© 2018 Xin Wang
ESSAYS ON STOCK EXCHANGES SPEED COMPETITION, DESIGNS AND HIGH-FREQUENCY TRADING

BY

XIN WANG

DISSERATION
Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics in the Graduate College of the University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:
Professor Dan Bernhardt, Chair
Assistant Professor Mao Ye
Professor Neil Pearson
Professor Charles Kahn
Abstract

The first chapter shows that a key driver of stock exchanges’ competition on order-processing speeds is the Order Protection Rule, which requires an exchange to route its customers’ orders to other exchanges with better prices. Faster exchanges attract more price-improving limit orders because the probability of being bypassed by trades with inferior prices on other exchanges is reduced. When all exchanges speed up, this probability can increase, potentially harming the welfare of investors. In contrast, increasing connection speeds between exchanges raises investor welfare by reducing this probability. Nevertheless, no exchange wants to improve connection speeds because this will reduce its trading volume. I provide empirical evidence showing that slow exchanges lose trading volume to fast exchanges as the latter attract more price-improving orders. I first show that a slow exchange’s (IEX) market share of trading volume in stocks with a five-cent tick, the minimum price movement, increases by 13 percent relative to one-cent tick stocks after the introduction of Tick Size Pilot Program in 2016, because price improving is less likely with larger tick size. I then show that after switching from a dark pool to a public exchange, IEX attracts more trading volume in stocks that are more likely to have one tick bid-ask spread as price improving is impossible with binding spread.

To reduce high-frequency trader’s speed advantage, new stock exchange designs such as frequent batch auctions and several order delay designs have been proposed to slow down trading speed and eliminate the speed arms race among high-frequency traders. In the second chapter, I investigate how newly designed exchanges with these ‘speed bump’ features would compete against traditional exchanges. I find that among order delay proposals, the most effective design is to delay only liquidity taking orders as proposed by the Chicago Stock Exchange. Frequent batch auctions are shown not to improve liquidity when the degree of private information is high enough. Moreover, when frequent batch auctions are implemented, exchanges have incen-
tives to compete on the frequency at which batch auctions take place. Finally, I show that even when sniping is a significant problem for some stocks, exchanges with large market share of total trading volume may lack incentives to implement frequent batch auctions or order delays even when these innovative designs could improve long-term investor welfare. Therefore, the interests of exchanges may not be aligned with those of long-term investors with regard to how they value designs that alleviate sniping.

In the last chapter, I incorporate discrete tick size and allow non-high-frequency traders (non-HFTs) to supply liquidity in the framework of Budish et al. (2015). When adverse selection risk is low or tick size is large, the bid-ask spread is typically below one tick, and HFTs dominate liquidity supply. In other situations, non-HFTs dominate liquidity supply by undercutting HFTs, because supplying liquidity to HFTs is always less costly than demanding liquidity from HFTs. A small tick size improves liquidity, but also leads to more mini-flash crashes. The cancellation-to-trade ratio, a popular proxy for HFTs, can have a negative correlation with HFTs’ activity.
To my parents, my wife Yan, and my children Ziqian and Zijie.
Acknowledgments

First and foremost, I would like to express my sincere gratitude to my dissertation advisors, Professors Dan Bernhardt and Mao Ye, for their continuous support of my Ph.D. study and related research, and for their infinite patience, motivation, and encouragement. Their guidance helped me throughout my research and the writing of this thesis. I am particularly grateful to Professor Dan Bernhardt for his numerous comments and suggestions on the modeling work, and his generous help in editing the early draft of the first two chapters. I am also grateful to Professor Mao Ye for his continuous encouragement. I have learned a lot from him. In particular, his optimism inspired me to pursue the research in this thesis. Without Professor Bernhardt’s and Professor Ye’s support and continuous guidance, I could not have finished this dissertation.

Besides my advisors, I would like to thank the rest of my thesis committee: Professors Neil Pearson and Charles Kahn, for their insightful comments and encouragement. Discussions with them have given me new perspectives for my research.

My sincere thanks also go to the Economics Department at the University of Illinois for the generous summer research support and travel grants. I would also like to thank the Paul W. Boltz family for their fellowship in the summer of 2016. In addition, I also sincerely thank the department’s graduate secretaries, Carol Banks and Tera Martin-Roy, who have provided me with immeasurable assistance during my Ph.D. study.

I also want to thank my fellow classmates and friends at UIUC. It has been a great journey to pursue my Ph.D. with Sen Ma, Jianfeng Xu, Zening Li, Rui Fan, Hui Xu and Chang Sun.

Last but not least, I would like to thank my family: my parents and my wife Yan for supporting me spiritually throughout my writing of this thesis. I would also like to thank my two lovely boys Ziqian and Zijie. I had many happy moments with them during my Ph.D. study. Their laughter was the driving force of my work every day.
# Table of Contents

List of Tables ................................................................. viii

List of Figures ............................................................... ix

Chapter 1  Why Do Stock Exchanges Compete on Speed, and How? .................. 1
  1.1 Introduction .................................................................. 1
  1.2 Baseline Model .......................................................... 7
  1.3 Equilibrium Analysis of Exchange Speed Competition ..................... 11
  1.4 Empirical Analysis ...................................................... 33
  1.5 Conclusions ............................................................... 36

Chapter 2  Can Stock Exchanges with Speed Bump Designs Survive? ............ 43
  2.1 Introduction .................................................................. 43
  2.2 Model Setup ................................................................ 47
  2.3 Equilibrium Analysis .................................................... 50
  2.4 Conclusions ............................................................... 58

Chapter 3  Who Provides Liquidity and, When? ........................................ 60
  3.1 Introduction .................................................................. 60
  3.2 Model .......................................................................... 64
  3.3 Benchmark: Binding at one tick under a large tick size ...................... 66
  3.4 Equilibrium types under a small tick size ..................................... 71
  3.5 Make-take spread .......................................................... 72
  3.6 Flash equilibrium versus undercutting equilibrium ......................... 73
  3.7 Stub quotes and mini-flash ............................................. 79
  3.8 Predictions and policy implications ........................................ 82
  3.9 Conclusions ............................................................... 88

References ................................................................. 98

Appendix A  Appendix for Chapter 1 ........................................... 102
  A.1 Proofs ........................................................................ 102
  A.2 Equilibrium Analysis with Exchange Speed Heterogeneity When $\gamma < \bar{\gamma}$ .... 125

Appendix B  Appendix for Chapter 2 ........................................... 129
  B.1 Proofs ........................................................................ 129
Appendix C  Appendix for Chapter 3 .................................................. 135
C.1  Proofs .................................................. 135
List of Tables

1.1 Summary Statistics for Tick Size Pilot Test .................................................. 38
1.2 Impact of Tick Size on IEX’s Market Share of Trading Volume .................... 39
1.3 Impact of Switching to Public Exchange on IEX’s Market Share of Total Trading Volume ................................................................. 40
List of Figures

1.1 Event Time line of the Baseline Model .......................... 9
1.2 Potential Trade-Through Time Window .......................... 10
1.3 Latency Map ......................................................... 41
1.4 Plot of Coefficients on Interacting Terms of Pilot Dummy with All Time Dummies Except The First Period .................. 42

2.1 Trading Stage Event Time line .................................. 49

3.1 Pricing Grid under Large vs. Small Tick Sizes ..................... 90
3.2 Depth and the Adverse Selection Risk under a Binding Tick Size ......................... 91
3.3 Bid-ask Spread Quoted by HFTs under a Small Tick Size ......................... 92
3.4 The Undercutting and the Flash Trading Equilibrium ......................... 93
3.5 States and Profits for HFT Liquidity Suppliers with the $Q^{th}$ Position on the Ask Side ......................... 94
3.6 Value of Liquidity Supply and Stale-Queue Sniping and Queue Length ......................... 95
3.7 Flash Crash Intensity ............................................. 96
3.8 States and Profits for BATs on the Ask Side ............................................. 97
Chapter 1

Why Do Stock Exchanges Compete on Speed, and How?

1.1 Introduction

Technological innovation has led U.S. stock exchanges compete aggressively on the incredible speed with which they process orders. The round-trip order-processing time today is about 50 microseconds, and stock exchanges continuously highlight new speed records. This “arms race” in processing speed is so prevalent that researchers often use the speed enhancements of stock exchanges as instruments to address such questions as the impact of high-frequency traders (Hendershott et al. (2011) and Menkveld (2013)). However, neither the drivers nor the impact of this arms race have been studied.

The lack of understanding of the origin of the speed competition among exchanges leaves room for interpretations based on anecdotal evidence or conjecture. For example, in his New York Times best-selling book, Flash Boys, Michael Lewis posits that exchanges increase speed to collude with high-frequency traders (HFTs), and that their joint forces rigged U.S. stock markets. My paper contributes to the literature and broader understanding of the issue by providing theoretical foundations for the origins and consequences of speed competition among stock exchanges.

I show that a key driver of stock exchanges’ competition on order-processing speed is the Order Protection Rule, implemented as part of Regulation National Market Systems (Reg NMS) in 2007. The Order Protection Rule requires exchanges to prevent trade-through (i.e., to prevent a market order from being executed at an inferior price than the best price quoted on other exchanges). Preventing trade-through is vital as higher trade-through rates harm equity markets by increasing the possibility that investors will not receive best prices, discouraging investors from displaying their orders. To comply, exchanges must route orders to other exchanges with better prices.
The impacts of Order Protection Rule on inter-exchange competition depend on how fast each exchange is informed about the best prices quoted on other exchanges. This, in turn, depends on two speeds: the order-processing speeds of exchanges and the connection speeds between exchanges. If exchanges could process orders more quickly and send price information to other exchanges with low latency, each exchange would also be informed of the best prices on other exchanges more quickly. My paper asks: what incentives lead exchanges to invest or to avoid investing in these two speeds? How do these two speeds affect liquidity and the welfare of long-term investors? What are the policy implications for inter-exchange competition? I build a continuous time trading model to address these issues.

My model works as follows: A single security is traded on multiple exchanges. There are three types of traders: (1) liquidity providers (high-frequency traders or HFTs), who choose to which exchange they will provide liquidity by posting limit orders; \(^1\) (2) long-term investors, e.g., retail or institutional traders, who arrive stochastically with an inelastic need to buy or sell the security; (3) a liquidity provider called undercutting HFT arrives stochastically, and upon arrival undercutting HFTs submit price-improving orders that improve the current best price quotes by one tick, \(^2\) the smallest price increment. The incentives of an undercutting HFT reflect unmodeled shocks in their inventories or risk capacities. Liquidity providers face potential adverse-selection problems due to a publicly observable signal that stochastically arrives and shifts the asset’s value up or down upon arrival. After observing this signal, HFTs who do not provide liquidity race to trade at the old quotes to make profits, while liquidity providers race to send messages to cancel their stale limit orders. Since exchanges process orders sequentially, liquidity providers cannot always win the race to cancel their stale orders, which generates a cost for liquidity provision. Budish et al. (2015) called this phenomenon “sniping.” Competition among liquidity-providing HFTs pins down the equilibrium quoted price and the number of exchanges having the best price quotes.

The Order Protection Rule drives exchanges’ the arms race in order-processing speeds because the probability of trade-through is lower on fast exchanges. Fast exchanges can process under-

\(^1\) Nowadays, HFTs are the main liquidity providers in equity markets, as documented in Brogaard et al. (2014).
\(^2\) Currently, in the U.S. equity market, the tick size or the minimum price movement is one cent for stocks with prices above $1 per share. For a stock, if the current bid (highest buy) price=$10.00 and ask (lowest sell) price=$10.05, then the undercutting HFT is willing to sell at $10.04 or buy at $10.01.
cutting HFT’s orders more quickly, which means that other exchanges will be informed of best price quote more quickly, too. This raises an undercutting HFT’s payoff by increasing the opportunities for them to trade with investors, and reducing their exposure to sniping. In turn, because fast speed attracts more price-improving orders, more orders can be routed to fast exchanges to comply with the Order Protection Rule, and the trade volume on fast exchange rises, providing incentives for exchanges to compete on order processing speed.

I show that the size of the potential trade-through time window in which traders might not get best quotes depends on the differences between order-processing speeds, not their absolute levels. As a result, when all exchanges increase their processing speeds, the trade-through time window does not decrease. The probability of trade-through is increasing in this time window, and the number of exchanges having the current best price quotes. The latter can increase when all exchanges speed up. Because of exchanges’ faster order-processing speeds, liquidity-providing HFTs can more quickly adjust their quotes, and, hence, they are less subject to the risk of being sniped, which encourages them to provide their “fleeting” liquidity on more exchanges. This increases the possibility that long-term investors submit their orders to an exchange that is not chosen by the undercutting HFT; this, in turn, increases the probability of trade-through. This scenario may explain why investors have recently complained about the complexity of the equity markets. Due to the “fleeting” liquidity on almost all exchanges, it is hard for investors to discern which exchange might offer price improvements. As a result, when all exchanges speed up, the welfare of long-term investors may fall.

In sharp contrast, I show that increasing the connection speeds between exchanges can significantly reduce the overall trade-through rates, and improve the welfare of long-term investors; nonetheless, exchanges do not have incentives to increase connection speeds. With fast connection speeds, each exchange is informed of the current best prices from other exchanges more quickly, which reduces the probability of trade-through. In reality, however, exchanges do not have incentives to increase connection speeds because slower connection speeds reduce the competition, and increase an exchange’s trading volume. Intuitively, with slower connection speeds, liquidity-providing HFTs will not immediately cancel their orders, even if there is a better price on another exchange, because slow connection speeds result in more “separation” and, thus, less price competition among the exchanges. Since liquidity providers’ orders stay at exchanges for
longer time, the probability of sniping on these orders increases, which increases the overall trading volume for all exchanges. This observation underlies why exchanges have no incentives to increase connection speeds. The above analysis underscores a key observation: exchanges do not necessarily compete on liquidity-enhancing dimensions.

My results regarding order-processing speeds and the connection speeds between exchanges match the stylized facts that: exchanges are continuously increasing their order-processing speeds while the connection speeds between exchanges remain the same.\(^3\) I show that slow exchanges lose trading volume to fast exchanges because liquidity providers prefer to submit price-improving orders to fast exchanges. For this to occur, the two conditions must be met: 1) the stock’s bid-ask spread—the difference between the lowest quoted sell price and the highest quoted buy price—must exceed one tick; and 2) the Order Protection Rule must be present. I provide supporting empirical evidences for my theory.

My first empirical test shows that slow exchanges differentially lose trading volume to fast exchanges in stocks whose bid-ask spread is less likely to bind at one tick. When the spread binds, the lowest sell price is one tick above the highest buy price, so no liquidity providers can undercut the quotes. As a result, the trading volume on a faster exchange would increase by less than that for stocks where the price tick is less binding. To test this prediction, I exploit the Tick Size Pilot Program introduced by the U.S. Securities and Exchange Commission (SEC) in October 2016. The program increased the tick size from one cent to five cents for 1,200 randomly selected stocks with small capitalizations. The Investors Exchange (IEX) has a slower order-processing speed than other exchanges.\(^4\) Therefore, my theory predicts that IEX’s market share of total trading volume in stocks with a five-cent tick size should rise because for these stocks their bid-ask spreads are more likely to bind at one tick (five cents). I use a difference-in-differences approach to test this prediction. I find that IEX’s market share of total trading volume in stocks with a five-cent tick rises by 13 percent (from 1.77 percent to 2.00 percent) compared to stocks with a one cent tick.

My second test investigates whether, in the wake of shifting from a dark pool to a public ex-

\(^3\)Currently, despite the availability of high-speed microwave connectivity, stock exchanges still use fiber-optic cables to connect each other with latency of about 350 microseconds. Indeed, HFTs use this connectivity to reduce the latency in their connections between exchanges to about 100 microseconds.

\(^4\)IEX intentionally delays all incoming orders and messages to its matching engine by 350-microseconds.
change status, IEX attracted more trading volume in stocks with binding bid-ask spreads relative to stocks with non-binding bid-ask spreads. My model predicts that, as a result, IEX would attract more trading volume in binding stocks than in non-binding stocks because undercutting is possible for non-binding stocks, and price-improving orders are less likely to be on IEX. I compare IEX’s market share of total trading volume in binding and non-binding stocks (excluding those stocks in the Tick Size Pilot Program) three months before and after it became a public exchange; the comparison reveals that on average IEX gained 0.17 percentage points more in binding stocks than in non-binding stocks. This represents roughly 37 percent of IEX’s three-month average market share in non-binding stocks before becoming a public exchange.

Existing literature mainly focuses on speed competition among traders (Hoffmann (2014), Biais et al. (2015), Budish et al. (2015), Yao and Ye (2017) and Wang and Ye (2017)). My paper contributes to this literature by looking at the speed competition among stock exchanges. Pagnotta and Philippon (2016) maintain that traders prefer fast venues because they can realize their gains from trading earlier due to the time-discount factor. But they cannot explain why stock exchanges compete on a microsecond level because the time discount is not a factor on a sub-second basis. In my paper, fast exchanges attract liquidity-providing HFTs. The feature, in which HFTs usually post their orders on exchanges for tiny amount of time (e.g., below one millisecond), results in their demand for high-speed exchanges.

My paper also contributes to the new line of research on the competition and industrial organization of the securities market by providing a flexible inter-exchange competition model. To mitigate sniping, and to reduce the speed advantage of HFTs, several new exchange designs have been proposed: Budish et al. (2015) suggest switching from the current continuous trading process to a discrete time batch trading process. IEX delays all incoming orders by a short time, while the Chicago Stock Exchange (CHX) has proposed a similar design that only creates a short delay for orders that trade against resting orders on CHX. A common question raised in debates is: without any regulation, can an exchange that implements these designs survive when comp-

---

5IEX became a public exchange on September 2nd, 2016. Previously, IEX was a “dark pool.” That is, it did not publicly display orders. Orders in dark pools are matched within the exchange’s bid-ask spread. Orders submitted to dark pools are not protected by the Order Protection Rule.

6Menkveld and Zoican (2016) also look at how an exchange’s speed affects liquidity. But they work on a single exchange setup and cannot explain why stock exchanges become faster and faster.
peting against other faster exchanges? In current equity markets, HFTs typically submit and cancel their orders at the microsecond level. Such fast trading speeds entangled with the Order Protection Rule make modeling inter-exchange competition a challenge for researchers.

I overcome this challenge by specifically determining the potential trade-through time window for an order submitted to any exchange. Trade-through is only possible within this time window, which depends on all exchanges’ order-processing speeds, and the connection speeds between exchanges. In this way, I can determine exactly when an exchange must route orders out to comply with the Order Protection Rule. In Wang (2017), I use the same approach to explore how newly designed exchanges compete against other traditional exchanges for trading volume. Baldauf and Mollner (2017) also study these newly designed exchanges by assuming that the exchange’s goal is to reduce the bid-ask spread. In my model, exchanges maximize expected profit, which reflects per-unit time trading volume. In this setting, I can address a variety of inter-exchange competition questions.

My paper also has policy implications on recent tick-size debates. O’Hara et al. (2015), Yao and Ye (2017), and Wang and Ye (2017) suggest that the tick size should be reduced. Rindi and Werner (2017), Griffith and Roseman (2016), and Song and Yao (2016) have documented evidence that increasing tick size does not improve liquidity, at lease for small investors. In my paper, when the tick size is large, and when all exchanges speed up, overall trade-through rates are more likely to increase, which harms investor welfare. I find a new channel that large tick sizes may reduce liquidity through exchanges’ speed. Thus, my analysis also suggests that reducing tick size can improve liquidity.

The paper is organized as follows. Section 1.2 sets up the model. Section 1.3 studies speed competition between exchanges. Empirical tests are presented in Section 1.4. Section 1.5 concludes. All proofs are in the Appendix.

7In his 2017 AEA/AFA joint luncheon address, Eric Budish has discussed some issues on how frequent batch auctions exchange competes with traditional limit order book exchanges. More details could be found at https://www.aeaweb.org/webcasts/2017/luncheon.php.
1.2 Baseline Model

In this section, I first describe my trading model when the order processing and connection speeds are given exogenously. I then describe the potential trade-through time window. I endogenize an exchange’s speed investment in Section 1.3.

1.2.1 Model Setup

**Exchanges and limit order book.** $M$ exchanges use continuous limit order book to conduct trades. Traders can use either market or limit orders to trade. A market order only specifies the quantity and will be executed immediately at the best available price. A limit order is an order to buy or sell at a specified price or better. For example, a limit buy order indicates that the trader wants to buy the stated amount of the asset if the transaction price does not exceed the quoted price in the limit order. The remaining non-executed portion is posted on the exchange’s limit order book. All limit buy orders are stored on the bid side and all limit sell orders are stored on the ask side. The minimum sell price and highest buy price available at time $t$ are called the best ask price $a_t$ and best bid price $b_t$. The difference $s_t = a_t - b_t$ is the bid-ask spread. A larger bid-ask spread is a symptom of less liquidity because traders must pay a higher transaction cost.

**Traders.** There are infinite number of risk neutral HFTs choosing whether to post limit orders on exchanges to provide liquidity to fundamental investors who arrive randomly. Fundamental investors attach an exogenous intrinsic value to trade, reflecting, for example, a need to re-balance their portfolios. Fundamental investors include mutual funds, pension funds and retail traders.

**Price grids.** The smallest price increment or tick size is given by $d > 0$. In current equity markets, the tick size is one cent for stocks with price above $1$ per share. Let $\mathcal{P} = \{p^i\}_{i=-\infty}^{\infty}$ denote the discrete set of available prices for quoting and trading: the distance between any two con-

---

8Currently in U.S. equities market there are 12 active exchanges: NYSE, NYSE Arca and NYSE American owned by NYSE; EDGX, BATS BZX, BATS BYX, and EDGA owned by BATS; NASDAQ, NASDAQ BX and NASDAQ PSX owned by NASDAQ; the Investors Exchange (IEX) and Chicago stock exchange (CHX). Continuous limit order book is the most popular trading mechanisms used by most exchanges all over the world to organize trades including all public exchanges in U.S. equities market.
secutive prices in \( \mathcal{P} \) is \( d \).

**Timing and Asset.** Time runs continuously on \([0, \infty)\). There is a single risky asset that is traded on all \( M \) exchanges and one risk-free numeraire asset with price normalized to be 1. At the beginning of the trading game, the risky asset has an expected value of \( v_0 \). To ease presentation, I assume \( v_0 = (p^i + p^{i+1})/2 \) for some \( i \in \mathbb{N} \), i.e., \( v_0 \) is at the midpoint of a price grid. At \( t = 0 \), HFTs choose the exchanges on which they post their limit orders. Then, three events may occur:

1. An fundamental investor with intrinsic value \( \hat{\theta} \) to trade may arrive. I assume that the arrival time is exponentially distributed with intensity parameter \( \lambda_f \). Upon arrival, the investor will buy or sell one unit of the risky asset with equal probability and only use market orders. A buyer arriving at time \( t \) and paying \( y_t \) to buy one unit of the risky asset has utility or welfare \( w_t = v_t - y_t + \hat{\theta} \), where \( v_t \) is the risky asset’s value at time \( t \). A seller’s welfare is defined in a similar way. Further, I assume there are \( \gamma \) portions of investors are sophisticated investors. These investors will consider potential price improvements when choosing which exchange to trade although all exchanges may have the same observed quoted price. Other \( 1 - \gamma \) portions are unsophisticated investors. Upon arrival, they will randomly chose one exchange having the current best price quotes to trade with equal probability. The portion of sophisticated investors only matters when there is heterogeneity in exchanges’ order processing speeds because the probability of each exchange offering potential price improvements might be different.

2. Before fundamental investors arrive, a signal related to the risky asset’s common value may arrive. This is the sniping phenomenon that BCS analyze. I assume the arrival time of this signal is given by an exponential distribution with intensity parameter \( \lambda_j \). This signal is publicly observable by all traders at exactly the same time. With equal probability it is a good or bad signal. Conditional on good signal, the risky asset’s common value will increase by \( \sigma = kd \) for some \( k \in \mathbb{N} \). Similarly, if it is a bad signal, the risky asset’s common value will decrease by \( \sigma \). If \( \sigma \) exceeds the current half bid-ask spread, those HFTs who have posted limit orders at exchanges will run to cancel their stale limit orders while other HFTs will try to trade at the stale price.

3. Alternatively, after HFTs post their limit orders on exchanges, an undercutting HFT may arrive who will offer a one tick price improvement of the current prices quoted by other HFTs. The arrival time of undercutting HFTs is given by an exponential distribution with intensity
parameter $\lambda_U$. Upon arrival, with equal probability the undercutting HFT will submit (a) a limit buy order with price one tick above the current bid price; or (b) a limit sell order with price one tick below the current ask price. If at the time when the undercutting HFT arrives the bid-ask spread is binding at one tick, undercutting HFT will not post any order. Alternatively, one could model that the undercutting HFT will choose an exchange to quote based on the depth on each exchanges. This is outside the scope of the current paper.

The arrival process for the fundamental investor, public information, and undercutting HFT all assumed to be independently distributed. Figure 1.1 draws the event timeline of one stage trading game. The conditional probabilities of each event is shown in the graph. The stage trading game ends whenever trade occurs, at which point the next stage begins.

![Figure 1.1: Event Time line of the Baseline Model](image)

1.2.2 Exchanges Order Processing and Connection speeds

Let $\delta_i$ be the amount of time that it takes for exchange $i$ (for $i = 1, 2, \cdots, M$) to process an incoming order or cancellation message. A small $\delta_i$ indicates a faster order processing speed. At the cutting edge of technology, $\delta_i$ is about 50 microseconds. I denote the time that it takes to send price information between exchange $i$ and exchange $j$ by $\epsilon_{ij}$, where $\epsilon_{ij} = \epsilon_{ji}$. A smaller $\epsilon_{ij}$ indicates faster connection speeds between exchanges. Currently, $\epsilon_{ij}$ is about 350 microseconds in U.S. equity market.

Figure 1.2 draws the timeline of information flow, when an undercutting HFT arrives at time
$t$ and posts her price-improving order on exchange $i$. At time $t + \delta_i$, exchange $i$ completes its processing of this order. Exchange $i$ will disseminate this information to all traders and other exchanges. I assume HFTs are co-located at all exchanges as in reality. As a result, all HFTs learn of the existence of this new order at time $t + \delta_i$.

Another exchange $j$ will receive this new price information and know that exchange $i$ has better price at time $t + \delta_i + \epsilon_{ij}$. That is, it takes an additional $\epsilon_{ij}$ units of time for this information to arrive at exchange $j$. Since exchange $j$ needs $\delta_j$ units of time to process an incoming market order, the processing of any market order sent to exchange $j$ after $t + \delta_i + \epsilon_{ij} - \delta_j$ will be completed after $t + \delta_i + \epsilon_{ij}$. By then, exchange $j$ is informed about the best price on exchange $i$. So if exchange $i$ has a better price, exchange $j$ must route this market order to exchange $i$ in order to comply with the Order Protection Rule. If a market order arrives at exchange $j$ between $t$ to $t + \delta_i + \epsilon_{ij} - \delta_j$, exchange $j$ will immediately execute this order on its own platform, although a better price is available at exchange $i$. As a result, trade-through can occur between $t$ to $t + \delta_i + \epsilon_{ij} - \delta_j$, which I call the potential trade-through time window.

![Figure 1.2: Potential Trade-Through Time Window](image)

Figure 1.3 presents a more complete information flow and latency among exchanges and HFTs. Exchanges now use fiber optic cable to connect with each other. HFTs co-locate with all exchanges and the latency between an exchange and its co-located HFT is, in essence, zero. Currently, HFTs use microwave to send information between exchanges. This latency is denoted by $\zeta$ in the graph. Information flow among HFTs (red part) is faster than information flow among exchanges (blue part). My main analysis focuses on exchange’s order processing speeds and connection speeds.
Finally, an investor who does not co-locate with exchanges and does not buy the real time direct data feed from exchanges must rely on the quoting and trading information disseminated by Securities Information Processor (SIP) to make trading decisions. The SIP for tape A (listed on NYSE) and Tape B (listed on local exchanges) stocks is located at NYSE while SIP for Tape C stocks (listed on Nasdaq) is located at Nasdaq. All exchanges have to report its quoting and trading updating information to the specific SIP. Because SIP has to consolidate information from all exchanges. Its latency denoted by \( \eta \) in Figure 1.3 is larger than the latency among exchanges. Currently when the NYSE sends order updating information to the SIP in Nasdaq, it takes around 1000 microseconds. Therefore, HFTs can observe any changes in the market and respond to it before other exchanges getting these updates. An investor without co-location and direct data feed is the last one to observe market movements. That is, \( \zeta < \epsilon < \eta \).

**Remarks on model setup.** My baseline model is stylized but should not be interpreted literally. The role of the model is to deliver the main intuition in my paper. Compared to other traditional liquidity provision models, the new feature in my model is the undercutting HFT. Although I model it as an inventory shock, it could be interpreted more broadly. HFTs who specialize in liquidity provision must continuously monitor the status of the limit order book, their queue positions and learn information from other traders’ limit orders. They need to continuously readjust their limit orders as the status of the limit order book changes. This phenomenon has been empirically studied in Hasbrouck (2015) and has been modeled as market making HFTs playing mixed strategy in Baruch and Glosten (2016). I add this feature is to address how exchanges’ order processing and connection speeds affect HFTs’ liquidity provision.

1.3 Equilibrium Analysis of Exchange Speed Competition

In this section, I will first study the equilibrium at giving exchanges’ order processing and connection speeds. Then, I will endogenize exchange’s speed investments and identify under which conditions they are engaging in a speed investment arms race.
1.3.1 Exogenous Exchange Speed

In this subsection I assume all exchanges have exactly the same order processing speed denoting as $\delta_i = \delta$ and exchanges need the same units of time for sending price updating information between them denoting as $\epsilon_{ij} = \epsilon$ for all $i, j \in \{1, 2, \ldots, M\}$. The goal is to examine how these two different notions of speed affect exchange’s trading volume and investor welfare.

