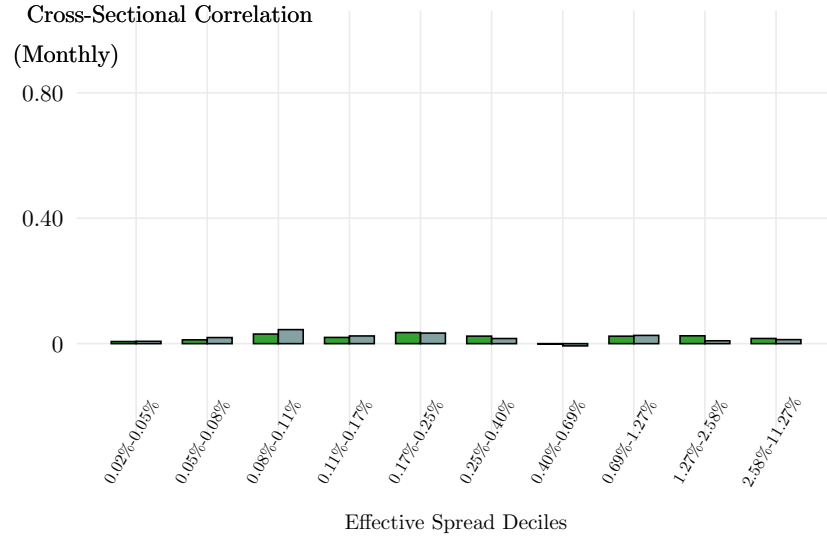
Figure B.10: Cross-Sectional Correlation: Changes in HL Estimates and Changes in Effective Spreads By Effective Spread Decile Size



In all effective spread size deciles, monthly averages of changes in close-high-low estimates without ad hoc adjustments correlate poorly with monthly changes of average effective spreads. Correlations from the estimates with the positives and zeros adjustments behave similarly, with the best performance in the least liquid stocks (above 30% correlation with effective spreads). The frequency of daily estimated spreads that turn out negative is above 50% in the most liquid stocks, decreasing as effective spread bands increase.

Notes: This figure shows cross-sectional correlations between three versions of the close-high-low spread estimator and effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

Figure B.11: Cross-Sectional Correlation: Changes in CHL Estimates and Effective Spreads By Effective Spread Decile Size



In all effective spread size deciles, monthly averages of changes in Roll estimates without ad hoc adjustments correlate poorly with monthly changes of average effective spreads. Correlations from the estimates with the positives and zeros adjustments behave similarly, with the best performance in the least liquid stocks (above 1% correlation with effective spreads). The frequency of daily estimated spreads that turn out negative is above 40% in the most liquid stocks, decreasing as effective spread bands increase.

Notes: This figure shows cross-sectional correlations between three versions of the Roll spread estimator and effective spreads from DTAQ US stock data. There are 3,398 stocks in total and sample construction details are described in Appendix (B.1).

Figure B.12: Cross-Sectional Correlation: Changes in Roll Estimates and Changes in Effective Spreads By Effective Spread Decile Size