**Equilibrium Spread and Depth.** Because all exchanges are homogeneous, undercutting HFT will randomly choose one exchange having current best price quotes with equal probability to submit her price-improving limit order. When exchanges have different order processing speeds, undercutting HFT’s trading strategy is presented in Lemma 1 in Section 1.3.2. In order to determine exchange’s per unit time trading volume, we need to pin down the equilibrium spread $s^*$ and consolidated market depth $M^*$ first. Since the game is symmetric, at $t = 0$ the equilibrium ask and bid price would be $v_o + s^*/2$ and $v_o - s^*/2$. Because investor’s trading size is one unit, at a specific exchange there is at most one limit order with unit size on the ask and bid side of its limit order book. As a result, the consolidated market depth $M^*$ indicates the number of exchanges that have the current best price quotes.

Specifically, suppose HFTs post limit sell orders at $v_o + \frac{s^*}{2}$ and limit buy orders at $v_o - \frac{s^*}{2}$ on $X$ exchanges among those $M$ exchanges, where $\frac{s^*}{2}$ denotes the half bid-ask spread. Denote $\pi(\frac{s^*}{2}, X)$ as the liquidity provision profit for a HFT who submits these limit orders on one of those $X$ exchanges. This profit depends on which event happens first: investor arrival, the risky asset’s common value jumping or undercutting HFT arrival. I denote the arrival time of these three events as: $t_I$, $t_J$ and $t_U$. For undercutting HFT, I define an indicator function as following:

**Definition 1** $\chi_i = 1$ if the undercutting HFT submits her order to exchange $i$ where $i \in \{1, 2, \ldots, M\}$.

I illustrate $\pi(\frac{s^*}{2}, X)$ as it is the liquidity provision profit on exchange 1 (so exchange 1 is one among those $X$ exchanges). If a fundamental investor arrives first, she will randomly choose one among those $X$ exchanges to trade with equal probability. The liquidity-providing HFT on exchange 1
has $\frac{1}{X}$ chance to earn the half spread:\(^9\)

\[
\pi(\frac{s}{2}, X| t_f < t_j, t_U) = \frac{1}{X} \frac{s}{2}
\]  

(1.1)

When the risky asset’s common value jumps first, the liquidity-providing HFT’s limit order on exchange 1 will be sniped because there are infinite number of sniping HFTs. In this case, liquidity-providing HFT on exchange 1 will lose $\sigma - \frac{s}{2}$. Denoted as:

\[
\pi(\frac{s}{2}, X| t_f < t_i, t_U) = -(\sigma - \frac{s}{2})
\]  

(1.2)

If undercutting HFT arrives first and sends her price-improving order to exchange 1, the liquidity-providing HFT on exchange 1 will know the existence of this new order exactly $\delta$ units of time after undercutting HFT’s arrival as shown in Figure 1.2. She will cancel her own order that has inferior price at this time and have liquidity provision profit:

\[
\pi(\frac{s}{2}, X| t_U < t_i, t_f; \chi_1 = 1) = \phi(\delta)[\frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1}{2} \frac{1}{2} \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s}{2})] +
\]

\[
[1 - \phi(\delta)][\frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1}{2} \frac{1}{2} \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} \frac{1}{2} (\sigma - \frac{s}{2})]
\]  

(1.3)

Where $\phi(\delta) = 1 - e^{-(\lambda_I + \lambda_J)\delta}$ is the probability that either an investor or signal jumping arrives within the $\delta$ units of times after undercutting HFT’s arrival. within this time, the liquidity-providing HFT on exchange 1 has not canceled her order. In this case, her profit is the first term in the right hand side of equation (1.3). There is $\frac{1}{2}$ in the revenue part because undercutting HFT has better price on either the bid or ask side. If no event happens within the $\delta$ units of time, the liquidity-providing HFT on exchange 1 will cancel her limit order that has inferior price than undercutting HFT’s order. So essentially after $\delta$ units of time, the original liquidity-providing HFT will only provide liquidity on one side of the market. This is the second term in the right hand side of equation (1.3).

If the undercutting HFT arrives first but she does not send her price-improving order to exchange 1, to simply exposition, in all the remaining analysis $t_f < t_j, t_U$ means $t_f < t_j$ and $t_f < t_U$. Other similar notations have the same meaning.
change 1, then the liquidity-providing HFT on exchange 1 will cancel her limit order that has inferior price than undercutting HFT’s order $\epsilon$ units of time after undercutting HFT’s arrival as shown in Figure 1.2. After $t_U + \epsilon$, because of the Order Protection Rule, the limit order with inferior price has no chance to trade with fundamental investors.\footnote{For example, if the undercutting HFT submits limit sell order at $v_0 + \frac{s}{2} - d$ to exchange 2 when she arrives, then the liquidity-providing HFT at exchange 1 will cancel her limit sell order at $v_0 + \frac{s}{2}$ exactly $\epsilon$ units of time after the undercutting HFT’s arrival. This can be seen clearly from Figure 1.2. Thus, if the market order arrives at exchange 1 after $t_U + \epsilon$, exchange 1 must reroute the order to exchange 2. This implies that, after $t_U + \epsilon$ limit sell order on exchange 1 at the price $v_0 + \frac{s}{2}$ has no chance to trade with an investor. As a result, liquidity-providing HFT at exchange 1 will cancel her limit sell order at $t_U + \epsilon$.} In this case, her profit from liquidity provision is similar:

\[
\pi\left(\frac{s}{2}, X|t_U < t_1, t_j; \chi_1 = 0\right) = \phi(\epsilon)\left[\frac{\lambda_I}{\lambda_I + \lambda_J X} \frac{1}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s}{2})\right] + \\
\left[1 - \phi(\epsilon)\right]\left[\frac{\lambda_I}{\lambda_I + \lambda_J X} \frac{1}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} \frac{1}{2} (\sigma - \frac{s}{2})\right]
\]

(1.4)

where $\phi(\epsilon) = 1 - e^{-(\lambda_I + \lambda_J)\epsilon}$. The first term in the right hand side of equation (1.4) is the liquidity-providing HFT’s profits when she does not cancel her order with inferior price. The second term is the profit after she cancels her order with inferior price.

Note that since all HFTs co-locate at all exchanges, the liquidity-providing HFT at exchange 1 knows the existence of the undercutting HFT at other exchanges at $t_U + \delta$ and she can adjust her quotes on exchange 1 at $t_U + \delta + \zeta$, where $\zeta$ is the time for HFTs to send an information from one exchange to another exchange as drawn in Figure 1.3. Here I implicitly assume: $\delta + \zeta < \epsilon$. This simply means that HFTs can respond to market movements faster than exchanges. This is what happens in practice as HFTs using microwaves to send information among exchanges while exchanges use fiber optic.

Combining all above cases, when $s/2 \leq \sigma$, $\pi\left(\frac{s}{2}, X\right)$ is defined as:

\[
\pi\left(\frac{s}{2}, X\right) = \frac{\lambda_I}{\Sigma \lambda} \pi\left(\frac{s}{2}, X|t_I < t_j, t_U\right) + \frac{\lambda_J}{\Sigma \lambda} \pi\left(\frac{s}{2}, X|t_I < t_j, t_U\right) + \\
\frac{\lambda_U}{\Sigma \lambda} \frac{1}{X} \pi\left(\frac{s}{2}, X|t_U < t_I, t_j; \chi_1 = 1\right) + \frac{\lambda_U}{\Sigma \lambda} \frac{1}{X} \pi\left(\frac{s}{2}, X|t_U < t_I, t_j; \chi_1 = 0\right)
\]

(1.5)

where $\Sigma \lambda = \lambda_I + \lambda_J + \lambda_U$. Competition among HFTs will drive this profit to zero, which pins
down the equilibrium spread and consolidated market depth given in the following proposition.

**Proposition 1** (Equilibrium Spread and Depth) When $\delta_i = \delta$ and $\epsilon_{ij} = \epsilon$ for all $i, j \in \{1, 2, \cdots, M\}$:

(i) The equilibrium bid-ask spread is given by:

$$s^*(\delta, \epsilon) = \begin{cases} d; & \text{if } \frac{\lambda_j}{\lambda_i + \lambda_j} \leq \frac{d}{2\sigma} \\ \min\{s|v_0 \pm \frac{s}{2} \in \mathcal{P}, \pi(\frac{s}{2}, 1) \geq 0\}; & \text{if } \frac{\lambda_j}{\lambda_i + \lambda_j} > \frac{d}{2\sigma} \end{cases}$$  \hspace{1cm} (1.6)

where $\mathcal{P}$ is the available price grids set and $\pi(\frac{s}{2}, 1)$ is defined as in equation (1.5).

(ii) The equilibrium consolidated market depth is given by:

$$M^*(\delta, \epsilon) = \begin{cases} M; & \text{if } \frac{s^*}{2} \geq \sigma \\ \max\{X|1 \leq X \leq M, \pi(\frac{s^*}{2}, X) \geq 0\}; & \text{if } \frac{s^*}{2} < \sigma \end{cases}$$  \hspace{1cm} (1.7)

Where the half bid-ask spread $s^*/2$ is determined in (1.6) and $X \in \mathbb{N}$.

Note that equilibrium bid-ask spread is pinned down by HFT’s liquidity provision profit on a single exchange $\pi(\frac{s}{2}, 1)$. Because price is discrete, at the equilibrium bid and ask prices, HFTs may be able to provide liquidity on multiple exchange. This consolidated market depth is given by (1.7). Although, the definition for (1.5) need to be adjusted for $s/2 > \sigma$ because of no sniping. But since (1.5) is positive for all $s/2 > \sigma$ and is strictly increasing in $s$, the equilibrium spread can be uniquely determined in (1.6) even when $s^*/2 > \sigma$. When $\frac{\lambda_j}{\lambda_i + \lambda_j} \leq \frac{d}{2\sigma}$ the equilibrium spread is binding at one tick $d$, in order to let undercutting HFTs playing a role, for the all remaining analysis I assume:

**Assumption 1** $\frac{\lambda_j}{\lambda_i + \lambda_j} > \frac{d}{2\sigma}$

Note that this assumption does not conflict with the observation that for many stocks their bid-ask spreads are at one tick very often. Under Assumption 1, in my model if $\lambda_U$ is large or HFTs are quite often to undercut each other, then the bid-ask spread could also be at one tick very often. The difference is that if $\frac{\lambda_j}{\lambda_i + \lambda_j} \leq \frac{d}{2\sigma}$, bid-ask spread would be binding at one tick all the time while Assumption 1 implies that sometimes the spread is binding at one tick and it may be wider than one tick during other times. This is certainly more close to reality. Now we can examine
how exchange’s order processing and connection speeds affect the equilibrium spread and depth. These results are summarized in the following corollary.

**Corollary 1** (Comparative Analysis on Equilibrium Spread and Depth)

(i) *Equilibrium bid-ask spread* \( s^*(\delta, \epsilon) \) *is weakly increasing in* \( \delta \) *and is independent of* \( \epsilon \);

(ii) *Equilibrium depth* \( M^*(\delta, \epsilon) \) *is weakly increasing in* \( \epsilon \);

(iii) If for some \( \delta_F < \delta_S \) and \( s^*(\delta_F, \epsilon) = s^*(\delta_S, \epsilon) \), then \( M^*(\delta_F, \epsilon) \geq M^*(\delta_S, \epsilon) \).

(i) implies that fast exchanges can reduce the cost of liquidity provision. Liquidity-providing HFTs can respond to any news or changes in the limit order book more quickly on a faster exchange. This reduces the adverse selection cost for liquidity-providing HFTs. Because price is discrete, equilibrium bid-ask spread \( s^* \) is weakly increasing in \( \delta \). The equilibrium spread does not depend on connection speed because the equilibrium bid-ask spread is pinned down by HFT’s liquidity provision profit on a single exchange, in which connection speed between exchanges can not play a role.

(ii) points out that when information flow between exchanges is slow, HFTs will provide liquidity on more exchanges at the equilibrium bid and ask prices because each HFT faces less price competition from HFTs on other exchanges. In other words, market is more fragmented if the connection speed between exchanges is slow.

(iii) simply states that when exchanges increase their order processing speeds, if the equilibrium bid-ask spread stays the same due to price discreteness, HFTs’ liquidity provision profit will increase. This increased liquidity provision profits result in HFTs to provide liquidity on more exchanges.

**Investor Welfare.** My welfare analysis focuses on fundamental investors. They are mutual funds, pension funds or retail investors. Their welfare or transaction cost is an important measure of the efficiency of equity markets. Ideally, one could use the equilibrium bid-ask spread to measure investor’s transaction cost as in Glosten and Milgrom (1985) and among others. But in my model, because an investor may arrive at the market after an undercutting HFT, the investor may get better price than the equilibrium bid or ask prices. As a result, in order to properly measure investor welfare, we need to take into account the different limit order book status at the
time when the investor arrives. Specifically, what I want to measure is: at \( t = 0 \) before trading starts, what is the ex ante average transaction cost for an investor to buy or sell one unit of the risky asset when all exchanges have the same order processing speed \( \delta \) and connection speed \( \epsilon \). I denote this cost as \( TC(\delta, \epsilon) \). Because the game is symmetric, the transaction cost is the same for an investor to buy or sell.

Note that, at \( t = 0 \) we do not know when the investor will arrive. We have:

\[
TC(\delta, \epsilon) = \lambda_I \frac{s^*}{2} + \lambda_J TC(\delta, \epsilon) + \lambda_U \frac{\lambda_U}{\lambda} \text{Prob}(t_I \geq t_J | t_I, t_J > t_U)TC(\delta, \epsilon) + \\
\lambda_U \frac{\lambda_U}{\lambda} \text{Prob}(t_I < t_J, t_I \leq t_U + \epsilon | t_I, t_J > t_U)\left[ \frac{1}{2} s^* + \frac{1}{2} (\frac{M^* - 1}{M^*} \frac{s^*}{2} + \frac{1}{M^*} (\frac{s^*}{2} - d)) \right] + \\
\lambda_U \frac{\lambda_U}{\lambda} \text{Prob}(t_I < t_J, t_I > t_U + \epsilon | t_I, t_J > t_U)\left[ \frac{1}{2} s^* + \frac{1}{2} \left( \frac{s^*}{2} - d \right) \right]
\]  \hspace{1cm} (1.8)

where \( t_I, t_J \) and \( t_U \) denote the arriving time of the investor, the risky asset’s common value jumping and the undercutting HFT. \( \text{Prob}(t_I \geq t_J | t_I, t_J > t_U) \) is the probability of \( t_I \geq t_J \) or the risky asset’s common value jumps before the investor arrival conditional on undercutting HFT arrives first. Others are defined in the similar way. \( s^* \) and \( M^* \) are the equilibrium bid-ask spread and consolidated market depth given in Proposition 1. With probability \( \lambda_I/\lambda \) the investor arrives at the market first, in this case her transaction cost is the half bid-ask spread \( s^*/2 \). This is the first term in the right hand side of equation (1.8).

With probability \( \lambda_J/\lambda \) the risky asset’s common value jumps first, the game will move to next stage. So the investor’s expected transaction cost in this new stage game is the same \( TC(\delta, \epsilon) \). This is the second term in the right hand side of equation (1.8).

With probability \( \lambda_U/\lambda \) an undercutting HFT arrives first, the investor’s transaction cost depends on the time she arrives at the market. If she arrives after the risky asset’s common value jumps, the game will also move to a new stage game. The investor’s transaction cost would be \( TC(\delta, \epsilon) \) again. This is the third term in the right hand side of equation (1.8). If the buyer arrives before the risky asset’s common value jumps and is within the \( \epsilon \) units of time after the undercutting HFT’s arrival, trade-through is possible. Specifically, if the investor is a buyer and the undercutting HFT is a seller, the probability for the investor to trade with the undercutting HFT is \( 1/M^* \). Because the investor sends her market buy order to one among those \( M^* \) exchanges with equal probability.
and undercutting HFT only provides liquidity on one exchange. The transaction cost to trade with this undercutting HFT is $s^*/2 - d$. Otherwise, it would be $s^*/2$. If the undercutting HFT is a buyer too, then there is no price improvement opportunity available for the buyer investor. In this case, her transaction cost is $s^*/2$. The undercutting HFT has equal probability to be a buyer or seller. This explains the forth term in the right hand of equation (1.8). If the buyer arrives after $t_U + \epsilon$ and before the risky asset’s common value jumps, there is no trade-through. If the undercutting HFT and the investor are at the opposite side of the market (one is a seller and the other one is a buyer and vice versa), the transaction cost for the investor would be $s^*/2 - d$. Otherwise, no price improvement and the transaction cost for the investor is $s^*/2$. This is the last term in equation (1.8).

Since an investor realizes a private value $\bar{\theta}$, if she trades one unit of the risky asset, we can define the investor ex ante expected welfare as:

$$W(\delta, \epsilon) = \bar{\theta} - TC(\delta, \epsilon)$$  \hspace{1cm} (1.9)

By solving equation (1.8), we can have a closed form of $TC(\delta, \epsilon)$. I summarize these result in the following proposition.

**Proposition 2** (Investor Welfare) When $\delta_i = \delta$ and $\epsilon_{ij} = \epsilon$ for all $i, j \in \{1, 2, \cdots, M\}$, then:

(i) An investor has ex ante expected welfare:

$$W(\delta, \epsilon) = \bar{\theta} - \left( \frac{\lambda_I^2 + \lambda_J s^*}{\sum \lambda^2} \right) + \frac{\lambda_U}{\sum \lambda} \left[ \phi(\epsilon)A + (1 - \phi(\epsilon))B \right]$$  \hspace{1cm} (1.10)

where $A = \frac{1}{2} s^* + \frac{1}{2} \left[ \frac{M^* - 1}{M^*} s^* + \frac{1}{M^*} (s^* - d) \right]$, $B = \frac{1}{2} s^* + \frac{1}{2} (s^* - d)$, and $\phi(\epsilon) = 1 - e^{-(\lambda_I + \lambda_J)\epsilon}$. $s^*$ and $M^*$ are the equilibrium spread and depth given in Proposition 1;

(ii) $W(\delta, \epsilon)$ is strictly decreasing in $\epsilon$ if $M^* \geq 2$ and is independent of $\epsilon$ if $M^* = 1$;

(iii) If for some $\delta_F < \delta_S$ and $s^*(\delta_F) = s^*(\delta_S)$, then $W(\delta_F, \epsilon) \leq W(\delta_S, \epsilon)$.

Proof of (i) is in appendix. (ii) points out that if multiple exchanges have the same best bid and ask price quotes, trade-through is possible. The probability of trade-through is strictly increasing in the latency among exchanges. Increasing the connection speeds between exchanges can strictly increase investor welfare. (iii) points out a surprising result: if all exchanges speed up,
it does not necessarily increase investor welfare. Due to price discreteness, after all exchanges increase their order processing speeds, the equilibrium bid-ask spread may stay same. According to Corollary 1, when the equilibrium spread stays the same, the consolidated market depth or number of exchange having the best price quotes may increase. This will increase the probability of trade-through. This is why if all exchanges become faster and faster, investor welfare can fall.

**Exchange Per Unit Time Trading Volume.** Since there is no heterogeneity among exchanges, in the current framework all exchanges have exactly the same trading volume. I denote \( Q(\delta, \epsilon) \) as the per unit time trading volume for an exchange when all exchanges have the same order processing speed \( \delta \) and connection speed \( \epsilon \). The way I calculate this per unit time trading volume is to look at how many paths there are from \( t = 0 \) moving to a new stage game. I calculate the expected time and exchange’s expected trading volume for each path. By averaging them, I have the following results. The detailed proof is in the appendix.

**Proposition 3** (Trading Volume) When \( \delta_i = \delta \) and \( \epsilon_{ij} = \epsilon \) for all \( i, j \in \{1, 2, \cdots, M\} \), and \( s^*/2 < \sigma \) then:

(i) Each exchange has the same ex ante expected per unit time trading volume:

\[
Q^*(\delta, \epsilon) = \lambda_j \frac{1}{M} + \lambda_j \frac{M^*(\delta, \epsilon)}{M} + \frac{\lambda_U \lambda_j}{2 \Sigma \lambda} \phi(\delta) \frac{1}{M} - \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U \lambda_j}{\Sigma \lambda} \frac{M^*(\delta, \epsilon) - 1}{M} \tag{1.11}
\]

where \( M^*(\delta, \epsilon) \) is the equilibrium depth as determined in equation (1.7) and \( \phi(\epsilon) = 1 - e^{-(\lambda_I + \lambda_J)\epsilon} \);

(ii) \( Q^*(\delta, \epsilon) \) is strictly increasing in \( \epsilon \) if \( M^*(\delta, \epsilon) \geq 2 \);

(iii) If for some \( \delta_F < \delta_S \) and \( s^*(\delta_F, \epsilon) = s^*(\delta_S, \epsilon) \), then \( Q^*(\delta_F, \epsilon) > Q^*(\delta_S, \epsilon) \) if \( M^*(\delta_F, \epsilon) > M^*(\delta_S, \epsilon) \)

and \( Q^*(\delta_F, \epsilon) < Q^*(\delta_S, \epsilon) \) if \( M^*(\delta_F, \epsilon) = M^*(\delta_S, \epsilon) \).

The result in equation (1.11) is intuitive. Within one unit of time, when an investor arrives, she will trade one unit of the risky asset. When the risky asset’s common value jumps, all stale limit orders are taken by sniping HFTs. Therefore, there would be \( M^* \) units trading volume. If undercutting HFT arrives before the value jumps, since other liquidity-providing HFTs will cancel their being undercut limit orders \( \epsilon \) units of time after undercutting HFT’s arrival, the exchange’s trading volume would be reduced in this case. This is the negative term in equation (1.11). When the undercutting HFT arrives, her limit order might be sniped too if followed by the risky asset’s
common value jumping. This is the third term in equation (1.11).

(ii) shows that if exchanges prefer larger trading volume, they do not have incentives to increase the connection speeds between exchanges for two reasons: 1) As shown in Corollary 1, the equilibrium depth $M^*$ is weakly increasing in $\epsilon$. As a result, exchange’s trading volume increases when the connection speed is slow; 2) With slow connection speed, liquidity-providing HFTs will keep their being undercut orders in the limit order book for a longer time. The probability of sniping on these orders increase. This increases trading volume for all exchanges. But as shown in Proposition 2, investor welfare could be strictly improved with faster connection speed. This result has important policy implications. Exchanges’ goal does not necessarily coincide with long-term investor’s welfare. It can not simply rely on the market to mitigate the cost of trade-through.

Since exchanges have exactly the same order processing speed, exchange’s speed can only affect its trading volume through the equilibrium depth. Certainly, when depth is larger, each exchange would have larger trading volume. Based on the results in Corollary 1 when all exchanges become faster, exchange’s trading volume can increase when equilibrium bid-ask spread stays the same.

### 1.3.2 Endogenous Exchange Speed

I will first look at when some exchanges have faster order processing speeds than others, how that affects fast and slow exchange’s trading volume. Then I will introduce exchange’s fee structures to endogenize exchange’s investment in order processing speed. As shown in Proposition 3, exchanges do not have incentives to increase connection speeds. In fact they have incentives to do the opposite. Thus current connection speeds are determined by regulation, pinned down by the slowest connection speed that the regulation allows. Consequently, $\epsilon_{ij} = \epsilon$ for all $i, j \in \{1, 2, \cdots, M\}$. I drop $\epsilon$ in most notation for concreteness.

**Exchange Trading Volume Under Speed Heterogeneity.** Suppose $K$ exchanges have the same fast order processing speed $\delta_F$ and other $M - K$ exchanges have the same slow order processing speed $\delta_S$, where $\delta_F < \delta_S$ and $1 \leq K \leq M - 1$ (in the case when $K = 0$ or $K = M$, all
exchanges have the same order processing speed $\delta_S$ or $\delta_F$. The results in last subsection can be directly applied). I will study how HFTs provide liquidity on these exchanges. The equilibrium spread is determined exactly the same way as in Proposition 1. If HFTs provide liquidity on fast exchanges, the smallest possible spread would be $s^*(\delta_F)$ (note that the equilibrium spread also depends on $\epsilon$ as given in Proposition 1. I drop it for easy exposition). If HFTs provide liquidity on slow exchanges, the smallest possible spread would be $s^*(\delta_S)$.

According to Corollary 1 (i), $s^*(\delta_F) \leq s^*(\delta_S)$. This is because providing liquidity on fast exchanges has smaller adverse selection cost than on slow exchanges. Later, I will show that the equilibrium spread would be either $s^*(\delta_F)$ or $s^*(\delta_S)$.

When all exchanges have the same order processing speeds and HFTs provide liquidity on multiple exchanges at the lowest spread, undercoting HFTs will randomly choose one among those exchanges with equal probability to submit her price-improving limit order. But when those exchanges with best price quotes have different order processing speeds, Lemma 1 shows that it is always optimal for the undercoting HFT to submit her order to a fast exchange.

**Lemma 1** If the best price quotes are available on some exchanges with fast order processing speed $\delta_F$, then it is always optimal for an undercoting HFT to submit her price-improving order to one among them.

This is because the probability of being traded-through is smaller on fast exchanges. Because fast exchanges can process undercoting HFT’s order more quickly. As a result, other exchanges and investors can observe this new better priced limit order with shorter delay. This increases the probability of the undercoting HFT’s order to trade with an investor and reduce its exposure to sniping.

For an investor, when all exchanges have the same order processing speeds, the investor will randomly choose one among those exchanges with the best price to trade with equal probability. This is reasonable because exchanges are homogeneous for investors. But if some exchanges have faster order processing speeds than others, undercoting HFTs strictly prefer faster exchange to submit her price-improving order. As a result, it is not reasonable to still assume that investors

\[1\] Remember that $s^*(\delta_F)$ or $s^*(\delta_S)$ are the smallest spread such that liquidity-providing HFTs can earn non-negative profits on a fast or slow exchange (\((1.6))\).
will randomly choose an exchange with better price to trade. In reality, some investors may consider the potential price improvements when make decisions on which exchange to trade. They might want to trade on those exchanges that undercutting HFTs prefer. Thus the $\gamma$ proportions of sophisticated investors will also prefer to trade on fast exchanges if best price is available. The remaining $1 - \gamma$ proportions of unsophisticated investors will still randomly choose one among those exchanges with best price to trade with equal probability. The sophisticated investors will always trade on fast exchanges if current best price quotes are available on some fast exchanges because undercutting HFT also submits her order to one among these fast exchanges. This dichotomy of investors to be sophisticated and unsophisticated has the same feature as in Foucault and Menkveld (2008). In their two competing exchanges setup, brokers responsible for routing investor’s orders have two types: the smart brokers will send order to both exchanges for best prices while the non-smart brokers only send orders to the incumbent exchange and ignore potential better price on the entrant exchange.

In reality there are several reasons why some investors might not consider potential price improvements and make trading decisions only based on the current available prices. First, even an exchange can process orders faster than other exchanges, some investors may not recognize it or they simply do not know how this might affect their transaction cost; Second, some investors rely on brokers to send their orders to exchanges. Brokers may have agreements with a particular exchange, and they will send orders to that exchanges if having the current best prices. The proportions of sophisticated investors will affect fast exchange’s trading volume. I will construct the equilibrium now.

Suppose HFTs provide liquidity on $X$ exchanges including $X_F$ fast exchanges and $X_S$ slow exchanges at half spread $s/2$. Thus, $X = X_F + X_S$. Denote $\pi_F(\frac{s}{2}, X_F, X_S)$ as HFT’s liquidity provision profit on one among those $X_F$ fast exchanges. Without loss of generality, we can still assume it is the liquidity provision profit on exchange 1. Thus, exchange 1 is one among those $X_F$ fast exchanges. I will construct $\pi_F(\frac{s}{2}, X_F, X_S)$ in a similar way as the profit in equation (1.5), which is HFT’s liquidity provision profit when all exchanges have the same order processing speed. Still

---

\textsuperscript{12}In reality, traders and brokers calculate transactions cost as a measure of execution quality on different exchanges. If they get price improvements more often on some particular exchanges, they will prefer to route their orders to these exchanges if best price quotes are available.
denote $t_I$, $t_J$ and $t_U$ as the arriving time of the investor, the risky asset’s common value jumping and undercuing HFT. We can also use the same indicator function defined in definition 1: $\chi_1 = 1$ if undercuing HFT submits her order to exchange 1. Otherwise, $\chi_1 = 0$. If the investor arrives first, liquidity-providing HFT on exchange 1 earns profit:

$$\pi_F\left(\frac{s}{2}, X_F, X_S| t_I < t_J, t_U \right) = (\frac{Y}{X_F} + \frac{1-Y}{X_F + X_S})\frac{s}{2}$$

(1.12)

This is because if the investor is sophisticated (with probability $\gamma$), she will choose one among those $X_F$ exchanges to trade because she knows that undercuing HFT also submits price-improving orders to fast exchanges. Thus, she might get better price on fast exchange than the current quotes. If it is a unsophisticated investor (with probability $1-\gamma$), she will randomly choose one among all those $X_F + X_S = X$ exchanges to trade. Liquidity-providing HFT earns half spread $s/2$ when her order is taken by an investor. If the asset’s common value jumps first, we have:

$$\pi_F\left(\frac{s}{2}, X_F, X_S| t_J < t_I, t_U \right) = -(\sigma - \frac{s}{2})$$

(1.13)

because sniping HFTs will take stale limit orders from all exchanges. If the undercuing HFT arrives first and submits her order to exchange 1, then liquidity-providing HFT on exchange 1 has profit:

$$\pi_F\left(\frac{s}{2}, X_F, X_S| t_I < t_J, t_U; \chi_1 = 1 \right) = \phi(\delta_F)\left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1}{2} \left( \frac{Y}{X_F} + \frac{1-Y}{X_F + X_S} \right) \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s}{2}) \right] + \
\left[ 1 - \phi(\delta_F) \right] \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1}{2} \left( \frac{Y}{X_F} + \frac{1-Y}{X_F + X_S} \right) \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} \frac{1}{2} (\sigma - \frac{s}{2}) \right]$$

(1.14)

The above profits could be explained in a similar way as in equation (1.3). Suppose the undercuing HFT is a seller. Thus, she is willing to sell at $v_o + s/2 - d$. The liquidity-providing HFT on exchange 1 will cancel her limit sell order at $t_U + \delta_F$. If an investor who is a buyer arrives at the market between $t_U$ to $t_U + \delta_F$ and submits her order to exchange 1, this buyer will trade with the undercuing HFT because the later sells at lower price. Thus, only when the investor is a seller, liquidity-providing HFT on exchange 1 can earn the half spread. This is why there is $\frac{1}{2}$ in the first term of equation (1.14). After $t_U + \delta_F$, liquidity-providing HFT on exchange 1 only provides
liquidity on the bid (buy) side of the limit order book. This is the second term in equation (1.14).
Similarly, if the undercutting HFT submits her order to other fast exchanges, liquidity-providing
HFT on exchange 1 has profits:

\[
\pi_F(\frac{s}{2}, X_F, X_S | t_U < t_I, t_J; \chi_1 = 0) = \phi(\epsilon)[\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{1 - \gamma}{X_F} + \frac{1 - \gamma}{X_F + X_S}) - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] +
\]

\[
[1 - \phi(\epsilon)][\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{1 - \gamma}{X_F} + \frac{1 - \gamma}{X_F + X_S}) - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] \quad (1.15)
\]

If undercutting HFT submits her order to other fast exchanges, liquidity-providing HFT on ex-
change 1 will cancel their being undercut limit order at \(t_U + \epsilon\) (see Figure 1.2). Thus, liquidity-
providing HFT on exchange 1 will provide liquidity on both side of the limit order book before
\(t_U + \epsilon\) and will only provide liquidity on one side of the limit order book after \(t_U + \epsilon\). Combining
the results in (13)-(16), we have:

\[
\pi_F(\frac{s}{2}, X_F, X_S) = \frac{\lambda_I}{\Sigma \lambda} \pi_F(\frac{s}{2}, X_F, X_S | t_I < t_J, t_U) + \frac{\lambda_J}{\Sigma \lambda} \pi_F(\frac{s}{2}, X_F, X_S | t_I < t_J, t_U) +
\]

\[
\frac{\lambda_U}{\Sigma \lambda} X_F \pi_F(\frac{s}{2}, X_F, X_S | t_U < t_I, t_J; \chi_1 = 1) + \frac{\lambda_U}{X_F} X_F \pi_F(\frac{s}{2}, X_F, X_S | t_U < t_I, t_J; \chi_1 = 0) \quad (1.16)
\]

Similarly, denote \(\pi_S(\frac{s}{2}, X_F, X_S)\) as HFT’s liquidity provision profit on one among those \(X_S\) slow
exchanges while HFTs also provide liquidity on other \(X_F\) fast exchanges at the same half spread
\(s/2\). We have:

\[
\pi_S(\frac{s}{2}, X_F, X_S) = \frac{\lambda_I}{\Sigma \lambda} \frac{1 - \gamma}{X_F} X_F + X_S \frac{s}{2} - \frac{\lambda_J}{\Sigma \lambda} (\sigma - \frac{s}{2}) + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_F + \epsilon - \delta_S)[\frac{\lambda_I}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] +
\]

\[
\frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] + \frac{\lambda_U}{\Sigma \lambda}[1 - \phi(\delta_F + \epsilon - \delta_S)][\frac{\lambda_I}{\lambda_I + \lambda_J}(\delta_F + \epsilon - \delta_S)] +
\]

\[
\frac{\lambda_J}{\lambda_I + \lambda_J}(\delta_F + \epsilon - \delta_S)] \quad (1.17)
\]

The difference between the liquidity provision profit on fast and slow exchange are: 1) only
unsophisticated investors may send their market orders to slow exchange and the proportions
of non-smart investors is \(1 - \gamma\); 2) Undercutting HFTs always submit her order to fast exchange. So
HFTs will cancel their being undercut orders on slow exchanges \(\delta_F + \epsilon - \delta_S\) (see Figure 1.2) units of
time after the undercutting HFT’s arrival. During this potential trade-through time window, HFTs
still provide liquidity at both sides of the limit order book on slow exchanges. After \(\delta_F + \epsilon - \delta_S\)
units of time, HFTs only provide liquidity at one side of the limit order book on slow exchange because they have canceled their being undercut orders. This explains equation (1.17).

Note that when \( X_S = 0 \), \( \pi_F(\frac{1}{2}, X_F, 0) = \pi(\frac{1}{2}, X_F|\delta = \delta_F) \). The later is the liquidity provision profit in equation (1.5) evaluated at \( \delta = \delta_F \). This is because when HFTs only provide liquidity on fast exchanges, the results under homogeneous order processing speed in Section 1.3.1 would apply. If HFTs only provide liquidity on one exchange, then their liquidity provision profit on fast exchange is always larger than on slow exchange for the same bid-ask spread \( (\pi(\frac{1}{2}, 1|\delta = \delta_F) > \pi(\frac{1}{2}, 1|\delta = \delta_S) \), see equation (1.5)). This is also why \( s^*(\delta_F) \leq s^*(\delta_S) \) (Corollary 1 (i)).

But when HFTs provide liquidity on both fast and slow exchanges, it is not necessary that HFTs have larger liquidity provision profit on fast exchanges than on slow exchanges. In other words, \( \pi_F(\frac{1}{2}, X_F, X_S) \) is not always larger than \( \pi_S(\frac{1}{2}, X_F, X_S) \). This is because undercutting HFTs always submit their price-improving limit orders to fast exchanges. Before it is canceled, the being undercut limit order on fast exchange has no chance to be taken by an investor but still subjects to sniping when the risky asset’s common value jumps. Liquidity-providing HFTs on slow exchanges do not have this cost because undercutting HFTs only submit orders to fast exchange. But since sophisticated investors always trade on fast exchange, when the proportions of sophisticated investors \( \gamma \) is large enough it is possible that \( \pi_F(\frac{1}{2}, X_F, X_S) \geq \pi_S(\frac{1}{2}, X_F, X_S) \) always holds. In this case, HFTs will provide liquidity on fast exchanges first and start to provide liquidity on slow exchanges only if no fast exchange is available and it is profitable to provide liquidity on slow exchanges. Proposition 4 summarizes the results when \( \gamma \) is large.

**Proposition 4** (Equilibrium with Exchange Speed Heterogeneity) *If there are \( K \) fast exchanges with order processing speed \( \delta_F \) and \( M - K \) slow exchanges with order processing speed \( \delta_S \), where \( 1 \leq K \leq M - 1 \) and \( \delta_F < \delta_S \). When \( \gamma \geq \bar{\gamma} = \frac{0.5\lambda_U\phi(e)}{\lambda_I + \lambda_I + 0.5\lambda_U} \), then:

(i) If \( M^*(\delta_F) \leq K \), HFTs provide liquidity on \( M^*(\delta_F) \) fast exchanges with bid-ask spread \( s^*(\delta_F) \) is the unique equilibrium;

(ii) If \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) < 0 \), HFTs provide liquidity on \( K \) fast exchanges with bid-ask spread \( s^*(\delta_F) \) is the unique equilibrium;

(iii) If \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) \geq 0 \), HFTs provide liquidity on \( K \) fast exchanges and
$M_S^*(K)$ slow exchanges with bid-ask spread $s^* (\delta_F)$ is the unique equilibrium, where:

$$M_S^*(K) = \max \{ X_S | \pi_S\left( \frac{s^* (\delta_F)}{2}, K, X_S \right) \geq 0, 1 \leq X_S \leq M - K \} \tag{1.18}$$

Note that $M^*(\delta_F)$ is the equilibrium depth when all exchanges have the same fast order processing speed (determined in equation (1.7)). In the appendix I show that when $\gamma \geq \bar{\gamma}$, HFT has the largest liquidity provision profit on fast exchanges for a given bid-ask spread and a given number of exchanges having the same price quotes. HFTs will run to provide liquidity on fast exchanges. Competition among HFTs will drive the bid-ask spread to its minimum $s^*(\delta_F)$ (determined in equation (1.6)). When the total number of fast exchanges $K$ is larger than $M^*(\delta_F)$, HFTs will only provide liquidity on fast exchanges. When $M^*(\delta_F) > K$, HFTs start to provide liquidity on slow exchanges until their liquidity provision profit on slow exchanges becomes negative. The depth on slow exchanges is determined in equation (1.18) while the depth on fast exchanges is always $K$.

In the Appendix B, I construct the equilibrium when $\gamma < \bar{\gamma}$. For small $\gamma$ the liquidity provision profit on fast exchanges is not necessarily larger than on slow exchanges. Therefore, when HFTs provide liquidity on both fast and slow exchanges it is possible that HFTs earn negative profits on fast exchanges when the maximum depth is reached on slow exchanges. To construct the equilibrium, I allow a single HFT to provide liquidity on multiple exchanges.¹³ When $s^*(\delta_F) < s^*(\delta_S)$ and if a single HFT can earn non-negative total profits by providing liquidity on some fast and slow exchanges, the equilibrium spread would be still $s^*(\delta_F)$ and HFTs always provide liquidity on fast exchanges too. More detailed analysis about this result could be found in the Appendix B.

In equity markets, most traders on exchanges are algorithm traders.¹⁴ These traders will monitor their transaction costs on each exchange. Based on their past trading costs, they will know whether a particular exchange has higher probability to offer potential price improvement than other exchanges or not. Thus, in reality $\gamma$ could be very high. For conciseness, in the remaining

¹³When $\gamma \geq \bar{\gamma}$, under the equilibrium in Proposition 4 liquidity-providing HFT earns non-negative profits on each exchange. Thus, it does not matter how many exchanges a single HFT provides liquidity on because the equilibrium spread and depth would be the same.

¹⁴Miller and Shorter (2016) estimates that HFTs account around 55% trading volume in U.S. equity market. HFTs are just a subset of algorithm traders.
analysis I will assume $\gamma \geq \bar{\gamma}$. I will first calculate each exchange’s expected per unit time trading volume in the same way as in Proposition 3.

In the first two cases of Proposition 4, HFTs only provide liquidity on fast exchanges. Thus, in these two cases slow exchanges have expected trading volume zero. Fast exchange’s per unit time trading volume can be directly implied from the results in Proposition 3 (trading volume with homogeneous order processing speed). When $M^*(\delta_F) \leq K$, HFTs provide liquidity on $M^*(\delta_F)$ fast exchanges. When $M^*(\delta_F) > K$ and $\pi_S(s^*(\delta_F)/2, K, 1) < 0$, HFTs provide liquidity on all those $K$ fast exchanges. Therefore, we can define $M^*_F(K) = \min\{M^*(\delta_F), K\}$ as the equilibrium depth in these two cases (depth on fast exchanges). We conclude that a fast exchange has expected per unit time trading volume:

$$Q^*_F(K) = \lambda I \frac{1}{K} + \lambda J \frac{M^*_F(K)}{K} + \frac{\lambda U \lambda J}{2 \sum \lambda} \phi(\delta_F) \frac{1}{K} - \frac{1 - \phi(\epsilon) \lambda U \lambda J}{2 \sum \lambda} \frac{M^*_F(K) - 1}{K}$$

(1.19)

when $M^*(\delta_F) \leq K$ or $M^*(\delta_F) > K$ and $\pi_S(s^*(\delta_F)/2, K, 1) < 0$. This is directly implied from equation (1.11) by simply replacing total number of exchanges $M$ with $K$, equilibrium depth $M^*$ with $M^*_F(K)$ and order processing speed $\delta$ with $\delta_F$. This is because HFTs only provide liquidity on fast exchanges and the facts that the total number of fast exchanges is $K$, each has order processing speed $\delta_F$ and the equilibrium depth on fast exchange is $M^*_F(K)$.

Note that in these two cases, slow exchanges have zero trading volume. This result should not be interpreted literally. In my model, investor (or liquidity trader) only buy or sell one unit of the risky asset. Thus if an exchange does not have the best price quotes, its trading volume would be zero. Since in reality some liquidity traders trade multiple units, HFTs usually provide liquidity on multiple price levels on each exchange. Therefore, a more appropriate way to interpret this result is that when $M^*(\delta_F) \leq K$ or $M^*(\delta_F) > K$ and $\pi_S(s^*(\delta_F)/2, K, 1) < 0$ slow exchanges are less often to be at the top of the consolidated limit order book across all exchanges. In other words, the national best bid and ask prices quotes occur on slow exchanges less often. In current equity markets, large institutional traders usually split their large order to many small orders. Thus few trades will take orders from multiple price levels. If an exchange is not at the top of the consolidated limit order book often, its trading volume would be small. Thus in reality although slow exchange’s trading volume is not zero but it would be smaller than the trading volume on
fast exchanges. It would be better to model multiple price levels on limit order book. But it is extremely hard to work on. For simplicity, I only model the best bid and ask prices on the limit order book, which can clearly deliver the intuition of my main results. I summarize all these trading volume results in Proposition 5.

**Proposition 5** (Trading Volume with Exchange Speed Heterogeneity) If there are \( K \) fast exchanges with order processing speed \( \delta_F \) and \( M - K \) slow exchanges with order processing speed \( \delta_S \), where \( 1 \leq K \leq M - 1 \) and \( \delta_F < \delta_S \). When \( \gamma \geq \gamma^* \) and \( s^*(\delta_F)/2 < \sigma \), then for each \( K \):

(i) The ex ante expected per unit time trading volume on a fast exchange \( Q_F^*(K) \) is determined in equation (1.19) if \( M^*(\delta_F) \leq K \) or \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) < 0 \). Otherwise,

\[
Q_F^*(K) = \lambda_l\left[\frac{\gamma}{K} + \frac{1 - \gamma}{M^*(K)}\right] + \lambda_J + \frac{\lambda_U\lambda_l[\phi(\delta_F) + (1 - K)(1 - \phi(\epsilon))] + \lambda_U\lambda_l[1 - \phi(\epsilon')]\gamma M^*_S(K)}{2K\lambda\Sigma\lambda} = \frac{M^*_S(K)}{K - 1}(\frac{\lambda}{K} + \frac{\lambda_J}{M^*(K)}) + \frac{\lambda_U\lambda_l[1 - \phi(\epsilon')]\gamma M^*_S(K)}{2K\lambda\Sigma\lambda} \tag{1.20}
\]

(ii) The ex ante expected per unit time trading volume on a slow exchange \( Q_S^*(K) \) = 0 if \( M^*(\delta_F) \leq K \) or \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) < 0 \). Otherwise,

\[
Q_S^*(K) = \frac{M^*_S(K)}{M - K}\left[\frac{\lambda_l(1 - \gamma)}{M^*(K)} + \lambda_J\right]\left[1 - \frac{\lambda_U}{2\Sigma\lambda}[1 - \phi(\epsilon')]\right] \tag{1.21}
\]

where \( \phi(\delta_F) = 1 - e^{-(\lambda_l + \lambda_J)\delta_F} \), \( \phi(\epsilon') = \phi(\delta_F + \epsilon - \delta_S) = 1 - e^{-(\lambda_l + \lambda_J)\delta_F + \epsilon - \delta_S} \) and \( M^*(K) = M^*_F(K) + M^*_S(K) \) is the total equilibrium depth.

Each exchange’s expected per unit time trading volume are calculated exactly in the same way as in Proposition 3. Because undercutting HFT always submits her price improving order to fast exchange and HFTs always provide liquidity on fast exchange, trading volume on fast exchange is always larger than on slow exchange when trading speed heterogeneity exists. When speed upgrading technology is available, whether exchanges have incentives to increase their order processing speed depending on how much additional trading volume it could attract. Specifically, when all exchanges have the same slow order processing speed \( \delta_S \), denote \( Q^*(\delta_S) \) as each exchange’s per unit time trading volume which is determined in equation (1.11). If one exchange becomes fast with order processing speed \( \delta_F < \delta_S \), then the per unit time trading volume for this fast exchange is \( Q_F^*(1) \) determined in equation (1.19) or (1.20). All remaining \( M - 1 \) exchanges with slow order processing speed \( \delta_S \) have expected per unit time trading volume \( Q^*_S(1) \) determined in equation (1.21) (or zero). We have the following result:
**Corollary 2** When \( \gamma \geq \bar{\gamma} \) and \( s^*(\delta_S)/2 < \sigma \):

(i) \( Q^*_S(K) < Q^*(\delta_F) < Q^*_F(K) \) for all \( 1 \leq K \leq M - 1 \);

(ii) \( Q^*(\delta_S) < Q^*_F(1) \) if \( M^*(\delta_S) < M \) or \( \phi(\delta_S) \leq M\phi(\delta_F) + (M - 1)[1 - \phi(\epsilon)] \).

(i) shows that as long as exchanges have different order processing speeds, fast exchanges always have large trading volume than slow exchanges. (ii) shows that under some general conditions \( Q^*_F(1) \) is larger than \( Q^*(\delta_S) \), thus exchanges always have incentive to invest in speed technology providing the speed cost is not too high. When the equilibrium depth \( M^*(\delta_S) \) is smaller than \( M \), It is quite intuitive that \( Q^*_F(1) \) is always larger than \( Q^*(\delta_S) \). If HFTs do not provide liquidity on all exchanges, faster order processing speeds is one way exchange could use to attract liquidity-providing HFTs. Exchanges can increase its trading volume through faster order processing speed.

When \( M^*(\delta_S) = M \) the reason why the trading volume on a fast exchange \( Q^*_F(1) \) is not always larger than \( Q^*(\delta_S) \) is because the probability of sniping decreases on fast exchange. When undercutting HFT submits price-improving limit order to the fast exchange, the other liquidity-providing HFT on the fast exchange will cancel her stale limit orders \( \delta_F \) units time after undercutting HFT’s arrival. If \( \delta_F \) is too smaller than \( \delta_S \), stale limit orders remain on the fast exchange for a very short time. The probability of sniping on these stale limit order decreases, which reduces fast exchange’s trading volume. Therefore, trading volume on fast exchange may not increase if \( \delta_F \) is too smaller than \( \delta_S \). But as long as \( \delta_F \) is close to \( \delta_S \), thus condition \( \phi(\delta_S) \leq \min\{M\phi(\delta_F) + (M - 1)[1 - \phi(\epsilon)], 2\phi(\epsilon')\} \) in Corollary 2 always holds, then an exchanges can always increase its trading volume through faster order processing speed because \( Q^*_F(1) > Q^*(\delta_S) \).

Corollary 2 points out a very interesting result. Normally, one will think that the speed arms race among exchanges would stop when all exchanges are fast enough. As trading speed getting faster and faster, a new available speed technology may not increase current trading speed too much. In other words, \( \delta_F \) is not too smaller than \( \delta_S \) when trading is already fast enough. It is natural to think that exchange may not invest in speed technology anymore because it can

---

\(^{15}\)Note that when \( \phi(\delta_S) \leq (M - 1)(1 - \phi(\epsilon)) \) the second condition in Corollary 2 (ii) always holds. When \( \phi(\delta_S) > (M - 1)(1 - \phi(\epsilon)), \phi(\delta_S) \leq M\phi(\delta_F) + (M - 1)[1 - \phi(\epsilon)] \) is equivalent as \( \Delta_{speed} \leq \delta_S - \phi^{-1}\left(\frac{\phi(\delta_S) - (M - 1)[1 - \phi(\epsilon)]}{\delta_F}\right) \) where \( \Delta_{speed} = \delta_S - \delta_F \).
not significantly enhance its trading speed. Surprisingly, Corollary 2 points out that exactly when \( \delta_F \) is not too smaller than \( \delta_S \), exchanges actually have stronger incentive to invest in speed technology because an exchange’s trading volume always increase if it has faster order processing speed than other exchanges. The reason is because it attracts HFTs to submit price-improving limit orders while the probability of sniping on the fast exchange does not decrease significantly.

So far I have shown that when a new fast speed technology is available and the conditions in Corollary 2 holds, exchanges have incentives to invest in this new speed technology if the cost is not too high. In other words, all exchanges remain their current slow order processing speeds is not an equilibrium anymore. In the next subsection, I will endogenize exchanges speed investment decisions and study the welfare implications for long-term investors.

**Exchanges Speed Arms Race.** I will add one more stage before the trading game starts. Specifically, at stage \( t = -1 \) all exchanges have opportunity to upgrade their order processing speeds from \( \delta_S \) to \( \delta_F \) at per unit time cost \( C_{\text{speed}} \), where \( \delta_F < \delta_S \).\footnote{I use the same notations as in BCS for the cost of speed investment. While in BCS \( C_{\text{speed}} \) is the per unit time cost for high speed traders, here it is the exchange’s per unit time cost if it invests in high speed order processing technology. In my model, all traders have exactly the same speed and my focus is on exchange’s order processing and connection speeds, not trader’s speed.} For simplicity, I assume exchanges make their speed investment decisions simultaneously. This assumption does not matter for my analysis. Later I will show that under some general conditions investing in the new speed technology is a dominant strategy for all exchanges. I model the speed cost as per unit time cost for exchanges is because maintaining a high speed exchange is costly. Exchanges may need to rent more space for their matching engines, and may have higher operating cost (such as cooling cost). As a result, it is more appropriate to model the speed cost as per unit time cost for exchanges.

Broadly speaking, exchange’s revenues come from three main sources: per-trade transaction fee, data and connection fee, listing and other services fee. In equity markets, the current maker-taker fee model generates the main per-trade revenue for exchanges. Exchanges pay rebates to traders who add liquidity (submit limit orders which are not immediately executable) and charge access fee for taking liquidity (submitting market or marketable limit orders). These fees are per share based. I follow similar notations as in Colliard and Foucault (2012) and Chao et al. (2017) to
define \( \tilde{f}_m \) and \( \tilde{f}_t \) as maker and taker fee. The total maker taker fee is defined as \( \tilde{f} = \tilde{f}_m + \tilde{f}_t > 0 \). The rebate is paid only when transaction occurs, and exchanges earn \( \tilde{f} \) per share traded. If an exchange has large trading volume, its revenue from transactions would increase.

Usually, an exchange with large trading volume could generate more revenue from data feeding fee. For example, in U.S. equity market allocation of the revenues from selling consolidated data is positively related to an exchange’s market share of total trading volume (see more details from Caglio and Mayhew (2012)). Exchange with large trading volume also attract more listings due to the positive externalities of liquidity. As a result, it is safe to conclude that an exchange’s revenue is increasing in its trading volume. For simplicity, I only model the per share transaction fee to study exchange’s speed investment decision, which is enough to deliver the main intuition of exchanges speed investment arms race.

Fortunately, all previous results still hold under fixed maker-taker fee as long as all exchanges have the same fee structure. Only the determination of equilibrium spread and depth need to be adjusted according to the liquidity rebates. For conciseness, here I assume \( \tilde{f}_m = 0 \) and \( \tilde{f} = \tilde{f}_t > 0 \), which means that only liquidity takers pay the transaction fee. In this way, the equilibrium spread and consolidated market depth stay the same and we can directly use all previous results as long as sniping HFTs still earns positive profits after paying liquidity taking fee. Now, we can define fast and slow exchange’s per unit time profit as:

\[
\pi_F(K) = (\tilde{f}_m + \tilde{f}_t)Q^*_F(K) - C_{speed}; \quad \pi_S(K) = (\tilde{f}_m + \tilde{f}_t)Q^*_S(K)
\]

Where \( \pi_F(K) (\pi_S(K)) \) denotes a fast (slow) exchange’s per unit time profit where there are \( K \) fast exchanges. Exchanges are trying to maximize their per unit time profit when make speed investment decisions. The equilibrium results are presented in the following proposition.

**Proposition 6** (Exchanges Speed Arm Races) When \( \gamma \geq \overline{\gamma}, s^*(\delta_S)/2 + \tilde{f}_t < \sigma \) and \( \phi(\delta_S) \leq M\phi(\delta_F) + (M - 1)[1 - \phi(\epsilon)] \), then for given \( \tilde{f} = \tilde{f}_t > 0 \):

(i) If \( \frac{C_{speed}}{\tilde{f}} < \min\{Q^*_F(1) - Q^*(\delta_S), Q^*(\delta_F) - Q^*_S(1)\} \), investing in the fast speed technology is a

---

\(^{17}\)For example, if exchanges pay 0.2 cents per share rebates for adding liquidity and charge 0.3 cents per share for taking liquidity we have \( \tilde{f} = 0.001 \), \( \tilde{f}_m = -0.002 \) and \( \tilde{f}_t = 0.003 \). Note that, \( \tilde{f}_m \) could also be positive. In this case, exchanges charge positive fee for providing liquidity while pay rebates for taking liquidity. This is called the “inverted” maker-taker pricing model currently adopted by Nasdaq BX and Bats BYX exchange.
dominant strategy for all exchanges;
(ii) If \( \frac{C_{\text{speed}}}{f} > Q^*(\delta_F) - Q^*(\delta_S) \), each exchange’s per unit time profits decrease when all exchanges speed up;

(iii) If \( s^*(\delta_F) = s^*(\delta_S) \) and \( M^*(\delta_F) > M^*(\delta_S) \), investor’s welfare (equation (1.10) minus taker fee) decreases when all exchanges speed up.\(^{18}\)

Proposition 6 shows that when the conditions in Corollary 2 holds and the speed cost is not too large, all exchanges will invest in high speed technology. This is not necessarily beneficial for exchanges. When all exchanges speed up, each exchange’s per unit time trading volume can decrease (it is possible that \( Q^*(\delta_F) < Q^*(\delta_S) \)), let along their profits. But if an exchange has slower order processing speed than other exchanges, it will loss trading volume significantly. This is why all exchanges have to make sure they have the current fastest order processing speed although it may not increase their profits.

Proposition 6 (iii) shows that when all exchanges speed up, investor welfare is not necessarily improved. This result shares the same intuition as in Proposition 2 (iii). When exchanges increase their order processing speeds by about the same amount, the overall trade-through rates can increase. Moreover, it is a common practice that institutional traders split their large orders to many small orders. As a result, only few trades will actually move price in the consolidated limit order book, which suggests that the cost of high trade-through rates could potentially be a significant portion of investor’s transaction cost.

One limitation of Proposition 6 is that maker-taker fee are exogenous. It would be definitely better to endogenize exchange’s fee structure. Modeling exchanges’ maker-taker fee competition is extremely complicated. Chao et al. (2017) has studied this question in a simple one round trading model without adverse selection cost. Even in their simple setup, no pure strategy equilibrium exists and the mixed strategy equilibrium is very complicated because it features two dimensions of the maker-taker fee distribution. But one lesson learned from their analysis and is important for us is that competition will never drive the total maker-taker fee to zero. Thus, trading volume still matters and the results in Proposition 6 would still be relevant even we endogenize exchange’s fee structure.

---

\(^{18}\)Investor’s welfare defined in equation (1.10) is the one without maker-taker fee. Since investors always take liquidity in my model, under maker-taker fee investor’s welfare is \( W(Buy|\delta_l = \delta_S, \epsilon_{ij} = \epsilon) - \bar{f}_i \).
1.4 Empirical Analysis

I show that slow exchanges lose trading volume to fast exchanges because liquidity providers prefer to submit their price-improving orders to fast exchanges. For this to occur, it needs: 1) the stock’s bid-ask spread, the difference between the lowest quoted sell price and the highest quoted buy price, to be larger than one tick; and 2) the Order Protection Rule to be present.

I provide two empirical tests that support this prediction. IEX is the only slow exchange due to its built in 350-microsecond delay. My first test investigates how IEX’s daily market shares of total trading volume of the stocks included in the recent Tick Size Pilot Program would change by exploring the exogenous increase in tick size. My second test examines how IEX’s monthly market shares change cross-sectionally after it became a public exchange in September 2016. Previously, IEX was a dark pool that did not publicly display quoted price and thus orders on IEX were not protected by the Order Protection Rule. After it became a public exchange, if IEX has better prices, the other exchanges must route their customers’ orders to IEX to comply with the Order Protection Rule. My model’s prediction is tested by whether IEX can attract price-improving orders or not after it became a public exchange.

1.4.1 Data Description

The Tick Size Pilot is a data-driven test to evaluate whether widening the tick size for securities of smaller capitalization companies would impact liquidity of those securities. The pilot consists of a randomly chosen control group and three test groups, with each test group having approximately 400 securities.

The first test group will be quoted in $0.05 increments, but will continue to trade at their current price increment. The second test group will be quoted and traded in $0.05 minimum increments, but would allow certain exemptions for midpoint executions, retail investor executions, and negotiated trades. The third test group will adhere to the requirements of the second test group, but will also be subject to a “trade-at” rule requirement, which requires off-exchange trading venues to offer significantly price improvement (i.e., one tick) to quoted price on public exchanges. The three treatment groups were gradually implemented on October 3rd to 31st, 2016. The pilot pro-
gram lasts for two years.

My sample period for the Tick Size Pilot test is September 2nd to December 30th, 2016. On September 2nd, 2016, IEX had fully transited from a dark pool to a public exchange. My test includes all stocks in the pilot program. IEX’s daily market share of total trading volume is calculated from the daily Trade and Quote (TAQ) data. Other variables such as daily closing price, share turnover, and market capitalization are drawn from the Center for Research in Security Prices (CRSP) data. Summary statistics of main variables are reported in Table 1.

My sample period in the second test is from June 1st to November 30th, 2016. Before September 2nd, 2016, IEX was a dark pool, and its trading volume in each stock is from the Rule 605 data downloaded from IEXfis website because TAQ does not report each dark pool’s trading volume. Since Rule 605 data is reported monthly, IEX’s market share of total trading volume in my second test is calculated monthly too. Other control variables are calculated from TAQ and CRSP, and are reported monthly too. I include all stocks reported in IEX’s Rule 605 report excluding stocks in the Tick Size Pilot Program and stocks with missing data.

One empirical challenge is to identify the volume of IEX before it became an exchange. TAQ data only separates volume across different stock exchanges. Before IEX became an exchange, its trading volume is under the category called trade report facilities (TRFs) with other non-exchange trading venues. It is impossible to compare the trading volume of IEX before and after it became to a public exchange. Fortunately, I am able to compile a proxy for the IEX volume using the SEC Rule 605 data.

SEC 605 data is well-known for comparing execution quality such as quoted spread and effective spread. However, SEC 605 data also includes the number of shares as the base to calculate this execution quality measure. In the United States, every trading venue needs to fill in the SEC 605 report, even if it is a dark pool. This feature allows me to construct the volume measure before IEX became a public exchange.

SEC Rule 605, formerly known as SEC 11Ac1-5 rule, requires market centers to disclose execution quality statistics on a monthly basis. Thus, I am able to observe the order execution in IEX before it becomes public exchange. See Bennett and Wei (2006), and Goldstein et al. (2008) for their study on market quality using the SEC 605 filing.
1.4.2 Empirical Results

**Tick Size Pilot Test.** In my model, faster exchanges attract more undercutting HFTs. However, when the bid-ask spread binds at one tick, no HFT can undercut the current quotes. As a result, the trading volume on a faster exchange for such stocks would increase by less than that for stocks where the price tick was less binding. To test this prediction, I exploit the tick size pilot program introduced by the SEC in October 2016. This pilot experiment increased tick size for 1,200 randomly selected stocks with small capitalizations from 1 cent to 5 cents. Since IEX has a slower order processing speed than other exchanges, with its 350 microseconds delay, my model predicts that IEX’s market share of total trading volume in those stocks with 5 cents tick size should increase. I test this prediction by running the following difference in differences test:

\[ y_{it} = \beta (Post_t \times Pilot_i) + X_{it}' \delta + \text{Stock FE}_i + \text{Time FE}_t + \epsilon_{it} \]  

(1.22)

\( y_{it} \) is IEX’s daily market share of trading volume on each stocks defined as the stock’s trading volume on IEX over total trading volume across all trading venues. Post and Pilot are two dummy variables for post treatment period and treatment stocks. \( X_{it}' \) are other control variables including the reverse of daily closing price, natural log of daily share turnover and natural log of market capitalization. I add both stock and time fixed effect. The estimation results are presented in Table 2. The coefficients on the Post×Pilot are positive and highly significant in all three treatment groups. Comparing to IEX’s market share in September 2016, its average market share in those treatment stocks increases by around 13 percent (from 1.77% to 2%). The third treatment group has the largest effect because more trading volume is driven from alternative trading systems (ATS) to public exchanges due to the “trade-at” rule.

**The Test of IEX Switching from Dark Pool to Public Exchange.** My model predicts that price improving orders will be submitted to exchanges with high order processing speeds. Thus, a slow exchange does not have a competitive advantage when price improvement is possible, and the Order Protection Rule is present. On September 2nd, 2016, IEX became the 12th public exchange. Meanwhile, IEX has slower order processing speed. My model predicts that IEX will attract less trading volume in those stocks with larger bid-ask spread relative to stocks with

35
binding spreads. I test this prediction by running the following test:

\[ y_{it} = \beta (Post_t \times NonBinding_i) + X'_{it} \delta + \text{Stock FE}_i + \text{Time FE}_t + \epsilon_{it} \quad (1.23) \]

The dependent variable \( y_{it} \) equals to IEXfis reported trading volume in Rule 605 Data divided by CRSP recorded total volume over all venues. I define \( NonBinding_i = 1 \) for those stocks with average effective spread larger than 1.25 cents during March 2016 to May 2016, (i.e. three months before the sample period). The tick Size Pilot related stocks are removed. \( Post_t \) equals one after September 1st, 2016, and zero otherwise. Thus, \( Post_t = 1 \) indicates that IEX is a public exchange. Covariates include natural log of Market Cap and monthly share turnover, and the reverse of nominal price which controls the relative tick size.

The regression result reported in Table 3 shows that after becoming public, IEX gained 0.17 percentage points more in binding stocks than non-binding stocks. This is consistent with my prediction that IEX can hardly attract price improvement market makers when tick size is not binding.\(^{20}\) The result is robust under various controls and fixed effects.

1.5 Conclusions

Over the past decade, trading at unfathomably high speeds has come to dominate U.S. equity markets. It is easy to understand why traders want to invest in technologies that allow them to trade at high speeds. Faster traders can exploit mispriced orders from slow traders by crossing against them, and they can withdraw their own mispriced orders before they themselves are exploited. It is less clear why exchanges want to process orders more quickly, but do not want to invest in increased connection speed between exchanges.

In this paper, I show that the Order Protection Rule, which requires an exchange to route its customers’ orders to other exchanges with better prices, is a key driver of stock exchanges’ competition on order-processing speeds. In particular, fast order-processing speeds attract more liquidity provision and, hence, more trading volume. I then show that when all exchanges in-

\(^{20}\) IEXfis overall market share increased because market makers would be happy to quote on an extra lit market for various reasons (Foucault and Menkveld (2008); Yao and Ye (2017)). However, those non-binding stocks do not benefit from these channels due to IEX’s slow order processing speed.
crease their order-processing speeds, it can harm investor welfare by increasing the probability of trade-through. By contrast, I show that increasing the connection speeds between exchanges can significantly increase investor welfare, but exchanges nonetheless prefer slow connection speeds. This is because, slower connection speeds reduce competition between exchanges, raising an exchange’s trading volume. As a result, stock exchanges do not necessarily compete on liquidity-enhancing dimensions. I provide two empirical tests of the theory. These tests support the prediction that slow exchanges differentially lose trading volume to fast exchanges that attract more price-improving orders when the bid-ask spread is less likely to bind.

For simplicity, the current model assumes exogenous exchanges’ fee structures. Consequently, exchanges maximize per unit time profit corresponds to maximize per unit time trading volume. Although trading volume is a good proxy for an exchange’s goal, a model that combines exchanges’ competition on speeds and fee structures merits further research.
Table 1.1: Summary Statistics for Tick Size Pilot Test

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Control Group (1168 Stocks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEX Market Share</td>
<td>94.414</td>
<td>1.78</td>
<td>3.00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Closing Price</td>
<td>94.398</td>
<td>23.82</td>
<td>28.22</td>
<td>0.365</td>
<td>485.6</td>
</tr>
<tr>
<td>Share Turnover</td>
<td>94.398</td>
<td>231.171</td>
<td>556.946</td>
<td>1</td>
<td>51,220,000</td>
</tr>
<tr>
<td>Market Cap</td>
<td>94.398</td>
<td>720.5</td>
<td>759.9</td>
<td>3.861</td>
<td>4129</td>
</tr>
<tr>
<td><strong>Panel B: Treatment Group 1 (393 Stocks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEX Market Share</td>
<td>31.765</td>
<td>1.89</td>
<td>2.91</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Closing Price</td>
<td>31.760</td>
<td>23.90</td>
<td>23.71</td>
<td>1.050</td>
<td>172.6</td>
</tr>
<tr>
<td>Share Turnover</td>
<td>31.760</td>
<td>237.763</td>
<td>560.753</td>
<td>7</td>
<td>50,260,000</td>
</tr>
<tr>
<td>Market Cap</td>
<td>31.760</td>
<td>711.2</td>
<td>756.1</td>
<td>4.048</td>
<td>3776</td>
</tr>
<tr>
<td><strong>Panel C: Treatment Group 2 (396 Stocks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEX Market Share</td>
<td>31.602</td>
<td>1.90</td>
<td>2.96</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Closing Price</td>
<td>31.599</td>
<td>23.56</td>
<td>23.26</td>
<td>1.250</td>
<td>203.8</td>
</tr>
<tr>
<td>Share Turnover</td>
<td>31.599</td>
<td>220.863</td>
<td>529.772</td>
<td>1</td>
<td>47,750,000</td>
</tr>
<tr>
<td>Market Cap</td>
<td>31.599</td>
<td>699.6</td>
<td>729</td>
<td>5.471</td>
<td>4019</td>
</tr>
<tr>
<td><strong>Panel D: Treatment Group 3 (390 Stocks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEX Market Share</td>
<td>31.470</td>
<td>1.95</td>
<td>2.82</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Closing Price</td>
<td>31.468</td>
<td>24.77</td>
<td>39.19</td>
<td>1.100</td>
<td>542.0</td>
</tr>
<tr>
<td>Share Turnover</td>
<td>31.468</td>
<td>247.085</td>
<td>622.077</td>
<td>2</td>
<td>34,880,000</td>
</tr>
<tr>
<td>Market Cap</td>
<td>31.468</td>
<td>732</td>
<td>772.5</td>
<td>5.276</td>
<td>4127</td>
</tr>
</tbody>
</table>

Note: IEX market share is defined as a stock’s daily trading volume on IEX over total trading volume across all trading venues. Trading volume data are from daily TAQ. Daily closing price, share turnover and market cap for each stocks are downloaded from CRSP. Market cap is measured in millions of dollars. Sample period is from September 2nd, 2016 to December 30th, 2016.
Table 1.2: Impact of Tick Size on IEX’s Market Share of Trading Volume

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Groups</td>
<td>All Groups</td>
<td>Group 1</td>
<td>Group 2</td>
<td>Group 3</td>
</tr>
<tr>
<td>Pilot×Post</td>
<td>0.23***</td>
<td>0.23***</td>
<td>0.20***</td>
<td>0.19***</td>
<td>0.330***</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.0590)</td>
</tr>
<tr>
<td>ln (Share Turnover)</td>
<td>0.16***</td>
<td>0.07***</td>
<td>0.07***</td>
<td>0.0923***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.0224)</td>
<td></td>
</tr>
<tr>
<td>Inverse of Share Price</td>
<td>-0.03</td>
<td>-0.12</td>
<td>0.06</td>
<td>-0.00710</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.30)</td>
<td>(0.31)</td>
<td>(0.30)</td>
<td>(0.319)</td>
<td></td>
</tr>
<tr>
<td>ln (Market Cap)</td>
<td>0.02</td>
<td>-0.04</td>
<td>-0.01</td>
<td>-0.0262</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.112)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>189,225</td>
<td>189,225</td>
<td>126,158</td>
<td>125,997</td>
<td>125,866</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.004</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>Number of Stocks</td>
<td>2,347</td>
<td>2,347</td>
<td>1,561</td>
<td>1,564</td>
<td>1,558</td>
</tr>
<tr>
<td>Time FE</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Stock FE</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Note: the above table reports estimations results from the regression that $y_{it} = \beta(\text{Post}_t \times \text{Pilot}_i) + \mathbf{X}'_i \delta + \text{Stock FE}_i + \text{Time FE}_t + \epsilon_{it}$, where $y_{it}$ is IEX’s daily market share of trading volume in each stock in the Tick Size Pilot Program, and defined as the stock’s trading volume on IEX over total trading volume across all trading venues. Post and Pilot are two dummy variables for post treatment period and treatment stocks in the Tick Size Pilot Program. $\mathbf{X}'_i$ are other control variables including the reverse of daily closing price, natural log of daily share turnover and market capitalization. Time period is from September 2nd, 2016 to December 30th, 2016. Robust standard errors in parentheses. *** \( p<0.01 \), ** \( p<0.05 \), * \( p<0.1 \).
Table 1.3: Impact of Switching to Public Exchange on IEX’s Market Share of Total Trading Volume

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) IEX Market Share</th>
<th>(2) IEX Market Share</th>
<th>(3) IEX Market Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonBinding×Post</td>
<td>-0.16***</td>
<td>-0.17***</td>
<td>-0.17***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>ln (Share Turnover)</td>
<td>-0.25***</td>
<td>0.36***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.07)</td>
<td></td>
</tr>
<tr>
<td>Inverse of Share Price</td>
<td>-0.13***</td>
<td>-0.07**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td></td>
</tr>
<tr>
<td>ln (Market Cap)</td>
<td>0.10***</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.02)</td>
<td></td>
</tr>
</tbody>
</table>

Observations: 30,493 30,493 30,493
R-squared: 0.033 0.140 0.604
Number of Stocks: 9,452 9,452 9,452
Time FE: NO NO YES
Stock FE: NO NO YES

Note: the above table reports estimations results from regression \( y_{it} = \beta (\text{NonBinding}_i \times \text{Post}_t) + X'_{it} \delta + \text{Stock FE}_i + \text{Time FE}_t + \epsilon_{it} \), where \( y_{it} \) is IEX’s monthly market share of trading volume on each stocks defined as the stock’s trading volume on IEX over total trading volume across all trading venues. \( \text{Post}_t \) and \( \text{NonBinding}_i \) are two dummy variables. \( \text{Post}_t = 1 \) for September, October and November of 2016, when IEX is a public exchange. \( \text{NonBinding}_i = 1 \) for stocks with monthly average effective spreads exceed 1.25 cents. \( X'_{it} \) are other control variables including the reverse of average monthly closing price, natural log of monthly share turnover and market capitalization. Sample period is from June, 2016 to November, 2016. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.
In the above picture, $\delta$ denotes exchange’s order processing speed. $\zeta$ is the time it takes for a HFT to send information from one exchange to another exchange. $\epsilon$ is the time it takes for an exchange to send its order updating information or route orders to other exchange. $\eta$ is the time it takes for an investor to receive any updates from exchanges through securities information processor (SIP). These latencies depend on specific exchanges. NYSE is located at Mahwah NJ. Nasdaq is located at Carteret NJ. BATS and IEX are in Equinix NY4/NY5 data center which is located at Secaucus NJ. At the cutting edge of technology $\delta$, $\zeta$, $\epsilon$ and $\eta$ are around 50, 100, 300 and 1000 microseconds.
I run regression $y_{it} = \alpha + \sum_{j\neq 1} \beta_j(Pilot_i \times I(t = j)) + X_{it}'\delta + \text{Stock FE}_i + \text{Time FE}_t + \epsilon_{it}$ and plot all estimated $\beta$ in above graph. The gray dash lines are the 90% confidence interval. $I(t = j)$ is a time dummy and is equal to one if it is at day $j$. The left red line at October 3, 2016 indicates the starting day of the pilot phase in period. The right red line indicates the pilot phase in ending time at October 31, 2016. After that, all treatment group stocks become active.
Chapter 2

Can Stock Exchanges with Speed Bump Designs Survive?

2.1 Introduction

The fact that high-frequency traders (HFTs) can observe market movements before slow traders has generated significant controversy and attention in the popular press, as in Michael Lewis’s book, Flash Boys. Several exchange design responses have been proposed to reduce the speed advantage of HFTs and hence their incentives to engage in a costly speed “arms race” that transfers resources away from liquidity consumers and may reduce liquidity. Budish et al. (2015) (henceforth BCS) suggests switching from the current continuous-trading process to a discrete time-batch trading process in which, for example, an exchange might run sealed-bid double auctions every 100 milliseconds to organize trades.¹ The Investors Exchange (IEX) intentionally delays all incoming orders and messages by about 350 microseconds. The Chicago Stock Exchange (CHX) has proposed a similar design that creates a 350-microsecond delay for those who trade against resting orders. A common question raised in debates on these new exchange designs is: Without any regulation, can an exchange that implements these designs survive when competing against other faster exchanges?²

The contribution of this chapter is to build a trading model with multiple exchanges to investigate this issue. My model builds on BCS. In BCS, time is continuous, and a security is traded on a single exchange. A Poisson arrival of liquidity traders seeks to buy or sell one unit of the asset to balance their portfolios, and a Poisson arrival of public information about the asset value shifts its value up or down. Fast HFTs choose whether to provide liquidity by posting limit orders, or

---

¹More details about this design can be found in Budish et al. (2014).
²In his 2017 AEA/AFA joint luncheon address, Eric Budish has discussed some issues on how frequent batch-auction exchanges compete with traditional limit-order book exchanges. More details can be found at https://www.aeaweb.org/webcasts/2017/luncheon.php.
hold back and wait for a public information arrival that will make non-cancelled quotes stale so they can make profits by sniping. The equilibrium bid-ask spread breaks even in expectations and trade-off profits from the liquidity provision with profits from sniping, leaving HFTs neutral as liquidity providers or snipers. HFTs racing to snipe or cancel stale limit orders encourages engagement in a speed arms race. HFTs using their speed to snipe stale orders increases the cost of the liquidity provision. The heightened liquidity costs are at the root of innovative new exchange designs aiming to mitigate sniping and reduce the speed advantage of HFTs.3

To this model, I add three new features. I first allow the single security to be traded on \( M(>1) \) exchanges. As in Chapter 1, each exchange uses the continuous limit order book to conduct trades. I will allow one exchange to implement frequent batch auctions or order delay designs, and examine how it competes with the remaining \( M-1 \) exchanges. Exchanges maximize their per-unit-time profit (or equivalently their per-unit-time trading volume) given a fixed per-share transaction fee. The second real-world feature that I integrate is a discrete price tick, so that equilibrium is pinned down by the number of HFT liquidity providers at equilibrium price ticks, rather than being a smooth adjustment in price. Currently in the U.S. equity market, the tick size or minimum price movement is 1 cent for stocks with prices above $1 per share. This matters because I also introduce the Poisson arrival of undercutting HFTs who offer a one-tick price improvement of the equilibrium quotes by other HFTs.4 This price undercutting captures the fact that HFTs who provide liquidity often need to adjust their quotes due to changes in their inventory positions or risk capacities. The stage trading game ends whenever the trade occurs and the next stage begins, creating a stationary framework.

My model offers several implications for how to design exchanges that mitigate sniping of the liquidity provider, reduce sniping results, improve liquidity and reduce the bid-ask spread. First, consider the frequent batch auctions (FBA) proposed in BCS. FBA can eliminate sniping if all information about asset value jumps reaches everyone at exactly the same time, as in Chapter

---

3 Another common term used in debates on these new exchange designs is “latency arbitrage,” which refers to the possibility that HFTs can explore the disparities of price quotes for the same asset on different exchanges to make profits. Latency arbitrage is a special example of sniping in BCS. A more detailed explanation about latency arbitrage can be found in Wah and Wellman (2013).

4 For instance, if the current bid (highest buy) price=$10.00 and ask (lowest sell) price=$10.05, the undercutting HFT is willing to sell at $10.04 or buy at $10.01. In BCS, liquidity-providing HFTs are homogeneous, and their model does not have undercutting HFTs.
1. When the liquidity provider’s cancellation message and other HFTs’ sniping orders arrive at the FBA exchange simultaneously, the exchange will always process cancellation messages first and then determine transaction prices and quantities from the remaining orders. As a result, liquidity providers can always successfully cancel their orders whenever the asset value jumps, and so sniping cannot occur in FBA when all traders observe the asset value jumps at exactly the same time.

In practice, however, information about value jumps does not always immediately become public information in this way. In particular, sometimes one trader will get the news before the others. In this case, the information about value jumps is private information, and the informed trader will play the role of the traditional informed trader, as in Kyle (1985) or Glosten and Milgrom (1985). In such a scenario, the liquidity provision costs for HFTs have two components: the traditional adverse selection cost on private news and the sniping cost on public news. Whether FBAs can thrive depends on the proportions of asset value jumps that are public or private news.

I show that FBAs improve liquidity only when most information arrival is public, reaching everyone at the same time. Otherwise, the FBA designs reduce liquidity: although FBAs eliminate sniping costs, they increase the traditional adverse-selection costs for liquidity-providing HFTs. This is because there is no pre-trade price transparency in FBAs. When an undercutting HFT arrives, the liquidity-providing HFTs want to adjust or cancel their quotes to reduce exposure to informed traders on private news. But in FBAs, HFTs cannot observe an undercutting HFT’s order, which increases the traditional adverse-selection costs for the liquidity-providing HFTs due to the increased exposure of their orders to the informed trader. As a result, when sniping is modest relative to the traditional adverse-selection costs, those costs dominate the gains from

---

5Another way to interpret this is to think of the risky asset’s value including two components: soft and hard information components. Soft information is usually related to inside news about the underlying firm. Only slow human traders may access soft information. In contrast, hard information is machine-readable information correlated with the asset’s value, such as any public macroeconomic news announcement or price information about related assets. Naturally, high-frequency trading firms are tracking and good at analyzing hard information. This soft and hard information dichotomy was first proposed by Petersen (2004) and has been applied in Jovanovic and Menkveld (2016) too.

6In frequent batch auctions, if all remaining orders are announced before the next trading round, traders with extremely fast speed have incentives to manipulate the market by submitting some limit orders and then canceling them at the last moment, before the exchange crosses the buy and sell limit orders. This will encourage speed competition among traders. For this reason, frequent batch auctions are designed in BCS without this pre-trade price transparency.
eliminating sniping. In other words, the FBA design reduces liquidity when most information arrivals reach some HFTs before others.

I show that when FBAs are implemented, exchanges have incentives to compete on trading frequency, because liquidity provision costs are lower on FBAs that have more frequent auctions. On more frequent batch auctions, liquidity-providing HFTs can observe quoting status and adjust their orders more frequently, which reduces their exposure to informed trading on private news. This causes the batch nature of the design to unravel. This result can explain why FBAs have not been implemented by any exchange despite the endorsement of BCS.

Among the different delay designs, I establish that a design that only delays liquidity-taking orders like CHX attracts more trading volume than one like that of IEX, which delays all incoming orders. This is because undercutting HFTs prefer to submit their price-improving limit orders to exchanges with faster order-processing speeds. My analysis suggests that the 350-microsecond delay for liquidity adding orders may make undercutting HFTs less likely to submit orders to IEX. This may explain why IEX’s market share of the total equity trading volume remained at 2% after it became a public exchange in September 2016.

Finally, I show that even when sniping is a significant problem for some stocks, exchanges with a large market share of the total trading volume may lack incentives to implement frequent batch auctions or order delays, even when these innovative designs could improve long-term investor welfare. This is because when sniping occurs, exchanges share revenues with the sniping HFTs. These exchanges will lose this trading volume and revenues if these order-delay or FBA designs are implemented. Therefore, the interests of the exchanges may not be aligned with those of long-term investors with regard to how they value designs that alleviate sniping.

**Related literature.** “Speed bump” designs such as frequent batch auctions and order delays all require slowing down exchanges’ order-processing speed. I will only discuss those papers which are related to exchanges’ order-processing speed or latency. Menkveld and Zoican (2016) studies the impacts of exchanges’ order-processing speed on the liquidity in a single-exchange setup. They find that increasing speed does not necessarily improve liquidity. Ye et al. (2013) finds that Nasdaq’s 2010 speed upgrade did not improve its liquidity. Pagnotta and Philippon (2016)

---

7There are many studies on HFTs, especially after the 2010 flash crash. Menkveld (2016) provides a good review of this topic.
studies trading venues’ speed competition. The speed differences in their models range from seconds to minutes. Traders prefer faster trading venues because their gains from trading can be realized earlier. Their models are well suited to studying speed differences like over-the-counter markets versus public exchanges. Quote-updating speed plays a key role in my model, while the speed of trades is crucial in Pagnotta and Philippon (2016). Thus, the results in this chapter are complementary to theirs. BCS suggests implementing frequent batch auctions (FBAs), and Du and Zhu (2017) finds that seconds or minutes per auction can improve allocation efficiency. This chapter can shed some light on why no exchange is currently trying to implement FBAs, although they might increase efficiency.

2.2 Model Setup

The model framework is similar to that in Chapter 1. The new feature is that the signal of the asset’s value jumping is not necessarily observable to all traders at exactly the same time. Some high-frequency traders may observe the signal earlier than others. An alternative way to interpret this is that the signal of the asset’s value change is related to the underlying firm’s inside news. Only some traders might be informed about this private news. The goal is to add the traditional adverse-selection component to the cost of the liquidity provision. The details of the model are as follows.

**Exchanges and limit order book.** $M$ exchanges use continuous limit order book to conduct trades, and among them one exchange consider to implement frequent batch auctions or several order delay designs. Traders can use either market or limit orders to trade.

**Traders.** There are infinite number of risk neutral HFTs choosing whether to post limit orders on exchanges to provide liquidity to fundamental investors who arrive randomly. Fundamental investors attach an exogenous intrinsic value to trade, reflecting, for example, a need to re-balance their portfolios. Fundamental investors include mutual funds, pension funds and retail traders.

**Price grids.** The smallest price increment or tick size is given by $d > 0$. In current equity mar-
kets, the tick size is one cent for stocks with price above $1 per share. Let $\mathcal{P} = \{p^i\}_{i=-\infty}^{\infty}$ denote the discrete set of available prices for quoting and trading: the distance between any two consecutive prices in $\mathcal{P}$ is $d$.

**Timing and asset.** Time runs continuously on $[0, \infty)$. There is a single risky asset that is traded on all $M$ exchanges and one risk-free numeraire asset with price normalized to be 1. At the beginning of the trading game, the risky asset has an expected value of $v_0$. To ease presentation, I assume $v_0 = (p^i + p^{i+1})/2$ for some $i \in \mathbb{N}$, i.e., $v_0$ is at the midpoint of a price grid. At $t = 0$, HFTs choose the exchanges on which they post their limit orders. Then, three events may occur:

1. An fundamental investor with intrinsic value $\bar{\theta}$ to trade may arrive. I assume that the arrival time is exponentially distributed with intensity parameter $\lambda_I$. Upon arrival, the investor will buy or sell one unit of the risky asset with equal probability and only use market orders. A buyer arriving at time $t$ and paying $y_t$ to buy one unit of the risky asset has utility or welfare $w_t = v_t - y_t + \bar{\theta}$, where $v_t$ is the risky asset’s value at time $t$. A seller’s welfare is defined in a similar way. Further, I assume there are $\gamma$ portions of investors are sophisticated investors. These investors will consider potential price improvements when choosing which exchange to trade although all exchanges may have the same observed quoted price. Other $1 - \gamma$ portions are unsophisticated investors. Upon arrival, they will randomly chose one exchange having the current best price quotes to trade with equal probability. The portion of sophisticated investors only matters when there is heterogeneity in exchanges’ order processing speeds because the probability of each exchange offering potential price improvements might be different.

2. Before fundamental investors arrive, a signal related to the risky asset’s common value may arrive. I assume the arrival time of this signal is given by an exponential distribution with intensity parameter $\lambda_J$. With probability $\mu$ this signal is only observable to one high frequency trader. And with probability $1 - \mu$, the signal is publicly observable by all traders at exactly the same time. With equal probability it is a good or bad signal. Conditional on good signal, the risky asset’s common value will increase by $\sigma = kd$ for some $k \in \mathbb{N}$. Similarly, if it is a bad signal, the risky asset’s common value will decrease by $\sigma$. If $\sigma$ exceeds the current half bid-ask spread, those HFTs who have posted limit orders at exchanges will run to cancel their stale limit orders while other HFTs will try to trade at the stale price.
3. Alternatively, after HFTs post their limit orders on exchanges, an undercutting HFT may arrive who will offer a one tick price improvement of the current prices quoted by other HFTs. The arrival time of undercutting HFTs is given by an exponential distribution with intensity parameter $\lambda_U$. Upon arrival, with equal probability the undercutting HFT will submit (a) a limit buy order with price one tick above the current bid price; or (b) a limit sell order with price one tick below the current ask price. If at the time when the undercutting HFT arrives the bid-ask spread is binding at one tick, undercutting HFT will not post any order.

The arrival process for the fundamental investor, public information, and undercutting HFT all assumed to be independently distributed. Figure 2.1 draws the event timeline of one stage trading game. The conditional probabilities of each event is shown in the graph. The stage trading game ends whenever trade occurs, at which point the next stage begins.

**Figure 2.1: Trading Stage Event Time line**

**Exchanges Order Processing and Connection speeds.** Let $\delta_i$ be the amount of time that it takes for exchange $i$ (for $i = 1, 2, \cdots, M$) to process an incoming order or cancellation message. A small $\delta_i$ indicates a faster order processing speed. I denote the time that it takes to send price information between exchange $i$ and exchange $j$ by $\epsilon_{ij}$, where $\epsilon_{ij} = \epsilon_{ji}$. A smaller $\epsilon_{ij}$ indicates faster connection speeds between exchanges.
2.3 Equilibrium Analysis

The 2010 Flash Crash has brought high-frequency trading (HFT) under the spotlight of many academic researches and public debates. BCS finds that as long as trading is continuous, the opportunity of sniping can not be competed away, which increases liquidity provision cost. Foucault et al. (2017) finds that latency arbitrage (one example of sniping) is also common in foreign exchange markets.

Several new exchange designs have been proposed to mitigate sniping and reduce HFTs speed advantage. BCS suggests switching from the current continuous trading process to a discrete time batch trading process. For instance, run sealed bid double auctions for every 100 milliseconds to conduct trades. The Investors Exchange (IEX) delays all incoming orders by 350 microseconds and it was approved by Securities and Exchanges Commissions (SEC) to be a public exchange in June 2016. Chicago Stock Exchange (CHX) has proposed a similar design that only creates a 350-microsecond delay for those who trade against resting orders on CHX. This proposal is still seeking approval from SEC.

In this section, I will study which proposal might be the most effective one to mitigate sniping, and how an exchange implementing these new designs competes against traditional exchanges for trading volume. I will first discuss the frequent batch auctions (FBAs) design and then several order delay proposals. But before that, I will relax one assumption in the baseline model setup. Specifically, in Section 1.3 the signal jumping related to the asset’s value immediately becomes public information and all traders can observe it at exactly the same time. However, it is impossible that information about value jumps immediately becomes public information in this way. Sometimes, one trader will get the news before others. To account for this, I assume that with probability $\mu$ a single trader receives this information before everyone else. As a result, she is the single informed trader as in Kyle (1985) or Glosten and Milgrom (1985). For simplicity, I assume that this single informed trader will send market orders to all exchanges with stale limit orders. After trade occurs, the information will be announced to the market. As a result, the single informed trader can only explore her private information once. I will show that the effectiveness of FBAs and order delay designs depend on the proportions of asset value jumps that are public or private signal.
2.3.1 Frequent Batch Auctions

FBAs can eliminate sniping because HFTs are able to cancel their stale quotes before next trading round, but it may cause higher adverse selection cost on private news trading for liquidity providers. How FBAs compete against continuous limit order book (LOB) exchange is an interesting research question. To conserve space, I will not investigate this question completely. What I will show is that since there is no pre-trade price transparency, the liquidity in FBAs might be worse than in a LOB exchange. FBAs exchanges also have incentive to compete on trading frequency.

Precisely, denote \( \tau \) as the length of each auction. Thus, the FBAs exchange will cross buy and sell orders and conduct trades for every \( \tau \) units of time. I assume the LOB exchange has order processing speed \( \delta \). I will study the equilibrium spread in these two exchanges separately. Denoting \( s^*_{\text{LOB}} \) and \( s^*_{\text{FBA}} \) as the equilibrium bid-ask spread for HFTs proving liquidity on the LOB and FBAs exchange respectively. We have the following results:

**Proposition 7** (Equilibrium Spread Comparison Between LOB and FBA) Assuming all events can only happen at most once within each auction and denote \( \phi(\delta) = 1 - e^{-(\lambda_I + \lambda_J)\delta} \), \( \Sigma \lambda = \lambda_I + \lambda_J + \lambda_U \) and \( f(\tau) = \frac{1}{1-e^{-\Sigma \lambda \tau}} \int_0^\tau \Sigma \lambda e^{-\Sigma \lambda x} [1 - e^{-(\lambda_I + \lambda_J)(\tau-x)}] dx \):

(i) when \( \frac{\lambda_I + \lambda_J + \lambda_U}{\lambda_I + \lambda_J + \lambda_U f(\tau)} \leq \mu < 1 \), \( s^*_{\text{LOB}} \leq s^*_{\text{FBA}}(\tau) \);

(ii) when \( \tau_1 \leq \tau \), \( s^*_{\text{FBA}}(\tau_1) \leq s^*_{\text{FBA}}(\tau_2) \).

The intuition is simple. When an undercutting HFT arrives, other HFTs who are not at the best price quotes need to adjust or cancel their limit orders because these orders have no chance to trade with an investor, but they are still subject to the pick off risk by the single informed trader. But in FBAs, because of no pre-trade price transparency, liquidity-providing HFTs cannot observe the arrival of an undercutting HFT and therefore do not adjust their quotes accordingly. This increases their exposure to the single informed trader and so is their liquidity provision cost on FBAs. When the portion of informed trading from private news is high (large \( \mu \)) or undercutting HFTs arrive more frequently, this cost will dominate the benefits of no sniping on FBAs. Consequently, liquidity on FBAs can be worse than on the continuous LOB. For the same reason, a FBAs conducting trades more frequently will suffer less from the loss of pre-trade price transparency. As a result, HFTs can provide liquidity at lower bid-ask spread on more frequent
FBAs exchange. This confirms the conjecture that exchanges will compete on trading frequency if all of them implement FBAs. They might converge to continuous trading.

To avoid manipulation, FBAs designed in BCS does not disseminate its current limit orders before next trading round. Alternatively, a random stop FBAs with pre-trade price transparency might work better because HFTs can immediately adjust or cancel their quotes after undercutting HFT’s arrival. I assume all events can only happen at most once within each auction. This extremely simplifies the analysis of FBAs. Although it can captures the additional liquidity provision cost on FBAs exchange, the optimal frequency of FBAs cannot be well analyzed under this assumption. For this reason, I do not consider how FBAs competes with LOB exchange and the optimal trading frequency. What Proposition 7 shows is that FBAs is not a cost free proposal to reduce high frequency trader’s speed advantage. In next section, I will consider how an exchange with order delays compete against the continuous LOB exchange.

2.3.2 Order Delays

I will keep the same model structure and notation as in Section 1.3, so that we could directly use previous results when they are applicable. Specifically, there are \( M - 1 \) exchanges and each has order processing speed \( \delta_F \). All these \( M - 1 \) exchanges use the traditional continuous limit order book (LOB) to conduct trades. There is another exchange which currently has order processing speed \( \delta_F \) too but considering to implement one of the following order delays with deterministic delaying time \( \Delta_{\text{speed}} = \delta_S - \delta_F > 0 \):

**Definition 2** Uniform Delay (UD): delaying all incoming orders and cancellation messages by a deterministic time \( \Delta_{\text{speed}} \);

**Definition 3** Non-cancellation Delay (ND): delaying all incoming liquidity taking and adding orders but not cancellations by a deterministic time \( \Delta_{\text{speed}} \);

**Definition 4** Liquidity Taking Delay (LTD): delaying only liquidity taking orders by a deterministic time \( \Delta_{\text{speed}} \);

Uniform delay and non-cancellation delay are defined in the same way as in Baldauf and Mollner (2017). Liquidity taking delay only delays those market or marketable limit orders which
can trigger immediate execution. Note that although IEX implements uniform delay, it indeed has some features as non-cancellation delay due to its primary peg order. Primary peg is a non-displayed order at or inside the national best bid and offer (NBBO). When IEX determines that current quotes is unstable, primary peg buy (sell) orders are automatically slid one tick below (above) the current national best bid (offer). When the asset’s value jumps, liquidity providers (primarily HFTs) need to adjust their quotes and sniping or latency arbitrage opportunity is usually available during this quote unstable time. By using the primary peg order, IEX helps its liquidity providers to avoid trading in the wrong direction as the market moves by automatically sliding back the primary peg quotes. This is essentially like to allow liquidity providers at or inside the current best price quotes to cancel their stale quotes and reprice their limit orders. For this reason, I will consider IEX as more in line with non-cancellation delay.

Among these three order delays, whether it could mitigate the problem of sniping depending on whether liquidity providers have the advantage to cancel their stale quotes before taken by other HFTs. Since the uniform delay exchange also delays cancellation message, liquidity providers on uniform delay exchange suffer the same sniping cost as in a regular limit order book exchange. This makes uniform delay less attractive. On both non-cancellation delay and liquidity taking delay exchanges, liquidity provision HFTs have the advantage to cancel their stale quotes before that are taken by other HFTs. As a result, sniping can not occur on exchanges with one of these two design features.

But since the non-cancellation delay exchange also delays liquidity adding orders, it is less attractive for an undercutting HFT comparing with the liquidity taking delay. For this reason, the non-cancellation delay exchange may suffer low trading volume when competing with other \( M - 1 \) regular limit order book exchanges. Now I will study the equilibrium spread and each delay exchange’s trading volume in more details.

For the exchange implementing uniform delay, it essentially has order processing speed \( \delta_F + \Delta_{speed} = \delta_S \). Therefore, the results in Propositions 4 and 5 under the case of \( K = M - 1 \) fast exchanges could be directly applied to here. Denoting \( s^*_{UD} \) and \( Q^*_{UD} \) as the equilibrium spread and the uniform delay exchange’s expected per unit time trading volume, we would have \( s^*_{UD} = s^*(\delta_F) \)

---

8More detailed explanations about IEX’s primary peg order can be found at https://www.iextrading.com/docs/IEX%20Primary%20Peg%20Upgrade%20Overview.pdf.
and \( Q_{UD}^* = Q_S^*(M - 1) \).

If the exchange implements liquidity taking delay (LTD), HFTs could potentially provide liquidity on this LTD exchange with smaller bid-ask spread. Precisely, if a HFT provide liquidity on LTD exchange with bid-ask spread \( s \) and this is the unique exchange has this best price quotes, then similar to equation (1.5) (at \( X = 1 \)) the liquidity provision HFT has profit:

\[
\pi_{LTD}(s/2, 1) = \frac{\lambda_I}{\Sigma \lambda} s - \frac{\mu \lambda_I}{\Sigma \lambda} (s - \frac{s}{2}) + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_F) \left[ \frac{\lambda_I}{\lambda_I + \lambda_J 2} \frac{1}{2} s - \frac{\mu \lambda_J}{\lambda_I + \lambda_J 2} (\sigma - \frac{s}{2}) \right] + \frac{\lambda_U}{\Sigma \lambda} \left[ 1 - \phi(\delta_F) \right] \left[ \frac{\lambda_I}{\lambda_I + \lambda_J 2} \frac{1}{2} s - \frac{\mu \lambda_J}{\lambda_I + \lambda_J 2} (\sigma - \frac{s}{2}) \right] \quad (2.1)
\]

The difference between \( \pi_{LTD}(s/2, 1) \) and \( \pi((s/2, 1) \delta = \delta_F) \) (equation (1.5) evaluated at order processing speed \( \delta_F \) and \( X = 1 \)) is that: when the risky asset’s common value jumps, liquidity-providing HFT will lose profit only when the signal is private (with probability \( \mu \) conditional on the asset’s value jumps). Because for public signal (with probability \( 1 - \mu \)), liquidity-providing HFT can cancel her stale limit orders. Thus, no sniping or latency arbitrage on this LTD exchange. It is straightforward that for any \( 0 \leq \mu < 1, \pi_{LTD}(s/2, 1) > \pi((s/2, 1) \delta = \delta_F) \). Thus, HFTs strictly prefer to provide liquidity on LTD exchange. For the same reason, if multiple exchanges have the best price quotes, the undercutting HFT will also prefer to submit her price-improving limit order to the LTD exchange if it is one among those exchanges with current best price quotes. Therefore, the equilibrium structure is exactly the same as in Proposition 4 under the case of \( K = 1 \) and the LTD exchange plays a similar role as the unique fast exchange in that case. Similar to equation (1.6), we can determine the equilibrium half bid-ask spread \( s_{LTD}^*/2 \) as:

\[
\frac{s_{LTD}^*}{2} = \min \left\{ \frac{s}{2} | \nu_0 \pm \frac{s}{2} \in \mathcal{P}, \pi_{LTD}(s/2, 1) \geq 0 \right\} \quad (2.2)
\]

if \( \frac{\mu \lambda_J}{\lambda_I + \mu \lambda_J} > \frac{d}{2\sigma} \). Otherwise, the equilibrium bid-ask spread is binding at one tick. HFTs will provide liquidity on LTD exchange and other regular limit order book (LOB) exchanges with half spread \( s_{LTD}^*/2 \) until their profits become to negative. Suppose HFTs provide liquidity on X

Note that we assume undercutting HFT will not adjust her quotes or take liquidity if the bid-ask spread is binding at one tick. Thus with half spread \( d/2, \pi_{LTD}(d/2, 1) = \frac{\lambda_I}{\lambda_I + \lambda_J 2} \frac{d}{2} - \frac{\mu \lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{d}{2}) \). If \( \pi_{LTD}(d/2, 1) < 0 \Leftrightarrow \frac{\mu \lambda_J}{\lambda_I + \mu \lambda_J} > \frac{d}{2\sigma} \), the equilibrium bid-ask spread will not be binding at one tick at the initial quoting stage.
exchanges including the LTD exchange with half spread \( s/2 \) and denote the liquidity provision profit on a regular LOB exchange as \( \pi_{LOB}(\frac{s}{2}, X) \), then similar to equation (1.17) we have:

\[
\pi_{LOB}(\frac{s}{2}, X) = \frac{\lambda_I}{\Sigma \lambda} \frac{1 - \gamma s}{2} - \frac{\lambda_J}{\Sigma \lambda} (\sigma - \frac{s}{2}) + \frac{\lambda_U}{\Sigma \lambda} \phi(\epsilon) \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1 - \gamma s}{X} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s}{2}) \right] + \\
\frac{\lambda_U}{\Sigma \lambda} [1 - \phi(\epsilon)] \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1 - \gamma s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s}{2}) \right] \tag{2.3}
\]

The only difference between equation (2.3) and (1.17) is that HFTs on LOB exchanges will cancel their order \( \delta_F + \epsilon - \delta_F = \epsilon \) units of time after the undercutting HFT’s arrival because all LOB exchanges have order processing speed \( \delta_F \). Note that the above profit is defined only for \( X \geq 2 \) because when \( X = 1 \), HFTs only provide liquidity on the LTD exchange. Thus, we denote \( \pi_{LOB}(\frac{s_{LTD}}{2}, 1) = 0 \). Denoting \( M^*_LTD \) as the equilibrium depth, we have:

\[
M^*_LTD = \max\{X | \pi_{LOB}(\frac{s_{LTD}}{2}, X) \geq 0, 1 \leq X \leq M\} \tag{2.4}
\]

If non-cancellation delay (ND) is implemented, HFTs do not necessarily prefer to provide liquidity on the ND exchange. We can see this from HFT’s liquidity provision profit on ND exchange. Similar to \( \pi_{LTD}(\frac{s}{2}, 1) \), denote \( \pi_{ND}(\frac{s}{2}, 1) \) as the liquidity provision profit for a HFT who provides liquidity on the ND exchange with half bid-ask spread \( s/2 \) and this is the only exchange has these best price quotes. The only difference between \( \pi_{ND}(\frac{s}{2}, 1) \) and \( \pi_{LTD}(\frac{s}{2}, 1) \) is that the ND exchange takes \( \delta_S \) units of time to process undercutting HFT’s limit order because ND exchange also delays liquidity adding orders. Thus, by replacing \( \delta_F \) with \( \delta_S \) in equation (2.1) we would have \( \pi_{ND}(\frac{s}{2}, 1) \).\(^{10}\)

When the delaying time \( \Delta_{speed} = \delta_S - \delta_F \) is large, it is possible that \( \pi_{ND}(\frac{s}{2}, 1) < \pi(\frac{s}{2}, 1) | \delta = \delta_F \) for some spread \( s \). Intuitively, HFTs would like to provide liquidity on the ND exchange because no sniping on ND exchange. But if the ND exchange takes longer time to process any incoming limit order, it will increase liquidity provider’s adverse selection cost. If the later cost is larger, HFTs will prefer to provide liquidity on regular limit order book exchanges which have fast order processing speed \( \delta_F \).

Therefore, if non-cancellation delay (ND) is implemented, the equilibrium depends on whether

\(^{10}\)Precisely, \( \pi_{ND}(\frac{s}{2}, 1) = \frac{\lambda_I}{\Sigma \lambda} \frac{1 - \gamma s}{2} - \frac{\lambda_J}{\Sigma \lambda} (\sigma - \frac{s}{2}) + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_S) \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1 - \gamma s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s}{2}) \right] + \\
\frac{\lambda_U}{\Sigma \lambda} [1 - \phi(\delta_S)] \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1 - \gamma s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s}{2}) \right]. \)
HFTs strictly prefer to provide liquidity on the ND exchange or not. If not, the equilibrium structure would be similar as in the uniform delay case. HFTs will run to provide liquidity on LOB exchanges and start to provide liquidity on the ND exchange only when all LOB exchanges have been already filled with current best price quotes. The ND exchange has small trading volume in this case. If HFTs prefer to provide liquidity on the ND exchange (i.e. when delaying time \( \Delta_{\text{speed}} \) is small), the equilibrium structure would be similar to the case of liquidity taking delay. The ND exchange could potentially reaches its maximum trading volume. I will mainly focus on the latter case because whether an exchange has incentive to implement non-cancellation order delay depends on the maximum trading volume the ND exchange could have.

Denote \( s_{\text{ND}}^* \) and \( Q_{\text{ND}}^* \) as the equilibrium spread and the expected per unit time trading volume for the ND exchange. When HFTs prefer to provide liquidity on the ND exchange, \( s_{\text{ND}}^* \) is determined exactly in the same way as in equation (2.2). Thus, \( s_{\text{ND}}^*/2 = \min\{(\xi/2)|\nu_0 \pm \xi/2 \in \mathcal{P}, \pi_{\text{ND}}(\xi, 1) \geq 0\} \) when \( \frac{\mu_i}{\lambda_i + \mu_i} > \frac{\delta}{2\sigma} \). Otherwise, it is binding at one tick. Since undercutting HFT also prefers to submit her price-improving limit order to the ND exchange, thus liquidity-providing HFTs on other LOB exchanges will cancel their stale limit order \( \delta_S + \epsilon - \delta_F \) units of time after the undercutting HFT’s arrival. Similar to equation (2.3), denote \( \pi'_{\text{LOB}}(\xi/2, X) \) as a HFT’s liquidity provision profit on one LOB exchange when HFTs provide liquidity on X exchanges including the ND exchange at bid-ask spread \( s \). Therefore, by replacing \( \epsilon \) with \( \delta_S + \epsilon - \delta_F \) in equation (2.3) we would have \( \pi'_{\text{LOB}}(\xi/2, X) \). Similar to equation (2.4), the equilibrium depth \( M_{\text{ND}}^* = \max\{X|\pi'_{\text{LOB}}(\xi/2, X) \geq 0, 0 \leq X \leq M\} \) and denote \( \pi'_{\text{LOB}}(\xi/2, 1) = 0 \). Now we summarize each delay exchange’s expected per unit time trading volume in Proposition 8.

**Proposition 8** (Order Delay Exchange’s Trading Volume) If there are \( M \) exchanges with order processing speed \( \delta_F \) and one among them considers to implement uniform delay (UD), non-cancellation

---

11Specifically, if HFTs prefer to provide liquidity on regular limit order book exchanges, the equilibrium bid-ask spread would be \( s'(\delta_F) \). Only when \( M'(\delta_F) > M - 1 \) and HFTs have non-negative liquidity provision profit on the ND exchange, which requires \( \frac{\lambda_i}{\lambda_i + \mu_i} \left(1 - \frac{1}{M} \right) \frac{\mu_j}{\lambda_j + \mu_j} (\sigma - \frac{s'(\delta_F)}{2}) + \frac{\mu_j}{\lambda_j + \mu_j} \phi(\delta_F + \epsilon - \sigma \delta_S) \left[ \frac{\lambda_i}{\lambda_i + \lambda_j} \left(1 - \frac{1}{M} \right) - \frac{\mu_j}{\lambda_j + \lambda_j} \frac{1}{M} (\sigma - \frac{s'(\delta_F)}{2}) \right] \right] > 0 \) the ND exchange has non-zero trading volume \( \left[ \frac{\lambda_i}{\lambda_i + \mu_i} \right] - \frac{\mu_j}{\lambda_j + \mu_j} \left[ 1 - \frac{\lambda_i}{\lambda_i + \lambda_j} \left(1 - \phi(\epsilon') \right) \right] \), where \( \epsilon' = \delta_F + \epsilon - \delta_S \). HFT’s liquidity provision profit on ND exchange and per unit time trading volume are similar to equation (1.17) and (1.21) by replacing the total equilibrium depth with \( M \) and the fact that no sniping on the ND exchange.

12Precisely, \( \pi'_{\text{LOB}}(\xi/2, X) = \frac{\lambda_i}{\lambda_i + \mu_i} (\sigma - \frac{s'(\delta_F)}{2}) + \frac{\mu_j}{\lambda_j + \mu_j} \phi(\delta_S + \epsilon - \delta_F) \left[ \frac{\lambda_i}{\lambda_i + \lambda_j} \left(1 - \frac{1}{M} \right) - \frac{\mu_j}{\lambda_j + \lambda_j} (\sigma - \frac{s'(\delta_F)}{2}) \right] + \frac{\mu_j}{\lambda_j + \mu_j} \left[ 1 - \phi(\delta_S + \epsilon - \delta_F) \right] \left[ \frac{\lambda_i}{\lambda_i + \mu_i} \left(1 - \frac{1}{M} \right) - \frac{\lambda_j}{\lambda_j + \lambda_j} \left(1 - \frac{1}{M} \right) \right] \left(\sigma - \frac{s'(\delta_F)}{2}\right). \)
delay (ND) or liquidity taking delay (LTD) with deterministic delaying time $\Delta_{\text{speed}} = \delta_S - \delta_F > 0$, denoting $\epsilon'' = \delta_S + \epsilon - \delta_F$ and $\epsilon' = \delta_F + \epsilon - \delta_S$, then when $\gamma \geq \tilde{\gamma}$, $\frac{\mu\phi}{\lambda F + \mu\phi} > \frac{d}{2\sigma}$ and $s'(\delta_F)/2 < \sigma$:

(i) If Uniform Delay is implemented, then $Q_{UD}^* = Q_S^*(M - 1)$;

(ii) If Liquidity Taking Delay is implemented, then:

$$Q_{LTD}^* = \lambda_I(1 - \frac{1 - \gamma}{M_{LTD}^*}) + \mu\lambda_J + \frac{\lambda_I\lambda_J}{2\Sigma \lambda} \mu\phi(\delta_F) + \frac{\lambda_I\lambda_F}{2\Sigma \lambda} [1 - \phi(\epsilon)] \frac{M_{LTD}^* - 1}{M_{LTD}^*} (1 - \gamma) \quad (2.5)$$

(iii) If Non-cancellation Delay is implemented, then either when 1) $s_{ND}^* < s'(\delta_F)$ or 2) $\mu\phi(\delta_F) \leq \phi(\delta_F) + (1 + 2\frac{\lambda_I + \lambda_J}{\lambda_U})(1 - \mu)$ and $(1 - \mu)\lambda_J(\sigma - \frac{S_{ND}^*}{2} + d) - \frac{\lambda_I}{M_{ND}^*}(\frac{S_{ND}^*}{2} - d)[(M_{ND}^* - 1)(\phi(\epsilon'') - \phi(\epsilon)) + \phi(\epsilon) - \phi(\epsilon')] \geq 0$, the ND exchange attains its maximum per unit time trading volume:

$$Q_{ND}^* = \lambda_I(1 - \frac{1 - \gamma}{M_{ND}^*}) + \mu\lambda_J + \frac{\lambda_I\lambda_J}{2\Sigma \lambda} \mu\phi(\delta_S) + \frac{\lambda_I\lambda_I}{2\Sigma \lambda} [1 - \phi(\epsilon'')][\frac{M_{ND}^* - 1}{M_{ND}^*} (1 - \gamma)] \quad (2.6)$$

Proof of the above results are similar to the proof of Propositions 4 and 5. (i) is directly implied from Proposition 5 because with uniform delay all LOB exchange has faster order processing speed $\delta_F$ while the UD exchange has slow order processing speed $\delta_S$. If liquidity taking delay is implemented and $\gamma \geq \tilde{\gamma}$, in the appendix I show that HFTs have larger liquidity provision profit on the LTD exchange than on other LOB exchanges at the same bid-ask spread. Thus HFTs will first provide liquidity on the LTD exchange and then provide liquidity on other LOB exchanges as long as they can earn non-negative liquidity provision profits. The LTD exchange’s per unit time trading volume is calculated similarly to the case of $K = 1$ fast exchange in Proposition 5.

In (iii), the ND exchange attains maximum trading volume when HFTs prefer to provide liquidity on the ND exchange. If $s_{ND}^* < s'(\delta_F)$, HFTs can provide liquidity on ND exchange with better price. Price competition among HFTs will drive them to provide liquidity on the ND exchange first. When HFTs can not provide liquidity on the ND exchange with better price but the first condition in (iii) 2) is satisfied, then HFTs will have larger liquidity provision profit on the ND exchange than on other LOB exchanges at the same bid-ask spread like the LTD case. This is because as long as the ND exchange will not delay limit orders for a longer time, the cost saving of no sniping on the ND exchange will dominate the cost of longer order processing time for liquidity provision HFTs. When the second condition in (iii) 2) is satisfied, undercutting HFTs
will strictly prefer to provide liquidity on the ND exchanges too.\footnote{We need two separate conditions because undercutting HFT has different liquidity provision profit from other HFTs who submit the initial quotes.} Note that when the delaying time $\Delta_{\text{speed}}$ is small, both conditions will hold. Now we can compare each delay exchange's trading volume and shed some light on whether current fast exchanges have incentive to implement order delay or not. We put these results in Corollary 3.

**Corollary 3 (Volume Comparison)** When $\gamma \geq \gamma_0$ and $s^*(\delta_F)/2 < \sigma$ if $0 < \mu \leq -\frac{\lambda_f[1+\frac{\mu}{\beta^*}(\delta_F)]-(M-1)\lambda_I}{\lambda fM[1+\frac{\mu}{\beta^*}(\delta_S)]}$, then $\max\{Q_{\text{LTD}}^*, Q_{\text{ND}}^*\} \leq Q^*(\delta_F)$.

This simple result has important policy implications. Many industry experts argue that if there is a problem in the equity trading, just let the market to fix it and no regulation is needed. Corollary 3 shows that even sniping is a significant problem (large $1 - \mu$) for some stocks, those exchanges with fastest order processing speed (thus will have large market share of trading volume in stocks with large $1 - \mu$) do not have incentives to implement any order delay designs if their goals are to maximize trading volume (recall that from Corollary 2, we have $Q_{\text{UD}}^* = Q_{\text{S}}^*(M-1) < Q^*(\delta_F)$ too).

The intuition is simple: whenever sniping occurs, those fastest exchanges share revenues with sniping HFT. If LTD or ND is implemented, these fastest exchanges will lose trading volume and revenue from sniping trades. Therefore, regarding of the sniping problem fastest exchanges do not necessarily stand on long-term investor’s side.

### 2.4 Conclusions

This chapter characterizes how different exchange designs such as frequent batch auctions and order delays affect trading outcomes. Given the costly arms race for speed and the harm that this speed can create for investors, the Securities and Exchange Commission (SEC) is considering alternative market designs. Budish et al. (2015) proposed that the current exchange design be replaced with frequent batch auctions (say auctions every 0.001 second); and Du and Zhu (2017) argue that frequent batch auctions can improve allocative efficiency. Yet no exchange employs a frequent batch design. The closest current design is the Investors Exchange’s (IEX), which has an order delay, a design whose impact is analogous to that of the frequent batch design considered
by Budish et al. (2015); however, the IEX has only drawn small amounts of trading volume. I provide explanations for why these alternative exchange designs proposed in the literature do not seem to work well in reality.

I incorporate more realistic features than does the literature, and show that they sharply alter the nature of the optimal exchange design. I show that when one speculator learns about shifts in a stock’s value before other speculators, frequent batch auctions will have low liquidity. The problem with frequent batch auctions is that HFTs cannot see an undercutting HFT’s order, which raises their risk of exposure to the informed speculator—and this adverse-selection cost can dominate concerns of sniping by liquidity-providing HFTs. I further show that exchanges with frequent batch auction designs have the incentive to conduct batch auctions more and more quickly in order to reduce the liquidity-provision costs. Thus, they will unravel. I also show that designs like the Chicago Stock Exchange (CHX), which only delays liquidity-taking orders, will attract more trading volume than designs that delay all incoming orders like IEX, because an exchange’s order-processing speed is crucial for attracting more price-improving limit orders, and thus more trading volume.
Chapter 3

Who Provides Liquidity and, When?

3.1 Introduction

In decades past, specialists on the New York Stock Exchange and dealers in NASDAQ supply liquidity to other traders, that is, they buy when other traders sell and sell when other traders buy. The transition to electronic trading not only destroyed these traditional liquidity suppliers, but also blurs the definition of liquidity supply. Everyone can supply liquidity, but no one is obligated to do so. Liquidity supply simply means to post a limit order, an offer to buy or sell at a certain price. A trade occurs when another trader (a liquidity demander) accepts the terms of a posted offer. Every trader has to decide whether to supply or demand liquidity in order to complete a trade. In this paper, we examine how the contemporary trading environment of voluntary liquidity supply and demand reaches its equilibrium. Who supplies liquidity and who demands liquidity? Can voluntary liquidity supply and demand lead to systemic risk such as a flash crash? And, if this is possible, what conditions lead up to it?

In this paper, we show how the equilibria in liquidity supply and demand depend on the characteristics of securities, market structures, and market conditions. Our model extends Budish et al. (2015) (BCS hereafter) along two dimensions. BCS include two types of traders: high-frequency traders (HFTs) and non-HFTs. In the BCS model, non-HFTs can only demand liquidity, while in our model we allow non-HFTs to provide liquidity. In addition, BCS consider a continuous price, whereas we consider a discrete price to reflect the tick size (minimum price variation) imposed by the U.S. Security and Exchange Commission’s (SEC’s) Regulation National Market Systems (Reg NMS) Rule 612, and to reflect the recent policy debate to increase the tick size from one cent to five cents.

Our model includes one security, whose fundamental value is public information. However,
liquidity suppliers in our model are subject to adverse selection risk, because they may fail to cancel stale quotes during value jumps. HFTs in our model have no private value to trade. They consistently monitor the market for profit opportunities. For example, they supply liquidity when the expected profit from doing so is positive, or snipe stale quotes after value jumps. Non-HFTs arrive at the market with a private value to buy or sell one unit of a security. We allow a fraction of non-HFTs to choose between providing or demanding liquidity. We call these non-HFTs “buy-side algorithmic traders” (BATs) to represent algorithms used by buy-side institutions (e.g., mutual funds and pension funds) to minimize the cost of executing trades in portfolio transition (Hasbrouck and Saar (2013); Frazzini et al. (2014)). BATs are major players in modern financial markets (O’Hara (2015)). We build the first theoretical model to study their trading behavior. Our model captures two main features of BATs. First, BATs are slower than HFTs (O’Hara (2015)). Second, BATs supply liquidity to minimize the transaction costs of portfolio rebalancing (Hasbrouck and Saar (2013)), not to profit from the bid-ask spread. As both BATs and HFTs are algorithmic traders (Hasbrouck and Saar (2013)), we call the fraction of non-HFTs who are not BATs non-algorithmic traders (non-algos).

As in BCS, the adverse selection risk increases with the arrival rate of value jumps and decreases with the arrival rate of non-HFTs. Supplying liquidity to non-HFTs leads to revenue, but value jumps lead to sniping cost. With the continuous price in BCS, the competitive bid-ask spread strictly increases with adverse selection risk. In our model, the tick size constrains price competition in the bid-ask spread. When adverse selection risk is low or the tick size is large, the competitive bid-ask spread can be less than one tick, which generate rents for liquidity supply. The rents are typically allocated to HFTs, because most U.S. stock exchanges use time to decide execution priority for orders quoted at identical prices. The market thus reaches equilibrium through queuing, not through price competition. In this first type of equilibrium, the queuing equilibrium, in which bid-ask spread is binding at one tick, HFTs dominate liquidity supply due to their speed advantage over BATs.

When the tick size does not bind, we find that BATs never demand liquidity from HFTs. Instead, they provide liquidity at more aggressive prices than HFTs. This result is surprising because Han et al. (2014), Hoffmann (2014), Bernales (2014), and Bongaerts and Van Achter (2016) maintain that HFTs cancel stale quotes faster, incur lower adverse selection cost, and quote more aggressive
prices than other traders. Brogaard et al. (2015), however, show that non-HFTs quote tighter bid-ask spreads than HFTs. Our model reconciles the contraction between previous channels of speed competition and the empirical results by including the opportunity cost of liquidity supply. BATs have to trade in our model. The outside option for BATs is to demand liquidity and pay the bid-ask spread. For BATs, supplying liquidity at a tighter bid-ask spread strictly dominate demanding liquidity from HFTs.

To show why BATs choose to supply liquidity, we develop a new concept: the make-take spread. Without loss of generality, consider the BATs’ decision to buy and HFTs’ decision to sell. HFTs quote an ask price above the fundamental value, and their difference, or the half bid-ask spread, reflects the compensation for adverse selection costs during value jumps. BATs pay the half bid-ask spread if they demand liquidity. BATs can reduce transaction costs by supplying liquidity slightly above the fundamental value. We call this type of limit order a flash limit order, because it immediately triggers HFTs to demand liquidity. Flash limit orders execute immediately like market orders, but with a lower transaction cost. Flash limit orders exploit the make-take spread, the price difference between HFTs’ willingness to make an offer and their willingness to accept one. HFTs accept a lower sell price when they demand liquidity, because when they immediately accept an order, they do not incur adverse selection costs during a value jump.

When the tick size does not impose a constraint for BATs to quote more aggressive prices than HFTs, our model has two types of equilibria: flash and undercutting. In the flash equilibrium, BATs use flash limit orders to supply liquidity to HFTs. In the undercutting equilibrium, BATs quote a buy limit order price below the fundamental value or a sell limit order price above the fundamental value. These regular limit orders stay in the LOB to supply liquidity to non-algos or other BATs. We find that undercutting equilibrium are more likely to occur when the adverse selection risk is low, because flash limit orders incur no adverse selection cost, whereas the cost of regular limit orders increases with the adverse selection risk.

We also examine mini-/flash crashes, which are sharp price movements in one direction followed by quick reversion (Biais et al. (2014)), and predict their cross-sectional and time series patterns. In the cross-section, mini-/flash crashes are more likely to occur for stocks with a smaller tick size or higher adverse selection risk. Because BATs can undercut HFTs for these stocks, HFTs’ limit orders face lower execution probability before value jumps. When the fraction of BATs is
large enough, HFTs have to quote stub quotes, a bid-ask spread wider than the maximum value of the jump, to protect against sniping. Yet BATs do not always supply liquidity on both sides of the market. Thus, an incoming market orders can hit HFTs’ stub quotes, causing a mini-flash crash. In time series, a downward (upward) mini-flash crash is more likely to occur immediately after a downward (upward) price jump, because such jumps can snipe all BATs’ limit orders on the bid (ask) side raising the probability that market orders hit stub quotes before BATs refill the limit order book (LOB).

Existing literature on HFTs focuses on the role of adverse selection. On the one hand, speed can allow HFTs to adversely select other traders, which harms liquidity; on the other hand, speed can reduce adverse selection costs for liquidity suppliers and improve liquidity [see Jones (2013), Biais et al. (2014), and Menkveld (2016) for surveys]. We contribute to the literature by identifying two new channels of speed competition, both of which are unrelated to adverse selection. For liquidity demand, we find that HFTs race to demand liquidity when BATs post flash limit orders, but HFTs impose no adverse selection cost on BATs. Instead, BATs prompt HFTs to demand liquidity to reduce their transaction costs. Thus, liquidity demand from HFTs need not be bad. Indeed, transactions costs are lower when HFTs demand liquidity than when they supply liquidity.

For liquidity supply, our queuing channel of speed competition rationalizes three contradictions between empirical evidence and existing theoretical channels that focus on adverse selection. If an HFT’s speed advantage primarily helps it to reduce adverse selection costs, HFTs should realize a comparative advantage in providing liquidity for stocks with higher adverse selection costs (Han et al. (2014), Hoffmann (2014), Bernales (2014), and Bongaerts and Van Achter (2016)). HFTs should also crowd out slow liquidity suppliers when the tick size is smaller, because a smaller tick size reduces the constraints to offer better prices (Chordia et al. (2013)). In addition, a higher cancellation-to-trade ratio likely indicates more liquidity supply from HFTs, because HFTs need to cancel many orders to avoid adverse selection risk [see Biais et al. (2014), and Menkveld (2016) for a survey]. Yet Jiang et al. (2014) and Yao and Ye (2017) show that non-HFTs dominate liquidity supply when adverse selection risk is high. O’Hara et al. (2015) and Yao and Ye (2017) show that a smaller tick size crowds out HFTsf liquidity supply. Yao and Ye (2017) show stocks with higher fractions of liquidity provided by HFTs have lower cancellation-to-trade ratios. The queuing channel of speed competition reconciles these three contradictions.
The tick size is more likely to be bind when adverse selection risk is low or the tick size is large. A binding tick size helps HFTs to establish time priority. HFTs dominate liquidity supply for stocks with larger tick sizes, but they also have less incentive to cancel orders. A smaller tick size or higher adverse selection risk allows BATs to increase liquidity provision by establishing price priority, but smaller tick size or higher adverse selection risk also leads to more frequent order cancellations. This theoretical intuition, along with the empirical evidence in Yao and Ye (2017), suggests that the cancellation-to-trade ratio should not be used as a cross-sectional proxy for HFT activities.1

Our model casts doubt on the recent policy proposal in the U.S. to increase the tick size, initiated by the 2012 Jumpstart Our Business Startups Act (the JOBS Act). In October 2016, the SEC started a two-year pilot program to increase the tick size from one cent to five cents for 1,200 less liquid stocks. Proponents to increase the tick size assert that a larger tick size should control the growth of HFTs and increase liquidity (Weild et al. (2012)). We find that an increase in tick size would encourage HFTs. We also find that an increase in tick size constrains price competition and reduces liquidity. A larger tick size may reduce mini-flash crashes, or very high volatility in liquidity, but such a reduction decreases liquidity in normal times. We argue that a more effective way to reduce a mini-flash crash is a trading halt after value jumps so that liquidity supply from BATs can resume.

3.2 Model

In our model, the stock exchange operates as a continuous limit order book (LOB). Each trade in the LOB requires a liquidity supplier and a liquidity demander. The liquidity supplier submits a limit order, which is an offer to buy or sell at a specified price and quality. The liquidity demander accepts the conditions of a limit order. Execution precedence for liquidity suppliers follows the price-time priority rule. Limit orders with higher buy or lower sell prices execute before less aggressive limit orders. For limit orders queuing at the same price, orders arriving earlier execute before later orders. The LOB contains all outstanding limit orders. Outstanding orders to buy are

1 The cancellation-to-trade ratio can still be a good time series proxy for HFTs’ activity (Hendershott et al. (2011); Angel et al. (2015); Boehmer et al. (2015)).
called “bids” and outstanding orders to sell are called “asks.” The highest bid and lowest ask are called the “best bid and ask (offer)” (BBO), and the difference between them is the bid-ask spread.

Our model has one security, $x$, whose fundamental value, $v_t$, evolves as a compound Poisson jump process with arrival rate $\lambda_j$. $v_t$ starts from 0, and changes by a size of $d$ or $-d$ in each jump with equal probability. As in BCS, $v_t$ is common knowledge, but liquidity suppliers are subject to adverse selection risk when they fail to update stale quotes after value jumps. Traders start with a small latency to observe the common value jump,² but can reduce the latency to 0 by investing in a speed technology with cost $c_{\text{speed}}$ per unit of time.

Our model includes HFTs and two types of non-HFTs: BATs and non-algo traders. HFTs place no private value on trading. They supply or demand liquidity as long as the expected profit is above 0. They submit a market order to buy (sell) $x$ when its price is below (above) $v_t$. HFTs supply liquidity as long as the expected profit from the bid-ask spread is above 0. Non-HFTs, who arrive with a compound Poisson jump process with intensity $\lambda_1$, have to buy or sell one unit of $x$, each with probability $\frac{1}{2}$. Non-HFTs do not invest in speed technology because they only arrive at the market once.

Our model extends BCS along two dimensions. First, non-HFTs in the BCS model submit only market orders. In our model, we allow a proportion $\beta$ of non-HFTs, BATs, to choose between limit and market orders to minimize transaction costs. The rest of the non-HFTs, non-algo traders, use only market orders. Second, BCS assume continuous pricing in their model, whereas we consider discrete pricing grids. The benchmark pricing grid in Section 3.3 $\{\cdots, -\frac{3d}{2}, -\frac{d}{2}, \frac{d}{2}, \frac{3d}{2}, \cdots\}$ has a tick size of $\Delta_0 = d$. This choice ensures that $v_t$ is always at the midpoint of two price levels at any time. In Sections 3.4 to 3.7, we reduce the tick size to $\Delta_1 = d/3$, which creates additional price levels, such as $d/6$ and $-d/6$. Figure 3.1 shows the pricing grids with large and small tick sizes.

Following the dynamic LOB literature (e.g., Goettler et al. (2005), Goettler et al. (2009), Roșu (2009) and Colliard and Foucault (2012)), we examine the Markov perfect equilibrium, in which traders’ actions condition only on state of the LOB and events at $t$. We assume that HFTs instantaneously build up the equilibrium LOB after any event. Under this simplification, six types of

---

²By small, we mean that no additional events, such as a trader arrival or a value jump, take place during the delay.
events trigger the transition of the LOB across states:

\[
\begin{align*}
\frac{1}{2}\beta \lambda_I & \quad \text{BAT Sells (BS)} \\
\frac{1}{2}\beta \lambda_I & \quad \text{BAT Buys (BB)} \\
\frac{1}{2}(1 - \beta) \lambda_I & \quad \text{Non-algo Sells (NS)} \\
\frac{1}{2}(1 - \beta) \lambda_I & \quad \text{Non-algo Buys (NB)} \\
\frac{1}{2} \lambda_J & \quad \text{Price jumps up (UJ)} \\
\frac{1}{2} \lambda_J & \quad \text{Price jumps down (DJ)}
\end{align*}
\]

(3.1)

BCS do not allow non-HFTs to supply liquidity. We extend their model by allowing BATs to submit limit orders. To convey the economic intuition in the most parsimonious way, we make a technical assumption that BATs can only submit limit orders when the price level contains no other limit orders. This assumption reduces the number of states of the LOB that we need to track. We can further relax the assumption in BCS by allowing BATs to queue for \( n > 1 \) shares, but such an extension only increases the number of LOB states without conveying new intuition. Non-HFTs in the BCS model never use limit orders, which can be justified by an infinitely large delay cost (Menkveld and Zoican (2017)). Our extension effectively reduces the delay cost to allow BATs to submit limit orders.\(^3\) The main intuition of our model stays the same as long as BATs do not queue for infinite length.

### 3.3 Benchmark: Binding at one tick under a large tick size

Our analysis starts from \( \Delta_o = d \). As in BCS, HFTs can choose to be liquidity suppliers, who profit from the bid-ask spread, or to be stale-quote snipers, who profit by demanding liquidity from stale quotes after a value jump. In BCS, the equilibrium bid-ask spread equalizes the HFTs’ expected profits from these two strategies, which are both zero after speed investment. Lemma 2

\(^3\)We can assume a finite delay cost so that BATs only queue for one share, and the results are available upon request. The value of the delay cost, however, conveys no intuition and only leads to a more complicated proof. In Section 3.5, we show that the exact size of the delay cost has little impact for BATs’ choice between limit orders and market orders.
shows that this break-even bid-ask spread is smaller than the tick size when adverse selection risk is low.

Lemma 2 (Binding Tick Size) When $\Delta_0 = d$ and $\lambda_I > 1$, HFTs’ profit from providing the first share at the ask price of $a_i^* = v_t + d/2$ and the bid price of $b_i^* = v_t - d/2$ is higher than HFTs’ profit from stale-quote sniping.

Because non-HFTs trade for liquidity reasons and value jumps lead to sniping cost for stale quotes, $\lambda_I$ measures adverse selection risk in our model. As in BCS and Menkveld and Zoican (2017), this adverse selection risk comes from the speed of the response to public information, not from exogenous information asymmetry (e.g., Glosten and Milgrom (1985) and Kyle (1985)). As the arrival rate of non-HFTs increases or the intensity of value jumps decreases, the adverse selection risk decreases and so does the break-even bid-ask spread. The break-even bid-ask spread drops below one tick when $\lambda_I > 1$, making liquidity supply for the first share more profitable than stale-quote sniping.\(^4\) The rents for liquidity supply then trigger the race to win time priority in the queue. As BATs do not have a speed advantage to win the race, they demand liquidity in the same manner as non-algo traders. As a result, Lemma 1 does not depend on $\beta$.\(^5\)

Under a binding tick size, price competition cannot lead to economic equilibrium. It is the queue that restores the economic equilibrium. Next, we derive the equilibrium queue length for the ask side of the LOB, and the bid side follows symmetrically.

We evaluate HFTs’ value of liquidity supply and stale-quote sniping for each queue position, though we allow an HFT to supply liquidity at multiple positions and to snipe shares in other positions where she is not a liquidity supplier. We denote the value of liquidity supply for the $Q^{th}$ share as $LP(Q)$. A market sell order does not affect $LP(Q)$ on the ask side, because HFTs immediately restore the previous state of the LOB by refilling the bid side. A market buy order moves the queue forward by one unit, thereby changing the value to $LP(Q - 1)$. A limit order execution leads to a profit of $d/2$ to the liquidity supplier, $LP(o) = d/2$. When $v_t$ jumps upward,\(^6\)

---

\(^4\)Throughout this paper, we consider $\frac{\lambda_I}{\lambda_j} > 1$ for expositional simplicity. When $\frac{\lambda_I}{\lambda_j} \leq 1$, $\Delta_0$ is no longer binding, and the equilibrium structure is similar to that in Sections 3.4 to 3.7, where we reduce the tick size to $\Delta_j = d/3$.

\(^5\)An order with less time priority has lower probability of execution and higher probability of being sniped, both of which reduce BATs’ incentives to queue. In addition, BATs have incentives to implement trades, and a positive delay cost would compel them to use market orders when the queue is long. We assume that BATs never queue after the first position to reflect these intuitions in a parsimonious way.
the liquidity providing HFT of the $Q^{th}$ share races to cancel the stale quote, whereas the other $N - 1$ HFTs (with $N$ determined in equilibrium) race to snipe the stale quote. The loss from being sniped is $d/2$, while the probability of being sniped is $(N - 1)/N$. When $v_t$ jumps downward, the liquidity supplier cancels the order and joins the race to supply liquidity at a new BBO.\footnote{We assume that the HFT liquidity supplier cancels the limit order to avoid the complexity of tracking infinite many price levels in the LOB.} $LP(Q)$ then becomes 0. Equation (3.2) presents $LP(Q)$ in recursive form and Lemma 3 presents the solution for Equation (3.2).

$$LP(Q) = \frac{\frac{1}{2}\lambda_I}{\lambda_I + \lambda_J} LP(Q) + \frac{\frac{1}{2}\lambda_I}{\lambda_I + \lambda_J} LP(Q - 1) - \frac{N - 1}{N} \times \frac{d}{2} + \frac{\frac{1}{2}\lambda_J}{\lambda_I + \lambda_J} \times 0 \quad (3.2)$$

**Lemma 3 (Value of Liquidity Supply)** The value of liquidity supply for the $Q^{th}$ position is:

$$LP(Q) = \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q d - \frac{N - 1}{N} \times \frac{d}{2} \left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]$$

$LP(Q)$ decreases in $Q$.

Intuitively, Lemma 3 reflects the conditional probability of value-change events for $LP(Q)$ and their payoffs. Since $LP(Q)$ stays the same after a market sell order, the conditional probabilities of value-changing events are $\frac{\lambda_I}{\lambda_I + 2\lambda_J}$ for a market buy, $\frac{\lambda_J}{\lambda_I + 2\lambda_J}$ for an upward value jump, and $\frac{\lambda_J}{\lambda_I + 2\lambda_J}$ for a downward value jump. The $Q^{th}$ share executes when $Q$ non-HFTs arrive in a row to buy, which has a probability of $\left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q$, and the revenue conditional on execution is $d/2$. Their product, the first term in Equation (3.3), reflects the expected revenue for liquidity suppliers. The $Q^{th}$ share on the ask side fails to execute with non-HFTs when an upward or downward value jump occurs, each with probability $\frac{1}{2} \left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]$. After an upward value jump, the liquidity supplier has a probability of $1/N$ to cancel the stale quote, but failure to cancel the stale quote before sniping leads to a loss of $d/2$. The expected loss is $\frac{N - 1}{N} \times \frac{d}{2} \left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]$, the second term in Equation (3.3). A downward value jump before the order being snipped or executed leads to a zero payoff for the liquidity supplier. $LP(Q)$ decreases in $Q$, because an increase in a queue position reduces execution probability and increases the cost of being sniped.
The outside option for supplying liquidity for the $Q^{th}$ share is to be the sniper of the share during the value jump. HFTs’ liquidity supply decision for the $Q^{th}$ share also needs to include this opportunity cost. With a probability of $\frac{1}{2} \left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]$, the $Q^{th}$ share becomes stale before it gets executed, and each sniper has a probability of $1/N$ to profit from the stale quote. The value for each sniper of the $Q^{th}$ share is:

$$ SN(Q) = \frac{1}{N} \frac{1}{2} \left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]d $$

(3.4)

$SN(Q)$ increases with $Q$, because shares in a later queue position offer more opportunities for snipers.

HFTs race to supply liquidity for the $Q^{th}$ position as long as $LP(Q) > SN(Q)$, because the winner’s payoff is higher than that of the losers. Equation (3.5) determines the equilibrium length:

$$ \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Qd - \frac{1}{2} \left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]d > 0 $$

(3.5)

The solution for Equation (3.5) is:

$$ Q^* = \max\{Q \in \mathbb{N}^+ s.t. \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Qd - \frac{1}{2} \left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]d > 0\} $$

$$ = \max\{Q \in \mathbb{N}^+ s.t. \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q > \frac{1}{3}\} $$

$$ = \left\lfloor \log_{\frac{\lambda_I}{\lambda_I + 2\lambda_J}} \frac{1}{3}\right\rfloor $$

(3.6)

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to $x$.

Figure 3.2 shows the comparative statics for equilibrium queue length. The queue length at BBO decreases with $\frac{\lambda_I}{\lambda_J}$, which indicates that, for stocks with a bid-ask spread binding at one tick, the depth at the BBO may serve as a proxy for adverse selection risk. Traditionally, bid-ask spreads serve as a proxy for adverse selection risk (Glosten and Milgrom (1985); Stoll (2000)). Yet, Yao and Ye (2017) find that bid-ask spread is one-tick wide 41% of time for their stratified sample of Russell 3000 stocks in 2010. Depth at the BBO then serves as an ideal proxy to differentiate the
level of adverse selection for these stocks.\textsuperscript{7}

To derive $N$, note that HFTs’ total rents come from the bid-ask spread paid by non-HFTs, because sniping only redistributes the rents among HFTs. Ex ante, each HFT obtains $1/N$ of the rents per unit of time. New HFTs continue to enter the market until:

$$\lambda I_d^2 - Nc_{speed} \leq 0$$

(3.7)

In Proposition 9, we summarize the equilibrium under a large binding tick size.

**Proposition 9 (Large Binding Tick Size)** When $\Delta_0 = d$ and $\frac{\lambda_1}{\lambda_2} > 1$, $N^*$ HFTs jointly supply $Q^*$ units of sell limit orders at $a_i^* = v_i + d/2$ and $Q^*$ units of buy limit orders at $b_i^* = v_i - d/2$, where:

$$Q^* = \left\lfloor \log_{\frac{\lambda_1}{\lambda_2 + 2d}} \frac{1}{3} \right\rfloor$$

And,

$$N^* = \max \{ N \in \mathbb{N}^+ \text{ s.t. } \lambda_1^2 - Nc_{speed} > 0 \} \tag{3.8}$$

BATs and non-algo traders demand liquidity when there is a large binding tick size.

In BCS, the depth at the BBO is one share, because the first share has a competitive price. The second share at that price, which faces lower execution probability and higher adverse selection costs, is not profitable. The discrete tick size in our model raises the profit of liquidity supply above the profit of stale-quote sniping for the first share, and generates a depth of multiple shares.

In BCS, the number of HFTs is determined by $\lambda_1 s^* - Nc_{speed} = 0$, where $s^*$ is the break-even bid-ask spread. In our model, $N$ is determined by $\lambda_1^d - Nc_{speed} > 0$. When tick size is binding, $d > s^*$, so tick size leads to more entries of HFTs. Taken together, our model contributes to the literature by identifying a queuing channel of speed competition, in which HFTs race for top queue positions to capture the rents created by tick size.

We assume that BATs do not queue after the first share to get the analytical solution of the queuing equilibrium. The intuition when BATs can queue more than one share, however, remains the same. As long as we do not allow BATs to queue for an infinitely long time, BATs will demand

\textsuperscript{7}Certainly, the comparison also needs to control for price, because stocks with the same nominal bid-ask spread may have a different proportional bid-ask spread.
liquidity with positive probability. In Section 3.4, we show that BATs always supply liquidity when tick size is small.

### 3.4 Equilibrium types under a small tick size

Starting from this section, we reduce the tick size to $d/3$. BATs then always choose to supply liquidity by establishing price priority over HFTs, except when the adverse selection risk is very low. Corollary 4 shows that a small tick size of $d/3$ is still binding when $\frac{\lambda_I}{\lambda_J} > 5$.

**Corollary 4 (Small Binding Tick Size)**: When $\Delta = d/3$ and $\frac{\lambda_I}{\lambda_J} > 5$, the equilibrium bid-ask spread equals the tick size. $N_s^*$ HFTs jointly supply $Q_s^*$ units of sell limit orders at $a_{s,t}^* = \nu_t + d/6$ and $Q_s^*$ units of buy limit orders at $b_{s,t}^* = \nu_t - d/6$, where:

$$Q_s^* = \max\{Q \in \mathbb{N}^+ \text{s.t.} (\frac{\lambda_I}{\lambda_I + 2\lambda_J})^Q \frac{d}{6} - \frac{1}{2} [1 - (\frac{\lambda_I}{\lambda_I + 2\lambda_J})^Q] \frac{5d}{6} > 0\}$$

$$= \max\{Q \in \mathbb{N}^+ \text{s.t.} (\frac{\lambda_I}{\lambda_I + 2\lambda_J})^Q > \frac{5}{7}\}$$

$$= \lceil \log_{(\frac{\lambda_I}{\lambda_I + 2\lambda_J})} \frac{5}{7} \rceil < Q^*$$

(3.9)

And,

$$N_s^* = \max\{N \in \mathbb{N}^+ \text{s.t.} \lambda_I \frac{d}{6} - N c_{\text{speed}} > 0\} < N^*$$

(3.10)

Compared with Proposition 9, a small tick size reduces revenue from liquidity supply from $d/2$ to $d/6$, increases the cost of being sniped from $d/2$ to $5d/6$, and reduces the queue length from $Q^*$ to $Q_s^*$. Figure 3.2 shows that $Q_s^*$ is approximately $1/3$ of $Q^*$. A small tick size also discourages the entry of HFTs. $N_s^*$ is approximately $1/3$ of $N^*$, because HFTs’ expected profit per unit of time decreases from $\lambda_I \frac{d}{2}$ to $\lambda_I \frac{d}{6}$.

When $1 < \frac{\lambda_I}{\lambda_J} < 5$, the break-even bid-ask spread is larger than one tick. To profit from the
bid-ask spread, HFTs have to quote the following bid-ask spread: \(^8\)

\[
\begin{align*}
\frac{d}{2} & \quad \frac{1}{1 - \beta} < \frac{\lambda_l}{\lambda_f} < 5 \\
\frac{5d}{6} & \quad \frac{1}{5(1 - \beta)} < \frac{\lambda_l}{\lambda_f} < \frac{1}{1 - \beta} \\
\frac{7d}{6} & \quad 1 < \frac{\lambda_l}{\lambda_f} < \frac{1}{5(1 - \beta)}
\end{align*}
\]

(3.11)

Figure 3.3 shows that the bid-ask spread quoted by HFTs weakly decreases with \(\frac{\lambda_l}{\lambda_f}\), because an increase in \(\frac{\lambda_l}{\lambda_f}\) decreases adverse selection risk. The bid-ask spread quoted by HFTs increases weakly with the fraction of BATs, because BATs’ strategies for minimizing transaction costs reduce HFTs’ expected profit from liquidity supply. Interestingly, when the adverse selection risk or the fraction of BATs is high, HFTs effectively cease supplying liquidity by quoting a bid-ask spread that is wider than the size of a jump. In the following sections, we elaborate the equilibrium types when tick size is not binding.

### 3.5 Make-take spread

In this section, we develop a new concept make-take spread, and we use the concept to explain why BATs never demand liquidity from HFTs when the tick size is not binding. Without loss of generality, we consider the decision for a BAT who wants to buy. We start from the case when \(\frac{1}{1 - \beta} < \frac{\lambda_l}{\lambda_f} < 5\), for which HFTs need to quote an ask price of \(v_t + d/2\) and a bid price of \(v_t - d/2\) to profit from the bid-ask spread.

A BAT can choose to accept the ask price of \(v_t + d/2\), but submitting a limit order to buy at \(v_t + d/6\) is always less costly, because a buy limit order above fundamental value immediately attracts HFTs to submit market orders to sell. This flash limit order immediately executes like a market order, but with lower cost.

Why do HFTs quote a sell price of \(v_t + d/2\), but are willing to sell at \(v_t + d/6\) using market orders? It is because HFTs’ limit price to sell includes the costs of adverse selection risk. An offer

---

\(^8\)We defer the derivation of the boundary condition for HFTs’ bid-ask spread to Sections 3.4 to 3.6. Another way to bypass tick size constraints is to randomize quotes immediately above and below the break-even bid/ask spread. In this paper, we consider only stationary HFT quotes.
to sell is more likely to be executed when $v_t$ jumps up. HFTs would accept a lower sell price when they demand liquidity, because immediate execution reduces adverse selection risk.

Flash limit orders exploit the make-take spread, which measures the price difference between the traders’ willingness to list an offer and their willingness to accept an offer conditional on the trade direction (e.g., sell). We discover make-take spread because liquidity suppliers can demand liquidity. This new feature reflects reality in contemporary electronic platforms. In most exchanges, every trader can supply liquidity and encounter very limited, if any restrictions when demanding liquidity (Clark-Joseph et al. (2017), Forthcoming).

BATs are able to quote more aggressive prices than HFTs because they have lower opportunity costs for supplying liquidity. BATs have to buy or sell, and they supply liquidity as long as its cost is less than demanding liquidity. BATs lose $d/6$ by using flash limit orders, but the cost of flash limit orders is lower than paying a half bid-ask spread $d/2$. O’Hara (2015) finds that sophisticated non-HFTs cross the spread only when it is absolutely necessary. The make-take spread provides one interpretation for why sophisticated non-HFTs seldom cross the bid-ask spread.

When $1 < \frac{\lambda}{\lambda_i} < \frac{1}{1-\beta}$, the half bid-ask spread quoted by HFTs are higher than $d/2$, leaving more price levels for BATs to use flash limit orders. Therefore, BATs never demand liquidity as long as HFTs quote a bid-ask spread that is wider than one tick.

### 3.6 Flash equilibrium versus undercutting equilibrium

In the previous section, we show that flash orders strictly dominate market orders. In this section, we show that, under some conditions, BATs can further reduce their transaction costs by submitting limit orders that do not cross the midpoint. These regular limit orders do not get immediate execution but stay in the LOB to wait for market orders.

We consider BATs’ choice between flash and regular limit orders. In the flash equilibrium, BATs use flash limit orders to supply liquidity to HFTs, and HFTs supply liquidity to non-algos. In the undercutting equilibrium, BATs use regular limit orders to supply liquidity to non-algos and other BATs, whereas HFTs follow complex strategies with frequent order additions and cancellations. For simplicity, we focus on the case when $\frac{1}{1-\beta} < \frac{\lambda}{\lambda_i} < 5$, for which HFTs need to quote an ask price of $v_t + d/2$ and a bid price of $v_t - d/2$ to profit from the bid-ask spread. In this case, BATs
only need to consider two price levels: a flash limit order (e.g., \( v_t + d/6 \) to buy) or a regular limit order (e.g., \( v_t - d/6 \) to buy).

3.6.1 Flash equilibrium

In Proposition 10, we characterize the flash equilibrium. Starting from now, we only characterize the equilibrium outcome. BATs’ response to off-equilibrium paths are defined in the proofs.

**Proposition 10 (Flash Equilibrium)** When \( \Delta_{1} = \frac{d}{3} \) and \( \frac{1}{1 - \beta} < \frac{\lambda_{I}}{\lambda_{J}} < \frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta} \), the equilibrium is characterized as follows:

1. BAT buyers submit limit orders at \( v_t + \frac{d}{2} \) and BAT sellers submit limit orders at price \( v_t - \frac{d}{6} \).

2. \( N^*_f \) HFTs jointly supply \( Q^*_f \) units of sell limit orders at \( a^*_f,t = v_t + \frac{d}{2} \) and \( Q^*_f \) units of buy limit orders at \( b^*_f,t = v_t - \frac{d}{2} \), where:

\[
Q^*_f = \max\{Q \in \mathbb{N}^+ \text{s.t.} (\frac{(1 - \beta)\lambda_{I}}{(1 - \beta)\lambda_{I} + 2\lambda_{J}})^Q \frac{d}{2} - \frac{1}{2} \left[1 - (\frac{(1 - \beta)\lambda_{I}}{(1 - \beta)\lambda_{I} + 2\lambda_{J}})^Q\right] \frac{d}{2} > 0\}
\]

\[
= \max\{Q \in \mathbb{N}^+ \text{s.t.} (\frac{(1 - \beta)\lambda_{I}}{(1 - \beta)\lambda_{I} + 2\lambda_{J}})^Q > \frac{1}{3}\}
\]

\[
= \lfloor \log_{\left(\frac{(1 - \beta)\lambda_{I}}{(1 - \beta)\lambda_{I} + 2\lambda_{J}}\right)} \frac{1}{3} \rfloor < Q^*
\]  

(3.12)

And,

\[
N^*_f = \max\{N \in \mathbb{N}^+ \text{s.t.} \beta \lambda_{I} \frac{d}{6} + (1 - \beta)\lambda_{I} \frac{d}{2} - N_{c\text{speed}} > 0\} < N^*
\]  

(3.13)

3. HFTs participate in three races: (1) HFTs race to fill the queue when the depth at \( v_t + \frac{d}{2} \) or \( v_t - \frac{d}{2} \) becomes less than \( Q^*_f \). (2) HFTs race to take the liquidity offered by flash limit orders. (3) After a value jump, HFTs who supply liquidity race to cancel the stale quotes, whereas stale-quote snipers race to pick off the stale quotes.

In Proposition 10, we first derive the boundary between the flash equilibrium and the undercutting equilibrium. Figure 3.4 illustrates the boundary in. BATs choose flash limit orders over regular limit orders when adverse selection risk is high. Intuitively, flash limit orders execute immediately, but it costs \( d/6 \) relative to the midpoint; regular limit orders capture a half bid-ask
spread of $d/6$ if executed against a non-HFT, but it is also subject to adverse selection risk. BATs tend to choose flash limit orders when the adverse selection risk is high. Figure 3.4 also shows BATs tend to choose regular limit orders when $\beta$ decreases. Intuitively, because non-algo traders use only market orders, a regular limit order on the book would have higher execution probability before a value jump as the fraction of non-algo traders increases.

Proposition 10 identifies a unique type of speed competition led by tick size: racing to be the first to take the liquidity offered by flash limit orders. If price is continuous, any buy limit order price above fundamental value would prompt HFTs to sell. In our model with discrete tick size, a BAT needs to place the buy limit order at $v_t + d/6$, which drives the speed race to capture the rent of $d/6$ through demanding liquidity.

In the literature, HFTs demand liquidity when they have advance information to adversely select other traders (BCS; Foucault et al. (2017), Forthcoming; Menkveld and Zoican (2017)). Consequently, HFTs’ liquidity demand often has negative connotations. Our model shows that HFTs can demand liquidity without adversely selecting other traders. Instead, the transaction cost is lower for BATs when HFTs demand liquidity than when HFTs supply liquidity. Therefore, researchers and policy makers should not evaluate the welfare impact of HFTs simply based on liquidity supply versus liquidity demand.

As BATs no longer demand liquidity from HFTs, HFTs respond to the reduced liquidity demand and higher adverse selection cost by decreasing their depth to $Q_f^*$. The profit to take liquidity from BATs, $d/6$, is less than the profit to supply liquidity to BATs at $d/2$ when the tick size is $\Delta_0$. A smaller tick size, $\Delta_1$, reduces the profit for HFTs, thereby reducing the number of HFTs.

3.6.2 Undercutting equilibrium

In flash equilibrium, the LOB only has one stable state. In the undercutting equilibrium, the LOB transits across different states. As indicated in Proposition 10 BATs choose regular limit orders over flash limit orders when adverse selection risk or $\beta$ is low. In the undercutting equilibrium, their limit orders stay in the LOB, and their decisions, as well as those of HFTs, depend on the state of the LOB. Our technical assumption that BATs never queue at the second position reduces the number of states. Still, the solution is complicated. We focus on deriving the equilibrium
strategies of HFTs, as Proposition 10 and its proof in the Appendix demonstrate the strategy of BATs in undercutting equilibrium. BATs choose regular limit orders over flash limit orders when
\[
\frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta} < \frac{\lambda_I}{\lambda_J} < 5.
\]

To show the equilibrium strategy of HFTs, we first define the state of the LOB as \((i, j)\). Here \(i\) represents the number of BATs’ limit orders on the same side of the LOB, and \(j\) denotes the number of BATs’ limit orders on the opposite side of the LOB. For example, for a HFT who wants to buy, \(i\) represents the number of BATs’ limit orders on the bid side, and \(j\) represents the number of BATs’ limit orders on the ask side. The LOB then has four states:

- \((0, 0)\): No limit order from BATs
- \((1, 0)\): A BAT limit order on the same side
- \((0, 1)\): A BAT limit order on the opposite side
- \((1, 1)\): BAT limit orders on both sides

When \(\frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta} < \frac{\lambda_I}{\lambda_J} < 5\), HFTs quote a half bid-ask spread of \(d/2\), as a half bid-ask spread of \(d/6\) loses money. Similar to the queuing equilibrium and the flash equilibrium, HFTs’ decision to supply liquidity depends on the payoff of the liquidity supply relative to the outside option of sniping. The new feature of the undercutting equilibrium is that HFTs’ decision also depends on the status of the LOB. We denote the payoff of the \(Q^{th}\) share to supply liquidity at half the bid-ask spread \(d/2\) as \(LP^{(i,j)}(Q)\), and the payoff to the snipers of the \(Q^{th}\) share as \(SN^{(i,j)}(Q)\). The HFT’s strategy depends on \(D^{(i,j)}(Q) = LP^{(i,j)}(Q) - SN^{(i,j)}(Q)\).

Figure 3.5 illustrates how \(D^{(i,j)}(Q)\) changes with the six types of events defined in Equation (3.1). For example, consider \(D^{(0,0)}(Q)\) for an HFT on the ask side of the LOB.

1) A BAT buyer submits a limit order at \(v_t - d/6\), which changes \(D^{(0,0)}(Q)\) to \(D^{(0,1)}(Q)\).
2) A BAT seller undercutts the ask side at \(v_t + d/6\), which changes \(D^{(0,0)}(Q)\) to \(D^{(1,0)}(Q)\).
3) A non-algo buyer submits a market buy order, which moves the queue position forward by one unit. \(D^{(0,0)}(Q)\) to \(D^{(0,0)}(Q - 1)\).
4) A non-algo seller submits a market sell order, which does not affect \(D^{(0,0)}(Q)\) as the LOB on the bid side is refilled immediately by HFTs.
5) In an upward value jump, a liquidity providing HFT on the ask side gains \(-\frac{d N - 1}{2 N}\), a stale-
quote sniper gains $d \frac{1}{2N}$, and the difference between them is $-d$.  

6) In a downward value jump, the liquidity supplier cancels the limit order, thereby changing the value of both the liquidity supply and stale-quote snipping to zero.  

These six types of events and the four states of the LOB are the key features of the undercutting equilibrium, which we summarize in Proposition 3. To simplify the notation, we use $p_1 \equiv \frac{1}{2} \hat{\beta} \lambda_l$ to denote the arrival probability of a BAT buyer or seller, $p_2 \equiv \frac{1}{2} (1 - \hat{\beta}) \lambda_l$ to denote the arrival probability of a non-algo trader to buy or sell, and $p_3 \equiv \frac{1}{2} \lambda_j$ to denote the probability of an upward or downward value jump.  

**Proposition 11** When $\Delta_t = \frac{d}{3}$ and $\frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta} < \frac{\lambda_l}{\lambda_j} < 5$, the equilibrium is characterized as follows:  

1. HFTs’ strategy:
   a. Spread: HFTs quote ask price at $v_t + \frac{d}{2}$ and bid price at $v_t - \frac{d}{2}$.
   b. Depth: The following system of equations determines the equilibrium depth in each state.

   \[
   \begin{align*}
   D^{(0,0)}(Q) &= \max\{0, p_1 D^{(1,0)}(Q) + p_2 D^{(0,1)}(Q - 1) + p_3 \left( -\frac{d}{2} \right) + p_3 \cdot 0 \} \\
   D^{(1,0)}(Q) &= \max\{0, p_1 D^{(1,0)}(Q) + p_2 D^{(0,0)}(Q) + p_3 \left( -\frac{d}{2} \right) + p_3 \cdot 0 \} \\
   D^{(0,1)}(Q) &= \max\{0, p_1 D^{(0,1)}(Q) + p_2 D^{(1,0)}(Q) + p_3 \left( -\frac{d}{2} \right) + p_3 \cdot 0 \} \\
   D^{(1,1)}(Q) &= \max\{0, p_1 D^{(1,1)}(Q) + p_2 D^{(0,1)}(Q) + p_3 \left( -\frac{d}{2} \right) + p_3 \cdot 0 \}
   \end{align*}
   \]  

   \[\text{(3.14)}\]

   ii. Difference in value for immediate execution: $D^{(0,0)}(\cdot) = D^{(0,1)}(\cdot) = \frac{d}{2}$.

   iii. Equilibrium depth as a function of the difference in value:

   \[
   Q^{(i,j)} = \max\{Q \in \mathbb{N}^+ | D^{(i,j)}(Q) > 0 \} \quad i = 0, 1; j = 0, 1
   \]

   c. In equilibrium there are $N_u^* < N^*$ HFTs.

2. BATs who intend to buy (sell) submit limit orders at price $v_t - \frac{d}{6}(v_t + \frac{d}{6})$ if no existing limit orders sit at the price level, or buy (sell) limit orders at price $v_t + \frac{d}{6}(v_t - \frac{d}{6})$ otherwise.\(^9\)

\(^9\)After an upward (downward) jump with size $d$, we assume BATs buy (sell) undercutting orders at $v_t - d/6(v_t +
The depth from HFTs depends on \( D^{(i,j)}(Q) \). \( D^{(i,j)}(Q) \) is defined using the system Equation (3.14), because the value difference in each state also depends on the value differences in other states. Equation (3.14) contain the max\{0, . \} as HFTs do not queue at the \( Q^{th} \) position once the expected payoff is below 0.

We present the solution for \( D^{(i,j)}(Q) \) for any \( i,j \), and \( Q \) in the Appendix. Here we use a numerical example to present the main intuition of the undercutting equilibrium. Figure 3.6 shows that the value of the liquidity supply decreases in \( Q \), while the value of stale-quote sniping increases in \( Q \). HFTs supply liquidity as long as \( LP^{(i,j)}(Q) > SN^{(i,j)}(Q) \). For example, in state \((o, o)\), the LOB has a depth of two shares.

Figure 3.6 also shows that \( LP^{(i,j)}(Q) \) and \( SN^{(i,j)}(Q) \) also depend on the state of the LOB. As the undercutting limit orders from BATs can change the states of the LOB, HFTs can add or cancel their limit orders even when the fundamental value stays the same. A comparison between Panel A and Panel B and between Panel C and Panel D of Figure 3.6 shows that an undercutting order reduces HFTs’ depth on the same side of the LOB by approximately one share. Intuitively, when a BAT submits an undercutting order, the execution priority for all HFTs on the same side of the book decreases by one share. An HFT who used to quote the last share at the half bid-ask spread \( d/2 \) has to cancel, because the share become unprofitable after the arrival of the undercutting order. For the same reason, once an undercutting order from a BAT executes, HFTs race to submit one more share at the half bid-ask spread \( d/2 \), because the execution priority in the LOB increases by one. One new feature of the undercutting equilibrium is the frequent order addition or cancellation of HFTs’ limit orders in the absence of a change in fundamental value.

One driver of HFTs’ frequent additions and cancellations is small tick size. When tick size is \( d/6 \) will be cancelled and resubmitted at price \( v_t + 5d/6(v_t - 5d/6) \) to follow the value jump. Alternative BATs strategy does not change the equilibrium.

An undercutting BAT order on the opposite side of the LOB has an indirect effect. For example, in state \((1, 1)\), a BAT buyer takes liquidity at price \( v_t + d/6 \) and changes the state to \((o, 1)\), which enables an HFT limit sell order at price \( v_t + d/2 \) to trade with the next buy market order from a non-algo trader. In state \((1, o)\), a BAT buyer chooses to submit a limit order at price \( v_t - d/6 \), which changes the state to \((1, 1)\). An HFT limit sell order at price \( v_t + d/2 \) then needs to wait at least one more period for execution. More generally, an undercutting BAT limit buy (sell) order may attract future BAT sellers (buyers) to demand liquidity, making future BATs less likely to undercut HFTs. In turn, the value of liquidity supply increases relative to sniping, thereby incentivizing HFTs to supply larger depth. This indirect effect is so small that it does not affect depth in our numerical example, because the number of shares is an integer. It is possible for a depth of \((1, 1)\) to be higher than \((1, o)\) for numerical values such as \( \lambda_1 = 4.9 \) and \( \beta = 0.06 \), and the results are available upon request.
binding, BATs cannot achieve execution priority over HFTs who are already in the queue. When tick size is small, BATs can achieve price priority over HFTs, which induces HFTs to cancel their earlier orders and to add new ones in response to the undercutting orders from BATs.

When \( \frac{1}{5(1 - \beta)} < \frac{\lambda_I}{\lambda_J} < \frac{1}{1 - \beta} \), HFTs quote \( \frac{5d}{6} \), and BATs’ strategies follow the intuition outlined above, where they choose between flash limit orders and regular limit orders. The only main difference is that the four price levels between \( v_t + \frac{5d}{6} \) and \( v_t - \frac{5d}{6} \) increase the states to \( 2^4 = 16 \). We do not report the results for brevity but they are available upon request. In Section 3.7, we discuss the case when the break-even spread equals \( \frac{7d}{6} \).

### 3.7 Stub quotes and mini-flash

In Proposition 12, we show that HFTs quote a bid-ask spread wider than the size of the jump when adverse selection risk is high or the fraction of BATs is large. We call such quotes stub quotes. A mini-flash crash occurs when a market order hits a stub quote. In our model, the size of the mini-flash crash is \( \frac{7d}{6} \), because the size of a value jump is \( d \). An increase in the support of jump size can lead to stub quotes further away from the midpoint, thereby creating mini-flash crashes of larger size. Such an extension adds mathematical complexity without conveying new intuition.

**Proposition 12 (Stub Quotes and Mini-Flash Crash)** When \( \delta = \frac{d}{3} \) and \( 1 < \frac{\lambda_I}{\lambda_J} < \frac{1}{5(1 - \beta)} \), the equilibrium is characterized as follows.

1. HFTs quote a half bid-ask spread of \( \frac{7d}{6} \).
2. A BAT buyer (seller) quotes \( v_t - \frac{5d}{6} \) if the price level has no limit orders. Otherwise, the BAT buyer (seller) submits a flash limit order at price \( v_t + \frac{d}{6} \) to provide liquidity.
3. Compared with the case when \( \Delta_0 = d \), the transaction cost for non-algo traders increases, but the average transaction cost for non-HFTs decreases.
4. The probability of mini-flash crashes decreases in \( \frac{\lambda_I}{\lambda_J} \). The probability of mini-flash crashes first increases in \( \beta \) and then decreases in \( \beta \).

Proposition 12 shows that HFTs are more likely to quote stub quotes when adverse selection risk is high. A higher adverse selection risk prompts HFTs to quote stub quotes through two
channels. First, HFTs have to quote a wider bid-ask spread to reach the break even point. Second, when HFTs’ quotes are wider than one tick, BATs are able to quote more aggressive prices than HFTs. HFTs then need to further widen the bid-ask spread due to reduced liquidity demand.

When HFTs quote stub quotes, BATs have six price levels to choose from. Fortunately, we are able to obtain analytical solutions for the BATs’ strategy. Consider the decision for a BAT buyer. We find that the buyer chooses to queue at $v_t - 5d/6$ if the price level contains no limit orders. The sniping cost is as low as $d/6$, and the BAT buyer can earn a half bid-ask spread of $5d/6$ if a non-algo trader arrives. When $v_t - 5d/6$ contains a limit order, the BAT buyer will use a flash limit order at $v_t + d/6$ to obtain immediate execution with a transaction cost of $d/6$.\(^{11}\) We show in the proof that BATs never quote at $v_t - d/2$ and $v_t - d/6$ as the execution cost is always higher than $d/6$. Flash buy limit orders at price $v_t + d/6$ also strictly dominate more aggressive flash limit orders of $v_t + d/2$ and $v_t + 5d/6$, because a limit order price of $v_t + d/6$ is aggressive enough to trigger immediate execution.

In Section 3.4, we find that the transaction costs for both BATs and non-algo traders are $d/2$ when tick size is $d$. A decrease in tick size to $d/3$ increases the transaction cost for non-algo traders. A non-algo trader pays $5d/6$ when an order is she executed against a BAT and pays $7d/6$ if a stub quote is encountered. Meanwhile, a decrease in tick size to $d/3$ decreases the transaction cost for BATs. BATs’ maximum transaction cost is $d/6$ if they use flash limit orders, although the cost is lower if they quote a half bid-ask spread of $5d/6$. Overall, we find that the average transaction cost decreases with tick size. Figure 3.3 shows that the proportion of BATs needs to be at least $4/5$ for stub quotes to occur. Non-algo traders’ maximum transaction cost is $7d/6$ if they hit stub quotes. The average transaction cost for non-HFTs is then at most $11d/30$ ($\frac{4}{5} \times \frac{d}{6} + \frac{1}{5} \times \frac{7d}{6}$), which is lower than $d/2$. Therefore, a reduction in tick size reduces non-HFTs’ average transaction costs, but increase the dispersion and volatility of their transaction costs.

An increase in adverse selection risk unambiguously increases the probability of mini-flash crashes. Figure 3.3 in Section 3.4 show that stub quotes are more likely to occur when there higher adverse selection risk. Conditional on stub quotes occurring, Figure 3.6 reveals another

\(^{11}\)This result is certainly a consequence of our simplifying assumption that BATS cannot queue for a second share. However, BATs should always have higher incentives to use flash limit orders when $v_t - 5d/6$ contains a limit order, because the second share has a lower probability of executing against a non-algo trader and a higher probability of executing against a sniper, whereas a flash limit order always incurs a constant cost of $d/6$. 

80
channel for adverse selection risk to increase the number of mini-flash crashes. An increase in adverse selection risk implies more value jumps relative to the arrival rate of non-algo traders. During an upward (downward) value jump, BATs’ limit orders on the bid (ask) side are all sniped and only stub quotes remain. If the limit orders from BATs fail to reconvene before a non-algo trader arrives, the market order from the non-algo trader hits the stub quote and causes a mini-flash crash.

The proportion of BATs, $\beta$, have an ambiguous effect on the probability of flash crashes because of two competing effects. On the one hand, Figure 3.3 in Section 3.4 shows that a larger $\beta$ increases the probability for stub quotes as HFTs face less liquidity demand. On the other hand, a larger $\beta$ decreases the probability of hitting stub quotes, because BATs never demand liquidity from HFTs. For example, mini-flash crashes never occur when $\beta = 0$ or $\beta = 1$. Therefore, mini-flash crashes need both BATs and non-algo traders. Figure 3.6 shows the simulated intensity of mini-flash crashes with respect to $\beta$. For each $\beta$, we first uniformly draw $100 \frac{\lambda_I}{\lambda_J}$ from $[1, 5]$, the support of the adverse selection risk in our paper. For each $\frac{\lambda_I}{\lambda_J}$, we simulate the first 100,000 trades. For all 10 million simulations, we count the number of trades that hit the stub quotes relative to the total number of trades.

Figure 3.7 shows that mini-flash crashes are most likely to occur when $\beta$ is approximately 0.95, and we normalize this crash intensity to 1. The black square line shows that the intensity is hump-shaped with respect to $\beta$. The circle line shows that majority of mini-flash crashes occur after a value jump. An upward value jump removes BATs’ limit orders from the ask side and a downward jump removes BATs’ limit orders from the bid side. If BATs’ limit orders do not reconvene in the LOB, a market buy (sell) order from non-algo trader would hit stub quotes. Therefore, most of the upward (downward) mini-flash crashes occur after an upward (downward) value jump. Only a small amount of crashes are due to BATs’ liquidity being used up by non-algo traders.

An effective way to prevent a mini-flash crash is a trading halt to let the trading interest of BATs reconvene. The triangle line in Figure 3.7 shows the intensity of mini-flash crashes with trading halts. We impose the trading halt after a value jump, and the market reopens after 10 orders arrive at the market. We find that such a trading halt reduces mini-flash crashes by about 90%.
3.8 Predictions and policy implications

Our model rationalizes a number of puzzles in the literature on HFTs and generates new empirical predictions that can be tested. In Section 3.8.1, we summarize the predictions on who supplies liquidity and when. In Subsection Section 3.8.2, we examine the predictions on liquidity demand. In Section 3.8.3, we evaluate the predictions on liquidity. In section 3.8.4, we discuss the use of the cancellation ratio as the cross-sectional proxy for HFTs’ activity.

3.8.1 Liquidity supply

Our model shows that who provides liquidity depends on the tick size, adverse selection risk, the motivation of the trade, and the speed of the trade. In Prediction 1, we posit that BATs dominate liquidity supply when tick size is not binding.

**Prediction 1** (*Price Priority*) When tick size is not binding, Non-HFTs are more likely to establish price priority in liquidity supply.

Speed advantages in the LOB reduce HFTs’ adverse selection costs (see Jones (2013) and Menkveld (2016) surveys), inventory costs (Brogaard et al. (2015)), and operational costs (Carrion (2013)). These reduced costs of intermediation raise the concern that “HFTs use their speed advantage to crowd out liquidity supply when the tick size is small and stepping in front of standing limit orders is inexpensive” (Chordia et al. (2013), p. 644). However, Brogaard et al. (2015) find that non-HFTs quote a tighter bid-ask spread than HFTs, and Yao and Ye (2017) find that non-HFTs are more likely to establish price priority over HFTs as the tick size decreases. We find that the opportunity cost of supplying liquidity can reconcile the contradiction between the empirical results and the channels of speed competition. BATs incur lower opportunity costs when supplying liquidity. When they implement a trade, they supply liquidity as long as it is less costly to demand liquidity. The make-take spread that we introduce in Section 3.5 indicates that BATs never demand liquidity from HFTs when tick size is not binding.

**Prediction 2** (*Queuing*) HFTs crowd out non-HFTs’ liquidity supply when tick size is binding, that is, when the tick size is large or adverse selection risk is low.
When tick size is binding, HFTs’ speed advantage allows them to establish time priority at the same price. Yao and Ye (2017) find that tick size is more likely to be binding when tick size increases. They also find that a large tick size crowds out non-HFTs’ liquidity supply. Both results provide evidence to support Prediction 2.

Hoffmann (2014), Han et al. (2014), Bernales (2014), and Bongaerts and Van Achter (2016) find that HFTs have lower adverse selection costs than non-HFTs. Yao and Ye (2017), however, find that HFTs do not have a comparative advantage in providing liquidity for stocks with higher adverse selection risk. In Prediction 2, we provide the economic mechanism to reconcile this inconsistency. Comparing Corollary 4 with Proposition 10 and Proposition 11, we find that the tick size is more likely to be binding when adverse selection risk is low. A binding tick size helps HFTs to supply liquidity through time priority. An increase in adverse selection risk raises the break-even bid-ask spread above one tick, allows non-HFTs to undercut HFTs, and decreases HFTs’ liquidity supply.

In Prediction 3, we address who provides liquidity during a mini-flash crash.

**Prediction 3** *(Stub Quotes and Mini-Flash Crashes)* A mini-flash crash is more likely to occur when the adverse selection risk is high or when the tick size is small. During a mini-flash crash, HFTs supply liquidity and non-HFTs demand liquidity. A downward (upward) mini-flash crash is more likely to follow a downward (upward) value jump.

A comparison of Propositions 9 and 12 shows that stub quotes are more likely to occur when the tick size is small. When the tick size is large, BATs cannot establish execution priority over HFTs. When the tick size is small, BATs can establish price priority over HFTs, which increases the adverse selection costs for HFTs through two channels. First, when BATs can undercut HFTs, they no longer demand liquidity from HFTs. HFTs then face reduced liquidity demand but the risk of value jump stay the same. Second, the undercutting orders by BATs reduce the execution priority of HFTs. In turn, HFTs’ limit orders face lower execution probability and higher sniping cost. When the adverse selection cost is high enough, HFTs effectively quit liquidity supply by quoting stub quotes. HFTs are more likely to quote stub quotes when adverse selection risk is high as higher adverse selection risk widens the break-even bid-ask spread; a wider break-even bid-ask spread also allows BATs to undercut HFTs, which further increases the adverse selection costs.
for HFTs. Because BATs do not continuously supply liquidity in the market, non-algo traders’ market orders can hit stub quotes and cause mini-flash crashes. A high adverse selection risk also implies more value jumps relative to the arrival rate of non-HFTs. Non-algo traders’ market orders are more likely to hit stub quotes after value jumps, because value jumps clear BATs’ limit orders on the side of the jump.

In cross-section, our model predicts that stocks with smaller tick sizes or higher adverse selection risk are more likely to incur mini-flash crashes. This cross-sectional pattern has not been tested. In time series, our model predicts that an initial downward (upward) jump increases the probability of a downward (upward) mini-flash crash. The downward (upward) jump clears the LOB on the bid (ask) side, making the market orders from non-algo traders more likely to hit stub quotes.

Brogaard et al. (Forthcoming) analyze the time series pattern of mini-flash crashes. They show that, 20 seconds before a mini-flash crash, HFTs neither demand nor supply liquidity, whereas non-HFTs demand and supply the same amount of liquidity; 10 seconds before a mini-flash crash, HFTs demand liquidity from non-HFTs; at the time of a mini-flash crash, HFTs supply liquidity to non-HFTs, but at a much wider bid-ask spread. The authors also find that the liquidity supply from the mini-flash crash is profitable. This evidence is consistent with the theoretical mechanism for mini-flashes crash that we document. (1) In normal times, non-HFTs dominate both liquidity supply and liquidity demand; (2) slightly before a mini-flash crash, HFTs demand liquidity and remove limit orders from BATs; (3) a mini-flash crash occurs when a non-algo trader’s market order hits HFTs’ stub quotes, thus HFTs profit when a mini-flash crash occurs.

Our interpretations of mini-flash crashes are consistent with both negative and positive framing of the role of HFTs in a mini-flash crash. Brogaard et al. (2017) suggest that HFTs supply liquidity in extreme price movements, while Ait-Sahalia and Sağlam (2017) suggest that HFTs withdraw liquidity supply when it is most needed. Both views, however, suggest that mini-flash crashes occur when the market orders of non-HFTs hit the stub quotes from HFTs.

Our interpretation of mini-flash crashes has two additional features that are consistent with economic reality. First, markets recover quite quickly from mini-flash crashes. In our model, mini-flash crashes disappear when the limit orders from BATs replenish the LOB. Second, Nanex,
the firm that invented the concept of mini-flash crash, finds that mini-flash crashes are equally likely to be upward as downward. Indeed, even during the famous Flash Crash on May 6, 2010, in which the Dow Jones plunged 998.5 points, some stocks, including Sotheby’s, Apple Inc., and Hewlett-Packard, increased in value to over $100,000 in price (SEC, 2010). In our model, upward and downward mini-flash crashes are equally likely, even though downward mini-flash crashes are more likely to occur conditional on an initial downward value jump.

### 3.8.2 Liquidity demanding

Our model discovers a new channel of speed competition to demand liquidity. In Prediction 4, we summarize the empirical implications of this new channel.

**Prediction 4** *(Speed Competition of Taking Liquidity)* Non-HFTs are more likely than HFTs to supply liquidity at price levels that cross the midpoint (flash limit orders). HFTs are also more likely to demand liquidity from flash limit orders, but they do not adversely select these orders.

Latza et al. (2014) find evidence consistent with Prediction 4. They classify a market order as “fast” if it executes against a standing limit order that is less than 50 milliseconds old. Because of the speed of taking liquidity, it is natural to expect that fast market orders are from HFTs. These authors also find that fast market orders often execute against limit orders that cross the midpoint, and they lead to virtually no permanent price impact.

In Prediction 4, we offer fresh perspectives on the liquidity demand from HFTs. Typically, HFTs demand liquidity when they employ a speed advantage to adversely select liquidity suppliers (BCS; Foucault et al. (2017); Menkveld and Zoican (2017)). Therefore, liquidity demand from HFTs generally has negative connotations of reducing liquidity (Jones (2013); Biais et al. (2014)). We find that HFTs’ liquidity demand does not necessarily adversely select slow traders. Instead, the liquidity demand from HFTs can reduce the transaction costs of non-HFTs. In the flash equilibrium, BATs pay \(d/2\) when HFTs supply liquidity, while BATs only pay \(d/6\) when HFTs demand liquidity.
3.8.3 Liquidity

On April 5, 2012, President Barack Obama signed into law the Jumpstart Our Business Startups (JOBS) Act. Section 106 (b) of the Act requires the SEC to examine the effect of tick size on initial public offerings (IPOs). On October 3, 2016, the SEC implemented a pilot program to increase the tick size from one cent to five cents for 1,200 small- and mid-cap stocks. Proponents of the proposal argue that a larger tick size can improve liquidity (Weild et al. (2012)). In Prediction 5, however, we posit that an increase in tick size decreases liquidity.

**Prediction 5** *A larger tick size increases the depth at the BBO, but it also increases the effective bid-ask spread, the transaction costs paid by liquidity demanders.*

Yao and Ye (2017) find evidence consistent with Prediction 5. Holding the BBO constant, an increase in depth at the BBO implies an increase in liquidity. Yet these authors also find that the quoted bid-ask spread increases after an increase in tick size. When both quoted bid-ask spread and depth increase, the most relevant liquidity measure becomes the effective bid-ask spread, the transaction cost paid by liquidity demanders (Bessembinder (2003)). Our model shows that constrained price competition increases the effective bid-ask spread, which is consistent with the findings in Yao and Ye (2017). Our model prediction, along with the evidence in Yao and Ye (2017), shows that an increase in tick size would not improve liquidity.

Advocates for an increase in tick size also argue that a wider tick size increases market-making profits, supports sell-side equity research and, eventually, increases the number of IPOs (Weild et al. (2012)). We find that a wider tick size increases market-making profits, but the profit belongs to traders with higher transaction speeds. Therefore, a wider tick size is more likely to result in an arms race in latency reduction than in sell-side equity research.

We also find that an increase in tick size harms non-HFTs. An increase in tick size also does not benefit HFTs as the cost of the speed investment dissipates when larger tick size generates higher rents. In our model, non-HFTs trade no matter how large the bid-ask spread may be. In reality, a wider spread may prevent investors with low gains from trading, leading to a further reduction in welfare.

An increase in tick size reduces mini-flash crashes, but it also increases the transaction costs for average trades. A more effective solution to prevent mini-flash crashes would be to slow down the
market, particularly during periods of market stress. In a standard Walrasian equilibrium, price is continuous and time is discrete. Modern financial markets exhibit exactly the opposite structure: price competition is constrained by the tick size, whereas time is divisible at the nanosecond level in electronic trading platforms (Ye et al. (2013)). Making price more continuous and time more discrete would improve liquidity and also prevent mini-flash crashes at the same time.

3.8.4 Cancellation-to-trade ratio as a cross-sectional proxy for HFT activity

The cancellation-to-trade ratio is widely used as a proxy for HFTs’ activities, particularly for HFTs’ liquidity supplying activities (Biais et al. (2014)). Yet Yao and Ye (2017) find that stocks with a higher proportion of liquidity provided by HFTs have a lower cancellation-to-trade ratio. In Prediction 6, we offer one interpretation for this surprising negative correlation.

**Prediction 6** (Cancellation-to-trade Ratio) Stocks with a smaller tick size and higher adverse selection risk have a lower proportion of liquidity provided by HFTs relative to non-HFTs but a higher cancellation-to-trade ratio.

A decrease in tick size decreases the proportion of liquidity provided by HFTs (Prediction 2), but it leads to more order cancellations. Under a large tick size in our model, HFTs do not need to cancel their orders when non-HFTs arrive, because non-HFTs cannot establish time priority over HFTs. A decrease in tick size increases the potential for non-HFTs to undercut HFTs. If non-HFTs submit flash limit orders, HFTs race to take liquidity, and the losers of the race cancel their orders. If non-HFTs submit regular limit orders, HFTs reduce their depth once non-HFTs undercut, and HFTs increase their depth once an undercutting order gets executed. These changes in depth lead to frequent order cancellations. We offer a new interpretation of flickering quotes. Yueshen (2014) shows that flickering quotes occur when new information causes the price to move to a new level. We show that HFTs can cancel orders in the absence of information. Periodic order additions and cancellations also differ from Baruch and Glosten (2013), who rationalize flicking quotes using a mixed-strategy equilibrium. An increase in adverse selection risk, defined as the intensity of value jumps relative to the arrival rate of non-HFTs, also lead to more order cancellations, but
HFTs also provide less liquidity for these stocks. Taken together, we suggest that the cancellation-to-trade ratio should not be used as a cross-sectional measure of HFTs’ activity.

### 3.9 Conclusions

In this paper, we extend BCS by adding two unique characteristics in financial markets: discrete tick size and algorithmic traders who are not HFTs. We discover a queuing channel of speed competition for liquidity supply. BATs are more likely to supply liquidity when tick size is small, because supplying liquidity is less costly than demanding liquidity from HFTs. A large tick size constrains price competition, creates rents for liquidity supply, and encourages speed competition to capture such rents through the time priority rule. Higher adverse selection risk increases the break-even bid-ask spread relative to tick size, which allows BATs to establish price priority over HFTs and reduces the fraction of liquidity provided by HFTs.

We also discover a new channel of speed competition in liquidity demand. HFTs race to demand liquidity from BATs when BATs post flash limit orders to buy above the fundamental value or to sell below the fundamental value. BATs incur lower transaction cost when HFTs demand liquidity than when HFTs supply liquidity. Thus, an evaluation of the welfare impact of HFTs should not be based solely on demand versus supply liquidity. Our results also indicate that the definition of providing versus demanding liquidity blurs in model electronic markets.

Yao and Ye (2017) find that the cancellation ratio, a widely used empirical proxy for HFTs’ activity, has a negative cross-sectional correlation with HFT liquidity supply. We provide a theoretical foundation for their surprising negative correlation. A large tick sizes induces HFTs to race for the top queue position, and HFTs are less likely to cancel orders once they secure this spot. HFTs cancel orders more frequently for stocks with smaller tick sizes, but they also supply less liquidity. Both theoretical and empirical evidence suggests that researchers should not apply the cancellation ratio as a cross-sectional proxy for HFT activity.

We also provide new predictions to be tested. We predict that 1) non-HFTs are more likely than HFTs to supply liquidity at price levels that cross the midpoint, and these limit orders are more likely to be taken by HFTs; 2) a mini-flash crash is more likely to occur for stocks with smaller tick sizes and higher adverse selection risk; 3) an upward (downward) mini-flash crash is more
likely to follow an initial price jump in the same direction.

Our model shows that a larger tick size increases transaction cost and negatively affects non-HFTs. Yet HFTs do not benefit from a larger tick size as an investment in high-speed technology dissipates the rents created by tick size. We challenge the rationale for increasing the tick size to five cents, and we encourage regulators to consider decreasing tick size, particularly for liquid stocks.

Our model is parsimonious. For example, BATs in our model do not have private information and they choose order types only upon arrival. It will be interesting to extending our model toward more realistic setups. Most studies in the finance literature ignore diversity among algorithm traders. We take the initial step to examine algorithmic traders who are not HFTs, and we believe that further examination on the relationship between HFTs and other algorithmic traders would prove to be fruitful.
Figure 3.1: Pricing Grid under Large vs. Small Tick Sizes

This figure demonstrates the pricing grids under a large tick size $d$ and a small tick size $d/3$. The fundamental value of the asset is $v_t$. 

$$v_t + \frac{d}{2}, \quad v_t - \frac{d}{2}$$

Large Tick Size

Small Tick Size
Figure 3.2: Depth and the Adverse Selection Risk under a Binding Tick Size

This figure demonstrates the relation between $Q$, the depth at the BBO, and $R = \frac{\lambda_i}{\lambda_j}$ under a binding tick size. An increase in the investor arrival rate ($\lambda_i$), or a decrease in intensity of jumps ($\lambda_j$), decreases the adverse selection risk and increases the depth. The solid line represents the depth under tick size $d$ and the dashed line represents the depth under tick size $d/3$. 
This figure demonstrates the half bid-ask spread quoted by HFTs as a function of $\beta$ (the fraction of BATs) and $R \equiv \frac{\lambda}{\lambda_f}$ (the arrival intensity of non-HFTs relative to the value jump, a measure of adverse selection risk). When $R \geq 5$, adverse selection risk is low and the tick size is binding. HFTs quote a half bid-ask spread $d/6$ and the spread is independent of the fraction of BATs. When $R < 5$, HFTs’ quoted bid-ask spreads weakly increase with the fraction of BATs and adverse selection risk.
This figure demonstrates two types of equilibrium, undercutting equilibrium and flash equilibrium, when HFTs’ ask price is at $v_t + d/2$ and their bid price is at $v_t - d/2$. In the undercutting equilibrium, BATs place limit buys at $v_t - d/6$ and limit sells at $v_t + d/6$. These limit orders undercut the BBO by one tick and establish price priority in the LOB. In the flash equilibrium, BATs place limit buys at $v_t + d/6$ and limit sells at $v_t - d/6$. These orders cross the midpoint and immediately attract market orders from HFTs. BATs are more likely to cross the midpoint when the fraction of BATs ($\beta$) is high or when the arrival intensity of non-HFTs relative to a value jump ($R \equiv \frac{\lambda_I}{\lambda_f}$) is low, because a high $\beta$ and a low $R$ reduce the potential for a limit order executing with non-HFTs before a value jump. To jumpstart an undercutting equilibrium, the expected transaction cost for a limit order that undercuts one tick must be lower than $d/6$. The short-dashed line, $C(1, o) = d/6$, illustrates the boundary for such a condition.

Figure 3.4: The Undercutting and the Flash Trading Equilibrium
This figure illustrates the dynamics of HFT queuing on $z_i + d/2$. In state $(i, j)$, the number of undercutting BAT orders on the ask side is $i$, while the number on the bid side is $j$. BB and BS represent the arrival of BATs’ buy and sell limit orders, NB and NS represent the arrival of non-algo traders’ buy and sell market orders, and UJ and DJ denote the upward and downward value jumps. The number next to the event is the immediate payoff of the event.
The x-axis is the value of HFT liquidity supply ($LP$) and stale-queue sniping ($SN$) for the four states of the LOB. In $Q(0, 0)$, no BATs undercut HFTs in the LOB. In $Q(1, 0)$, BATs undercut HFTs on the same side of the book. In $Q(0, 1)$, BATs undercut HFTs on the opposite side of the book. In $Q(1, 1)$, BATs undercut both sides of the book. $LP$ decreases in the queue position, while $SN$ increases in the queue position. HFTs supply liquidity as long as $LP > SN$. 

Figure 3.6: Value of Liquidity Supply and Stale-Queue Sniping and Queue Length
This figure shows the intensity of mini-flash crashes with respect to the fraction of BATs. We normalize the highest intensity as 1. For each $\beta$, we uniformly draw 100 samples from $[1,5]$ as $\lambda I$, which is the support of the adverse selection risk in our paper. For each $\lambda I$, we simulate 100,000 trades. For all these 10 million simulations, we count the number of trades hitting the stub quotes relative to the total number of trades. The line with squares shows the intensity for total crashes. The line with circles shows that the majority of mini-flash crashes occur after a value jump (and a small fraction of crashes occur after BATs’ liquidity being consumed by non-algos). The line with triangles shows that trading halts reduce the number of mini-flash crashes. We impose trading halts after each value jump, and the market reopens when the market receives 10 orders.
This figure illustrates the dynamics of the BAT seller who posts a limit order at $v_i + d/6$. State $(i, j)$ implies the number of BAT orders on the ask and bid sides if the BAT seller add a regular limit order. BB and BS imply the arrival of BAT buy and sell orders, respectively. NB and NS are arrivals of non-algo buy and sell orders, respectively, while UJ and DJ are upward and downward jumps, respectively. For example, submitting a sell limit order to an empty LOB leads to state $(1, 0)$, and the expected cost for the limit order is $C(1, 0)$. If a BAT submits a limit order when a limit order already exists on the opposite side of the LOB, the state after submission is $(1, 1)$ and the cost is $C(1, 1)$. 
References


Jones, C. M. (2013). What do we know about high-frequency trading?


Appendix A

Appendix for Chapter 1

A.1 Proofs

A.1.1 Proof of Proposition 1

I first show that when $s/2 < \sigma$, $\pi(s/2, X)$ is strictly decreasing in $X$. Insert equation (1.1), (1.2), (1.3) and (1.4) to (1.5), we have:

\[
\pi(s/2, X) = \frac{\lambda_I}{\Sigma \lambda} \frac{s}{2} + \frac{\lambda_J}{\Sigma \lambda} \left[-\left(\sigma - \frac{s}{2}\right)\right] + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta)\left[\frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1}{2} \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} \left(\sigma - \frac{s}{2}\right)\right] + \frac{\lambda_U}{\Sigma \lambda} \left[1 - \phi(\delta)\right] \left[\frac{1}{2} \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} \left(\sigma - \frac{s}{2}\right)\right]
\]

\[
\left(\pi(s/2, X) - \phi(\delta)\right)\left[\frac{1}{2} \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} \left(\sigma - \frac{s}{2}\right)\right] + \frac{\lambda_U}{\Sigma \lambda} \left[1 - \phi(\delta)\right] \left[\frac{1}{2} \frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} \left(\sigma - \frac{s}{2}\right)\right]
\]

Denote $\frac{1}{X} = x$, $C = \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{s}{2}$ and $D = \frac{\lambda_J}{\lambda_I + \lambda_J} \left(\sigma - \frac{s}{2}\right)$, then:

\[
\pi(s/2, x) = \frac{\lambda_I}{\Sigma \lambda} \frac{s}{2} - \frac{\lambda_J}{\Sigma \lambda} \frac{s}{2} \left[\sigma - \frac{s}{2}\right] + \frac{\lambda_U}{\Sigma \lambda} x \phi(\delta)\left(-x C - D\right) + \frac{\lambda_U}{\Sigma \lambda} x \left[1 - \phi(\delta)\right]\left(-x C - D\right) + \frac{\lambda_U}{\Sigma \lambda} \left(1 - x\right) \phi(\delta)\left(-x C - D\right) + \frac{\lambda_U}{\Sigma \lambda} \left(1 - x\right) \left[1 - \phi(\delta)\right]\left(-x C - D\right)
\]

Now we take derivative with respect to $x$:

\[
\frac{d\pi(s/2, x)}{dx} = \frac{\lambda_I}{\Sigma \lambda} \frac{s}{2} + \frac{\lambda_J}{\Sigma \lambda} \frac{\phi(\delta)\left(-x C - D\right)}{x} + \frac{\lambda_U}{\Sigma \lambda} \left[1 - \phi(\delta)\right]\left(-x C - D\right) + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta)\left(C + D - 2x C\right) + \frac{\lambda_U}{\Sigma \lambda} \left[1 - \phi(\delta)\right] x
\]

\[
\left[\frac{1}{x} C + \frac{1}{2} D - x C\right] = \frac{\lambda_I}{\Sigma \lambda} \frac{s}{2} + \frac{\lambda_J}{\Sigma \lambda} \left[\phi(\delta) - \phi(\delta)\right]\frac{1}{2} B + \frac{\lambda_U}{\Sigma \lambda} C\frac{1}{2} + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta)\left(-x C - D\right) + \frac{\lambda_U}{\Sigma \lambda} \left[1 - \phi(\delta)\right] x
\]

\[
\frac{1}{2} D + \frac{\lambda_U}{\Sigma \lambda} C\phi(\delta)\left(1 - x\right) > 0
\]
since \( x = \frac{1}{X} \leq 1 \) and \( \delta < \varepsilon \) (under assumption that \( \delta + \zeta < \varepsilon \)). Therefore, the profit function \( \pi(\frac{\delta}{2}, \frac{\varepsilon}{2}) \) is strictly decreasing in \( X \).

Competition among HFTs will drive the equilibrium bid-ask spread small enough such that liquidity provision profit is close to zero. Because \( \pi(\frac{\delta}{2}, X) \) is strictly decreasing in \( X \), the equilibrium bid-ask spread \( s^* \) is determined when \( X = 1 \). This is the result in equation (1.6). Intuitively, if \( \pi(\frac{\delta}{2}, 1) \) is positive, A HFT will submit limit orders with this spread to one exchange. We first need to determine under which parameters, the equilibrium bid-ask spread is binding at one tick. In this case, since there is no price grid available inside the current bid and ask price, the undercutting HFT would not generate any effects and can not affect the HFT’s liquidity provision profit at one tick bid-ask spread. Specifically, the expected profit by submitting limit sell at \( v_o + d/2 \) and limit buy at \( v_o - d/2 \) is \( \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{d}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{d}{2}) \). When investor arrives at the market first, the HFT earns liquidity provision revenue \( d/2 \), which is the first component. If the risky asset’s common value jumps first, the HFT will lose \( \sigma - d/2 \) which is the second component. The non-negative requirement of this profit needs \( \frac{\lambda_J}{\lambda_I + \lambda_J} \leq \frac{d}{\sigma} \). Otherwise, the equilibrium bid-ask spread is larger than one tick and the expected profit for providing liquidity at exchange 1 with bid-ask spread \( s \) is defined in equation (1.5) when \( X = 1 \). So the equilibrium bid-ask spread would be the minimum available price in the price grids such that \( \pi(\frac{\delta}{2}, 1) \) is nonnegative. This proves the result in equation (1.6).

If equilibrium half spread \( s^*/2 \) is larger or equal to the risky asset’s common value jumping size \( \sigma \), there is no adverse selection cost for liquidity provision HFTs. So they will provide liquidity at all \( M \) exchanges. If \( s^*/2 < \sigma \), HFTs will compete to provide liquidity at the equilibrium bid and ask price at multiple exchanges until their profits from liquidity provision is negative. This prove the results in Proposition 1.

We need to verify that undercutting HFT sends her price improving order to one among those \( M^* \) exchanges is optimal. Suppose this undercutting HFT arrives at time \( t \) and she is a seller. Thus she is willing to submit a limit sell order at price \( v_o + s^*/2 - d \) to one exchange. If she sends her order to one among those \( M^* \) exchanges, her payoff is:

\[
\frac{1}{2} \left\{ \phi(e) \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1}{M^*} \left( \frac{s^*}{2} - \frac{d}{2} \right) - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s^*}{2} + d) \right] + \left[ 1 - \phi(e) \right] \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{\lambda_J}{\lambda_I + \lambda_J} \frac{s^*}{2} - \frac{d}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s^*}{2} + d) \right] \right\}
\]

\[(A.4)\]
If she submits her pricing improving order to one exchange among those \(M - M^*\) exchanges, her payoff would be:

\[
\frac{1}{2}\{\phi(\epsilon) - \frac{\lambda_I}{\lambda_I + \lambda_J}[-(\sigma - \frac{s^*}{2} + d)] + [1 - \phi(\epsilon)][\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{s^*}{2} - d) - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s^*}{2} + d)]\} \quad (A.5)
\]

Clearly the payoff in equation (A.4) is larger than (A.5) because undercutting HFT will enter the market only when \(s^*/2 > d\). The reason is because at time \([t, t + \epsilon]\) investors and exchanges does not know the existence of the undercutting HFT’s pricing improving order, thus investors will still send their orders to one of those \(M^*\) exchanges if they arrives before \(t + \epsilon\). In order to increase the probability of trading with uninformed investors, it is optimal for the undercutting HFT to send her order to one of those \(M^*\) exchanges too.

A.1.2 Proof of Corollary 1

(i) Take derivative of \(\pi(\frac{\epsilon}{2}, 1)\) with respect to \(\delta\), we have:

\[
\frac{d\pi(\frac{\epsilon}{2}, 1)}{d\delta} = -\frac{1}{2}\phi'(\delta)\frac{\lambda_U}{\lambda_I + \lambda_J}\frac{\lambda_I}{\sigma - \frac{s^*}{2}} < 0 \quad (A.6)
\]

where \(\phi'(\delta) = \frac{1}{\lambda_I + \lambda_J}e^{-\frac{1}{\lambda_I + \lambda_J}\delta} > 0\). Since \(\pi(\frac{\epsilon}{2}, 1)\) is strictly increasing in \(\frac{\epsilon}{2}\), equation (1.6) implies that \(s^*\) is weakly increasing in \(\delta\). Since \(\pi(\frac{\epsilon}{2}, 1)\) is independent of \(\epsilon\) so as the equilibrium bid-ask spread \(s^*\).

(ii) Take derivative of \(\pi(\frac{\epsilon}{2}, X)\) with respect to \(\epsilon\), we have:

\[
\frac{d\pi(\frac{\epsilon}{2}, X)}{d\epsilon} = \frac{1}{2}\phi'(\epsilon)\frac{\lambda_U}{\lambda_I + \lambda_J}X - \frac{1}{\lambda_I + \lambda_J}\frac{1}{X}\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{X}\frac{s^*}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s^*}{2}) \quad (A.7)
\]

where \(\phi'(\epsilon) = \frac{1}{\lambda_I + \lambda_J}e^{-\frac{1}{\lambda_I + \lambda_J}\epsilon} > 0\). We must have \(\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{X}\frac{s^*}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s^*}{2}) > 0\), otherwise \(\pi(\frac{\epsilon}{2}, X) < 0\) for all \(X \geq 1\). So \(\frac{d\pi(\frac{\epsilon}{2}, X)}{d\epsilon} > 0\). Equation (1.7) implies that \(M^*\) is weakly increasing in \(\epsilon\).
(iii) From equation (A.1), we have $\pi(\frac{s^*}{2}, X | \delta = \delta_F) - \pi(\frac{s^*}{2}, X | \delta = \delta_S) =$

\[
\frac{\lambda_U}{\Sigma \lambda} \frac{1}{X} [\phi(\delta_F) - \phi(\delta_S)]\left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \cdot \frac{1}{2} \frac{s^*}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s^*}{2}) \right] + \frac{\lambda_U}{\Sigma \lambda} \frac{1}{X} [\phi(\delta_F) - \phi(\delta_S)]\left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \cdot \frac{1}{2} \frac{s^*}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s^*}{2}) \right] > 0 \quad (A.8)
\]

According to equation (1.7), we have $M^*(\delta_F) \geq M^*(\delta_S)$.

A.1.3 Proof of Proposition 2

Suppose the undercutting HFT arrives at $t_U$ and denoting $t_1 = t_I - t_U$ and $t_2 = t_J - t_U$, then $Prob(t_I \geq t_J | t_I, t_J > t_U) = Prob(t_i \geq t_2 | t_i, t_2 > o) = Prob(t_1 \geq t_2)$ where $t_1 \sim Exp(\lambda_I)$, $t_2 \sim Exp(\lambda_J)$ and they are independent for a given $t_U$ because of the memoryless property of exponential distribution. Similarly, $Prob(t_I < t_J, t_I \leq t_U + \epsilon | t_I, t_J > t_U) = Prob(t_1 < t_2, t_1 \leq \epsilon)$ and $Prob(t_I < t_J, t_I > t_U + \epsilon | t_I, t_J > t_U) = Prob(t_1 < t_2, t_1 > \epsilon)$. It is easy to see that $Prob(t_1 \geq t_2) = \frac{\lambda_I}{\lambda_I + \lambda_J}$. Now we show that:

\[
Prob(t_1 < t_2, t_1 \leq \epsilon) = \int_0^{\epsilon} \int_t^\infty \lambda_I e^{-\lambda_I t_1} \lambda_J e^{-\lambda_J t_2} dt_2 dt_1 = \int_0^{\epsilon} \lambda_I e^{-\lambda_I t_1} \int_t^\infty \lambda_J e^{-\lambda_J t_2} dt_2 dt_1 = \int_0^{\epsilon} \lambda_I e^{-\lambda_I t_1} e^{-\lambda_J t_1} dt_1 = -\frac{\lambda_I}{\lambda_I + \lambda_J} e^{-(\lambda_I + \lambda_J) t_1} \bigg|_0^\epsilon = \frac{\lambda_I}{\lambda_I + \lambda_J} \left[ 1 - e^{-(\lambda_I + \lambda_J) \epsilon} \right] = \frac{\lambda_I}{\lambda_I + \lambda_J} \phi(\epsilon) \quad (A.9)
\]

Similarly, we can show that $Prob(t_1 < t_2, t_1 > \epsilon) = \frac{\lambda_I}{\lambda_I + \lambda_J} [1 - \phi(\epsilon)]$. Insert these results into equation (1.8), we would have:

\[
TC(Buy | \delta_i = \delta, \epsilon_{ij} = \epsilon) = \frac{\lambda_I + \lambda_J \frac{s^*}{2}}{\Sigma \lambda} + \frac{\lambda_U}{\Sigma \lambda} \phi(\epsilon) \left( \frac{1}{2} \frac{s^*}{2} + \frac{1}{2} \left[ \frac{M^* - 1}{2} \frac{s^*}{2} + \frac{1}{2} \left( \frac{s^*}{2} - d \right) \right] \right) + \frac{\lambda_U}{\Sigma \lambda} (1 - \phi(\epsilon)) \left( \frac{1}{2} \frac{s^*}{2} + \frac{1}{2} \left( \frac{s^*}{2} - d \right) \right) \quad (A.10)
\]

Denoting $A = \frac{1}{2} \frac{s^*}{2} + \frac{1}{2} \left[ \frac{M^* - 1}{2} \frac{s^*}{2} + \frac{1}{2} \left( \frac{s^*}{2} - d \right) \right]$ and $B = \frac{1}{2} \frac{s^*}{2} + \frac{1}{2} \left( \frac{s^*}{2} - d \right)$ we would have the result in Proposition 2 (i). Since $A \geq B$ for all $M^* \geq 1$ and $A$ is increasing in $M^*$, it is obvious that $W(Buy | \delta_i = \delta, \epsilon_{ij} = \epsilon)$ is decreasing in $\epsilon$ because $M^*$ is weakly increasing in $\epsilon$ according to Corollary 1. Also from Corollary 1, $M^*(\delta_F) \geq M^*(\delta_S)$. Thus $W(Buy | \delta_F = \delta_1, \epsilon_{ij} = \epsilon) \leq W(Buy | \delta_i = \delta_2, \epsilon_{ij} = \epsilon)$. 

105
\( \delta_S, \epsilon_{ij} = \epsilon \).

### A.1.4 Proof of Proposition 3

Starting from \( t = 0 \), three events may occur: investor arrival, the risky asset’s common value jumping or undercutting HFT’s arrival. In average it takes \( \frac{1}{\sum \lambda} \) units time for one of these events to occur. With probability \( \frac{\lambda_I}{\sum \lambda} \) the investor arrives first, in this case total transaction is one unit of the risky asset. With equal probability this transaction can occur at one among those \( M \) exchanges.\(^1\)

Thus the ex ante expected trading volume for each exchange is \( 1/M \). With probability \( \frac{\lambda_I}{\sum \lambda} \) the risky asset’s common value jumps first. No matter it is public or private signal, all the limit orders at one side of those \( M^* \) exchanges are taken either by the informed investors or snipers. So in this case, the ex ante expected trading volume for each exchange is \( M^*/M \). With probability \( \frac{\lambda_U}{\sum \lambda} \) undercutting HFT arrives first, the expected trading volume in this case depends on whether the investor arrives before or after the risky asset’s common value jumps.

Specifically, after the undercutting HFT’s arrival only two events might occur: investor arrival or the risky asset’s common value jumps. In average it takes \( \frac{1}{\lambda_I + \lambda_J} \) units of time for one of these two events to occur. With probability \( \frac{\lambda_I}{\lambda_I + \lambda_J} \) an investor arrives first after the undercutting HFT’s arrival. In this case, total trading volume is one unit. Each exchange has expected trading volume \( 1/M \).

If the risky asset’s common value jumps first, trading volume depends on whether liquidity provision HFTs have canceled their being undercut limit orders or not. Denoting \( t_U, t_I \) and \( t_J \) as the first arriving time of an undercutting HFT, investor and the risky asset’s common value jumping, similar to the calculation in equation (A.9) With \( \text{Prob}(t_J < t_I, t_J \leq t_U + \delta | t_I, t_J > t_U) = \frac{\lambda_I}{\lambda_I + \lambda_J} \phi(\delta) \) liquidity provision HFTs have not canceled their being undercut limit orders when the risky asset’s common value jumps, where \( \phi(\delta) = 1 - e^{-(\lambda_I + \lambda_J)\delta} \). The total trading volume in this case depends on whether the undercutting HFT’s price improving limit order is at the same side with the asset’s value jumping or not. For example, if the undercutting HFT is a seller and the

\(^1\) Alternative, at the initial quoting stage a particular exchange has probability \( \frac{M^*}{M} \) to be chosen by liquidity provision HFT to submit their limit orders. When the investor arrives, each exchange among those \( M^* \) exchanges has probability \( \frac{1}{M^*} \) to be chosen by the investor. So ex ante each exchange has probability \( \frac{M^*}{M} \times \frac{1}{M^*} = \frac{1}{M} \) to facilitate the investor’s trade.
risky asset’s common value jumps up by \( \sigma \), in this case the undercutting HFT’s sell limit order would be sniped too. Thus the total trading volume would be \( M^* + 1 \). If the asset’s value jumps down by \(-\sigma\), the total volume would be \( M^* \) since the undercutting HFT’s sell limit order is not stale. In this case, the ex ante expected trading volume for each exchange is \([\frac{1}{2}(M^* + 1) + \frac{1}{2}M^*/M]\).

We also need to calculate the expected units of time it takes for the risky asset’s common value jumps conditional on it happens before investor arrival and liquidity provision HFTs have not canceled their orders. This is denoted as \( E(t_f - t_U|t_U < t_f, t_f \leq t_U + \delta) \). As in the proof of Proposition 2, we define \( t_1 = t_f - t_U \) and \( t_2 = t_f - t_U \), thus \( t_1 \sim Exp(\lambda_I) \) and \( t_2 \sim Exp(\lambda_J) \). In order to calculate \( E(t_f - t_U|t_U < t_f, t_f \leq t_U + \delta) \) we first calculate \( E(t_f - t_U|t_U < t_f, t_f \leq t_U + \delta) \). In this way we can directly use the result in equation (A.9). Specifically,

\[
E(t_f - t_U|t_U < t_f, t_f \leq t_U + \delta) = E(t_1|t_f < t_f, t_f \leq \delta) = \\
\int_0^\delta \int_0^\infty \frac{1}{\text{Prob}(t_1 < t_f, t_f \leq \delta)} t_f \lambda_I e^{-\lambda_I t_1} e^{-\lambda_f t_2} dt_1 dt_2 = \int_0^\delta \int_0^\infty \frac{1}{\lambda_I[1 - e^{-(\lambda_I + \lambda_J)\delta}]} t_f \lambda_I e^{-\lambda_I t_1} \lambda_J e^{-\lambda_J t_2} dt_1 dt_2
\]

\[
= \frac{\lambda_I + \lambda_J}{\lambda_I[1 - e^{-(\lambda_I + \lambda_J)\delta}]} \int_0^\delta t_f \lambda_I e^{-\lambda_I t_1} \int_0^\infty e^{-\lambda_J t_2} dt_2 dt_1 = \frac{\lambda_I + \lambda_J}{\lambda_I[1 - e^{-(\lambda_I + \lambda_J)\delta}]} \int_0^\delta t_f \lambda_I e^{-\lambda_I t_1} e^{-\lambda_J t_2} dt_1 dt_2
\]

\[
= \frac{\lambda_I + \lambda_J}{\lambda_I[1 - e^{-(\lambda_I + \lambda_J)\delta}]} [-\lambda_I \int_0^\delta t_f e^{-(\lambda_I + \lambda_J) t_f} dt_f + \int_0^\delta e^{-(\lambda_I + \lambda_J) t_f} dt_f] = \frac{\lambda_I + \lambda_J}{\lambda_I[1 - e^{-(\lambda_I + \lambda_J)\delta}]} \times
\]

\[
\left[-\frac{\lambda_I}{\lambda_I + \lambda_J} \delta e^{-(\lambda_I + \lambda_J)\delta} - \frac{\lambda_I}{(\lambda_I + \lambda_J)^2} e^{-(\lambda_I + \lambda_J)\delta} + \frac{\lambda_I}{\lambda_I + \lambda_J}\right] = \frac{1}{\lambda_I + \lambda_J} \left[-\delta e^{-(\lambda_I + \lambda_J)\delta} - \frac{\lambda_I}{\lambda_I + \lambda_J} \delta e^{-(\lambda_I + \lambda_J)\delta} + 1 - \phi(\delta) \delta \right] (A.11)
\]

Therefore,

\[
E(t_f - t_U|t_U < t_f, t_f \leq t_U + \delta) = E(t_2|t_2 < t_f, t_f \leq \delta) = \frac{1}{\lambda_I + \lambda_J} - \frac{1 - \phi(\delta)}{\phi(\delta)} \delta (A.12)
\]

This is because of the symmetric position of \( t_1 \) and \( t_2 \) in the calculation of \( E(t_1|t_f < t_f, t_f \leq \delta) \) or \( E(t_2|t_f < t_f, t_f \leq \delta) \) and the final result in equation (A.11) does not depend on the order of \( \lambda_I \) and \( \lambda_J \).

With \( \text{Prob}(t_f < t_f, t_f + \delta < t_f \leq t_f + \epsilon|t_f, t_f > t_U) \) the risky asset’s common value jumps after \( \delta \) units of time but within \( \epsilon \) units of time after the undercutting HFT’s arrival, since the liquidity provision HFT will cancel her being undercut limit order at \( t_U + \delta \) if she is at the same exchange.
with the undercutting HFT and will cancel her being undercut limit orders at \( t_U + \epsilon \) if they are not at the same exchange with the undercutting HFT, in this case there are always total \( M^* \) limit orders at the both side of the limit order books. Therefore, \( M^* \) limit orders would be taken by snipers or informed traders. In this case each exchange has ex ante expected trading volume \( M^* \).

We also need to calculate the probability of this case similarly to (A.9):

\[
\begin{align*}
Prob(t_j < t_i, t_U + \delta < t_j \leq t_U + \epsilon | t_i, t_j > t_U) &= Prob(t_2 < t_1, \delta < t_2 \leq \epsilon) = \int_\delta^\epsilon \int_t^\infty \lambda_i e^{-\lambda_i t_i} \lambda_j e^{-\lambda_j t_j} dt_i dt_2 = \int_\delta^\epsilon \frac{\lambda_i}{\lambda_i + \lambda_j} \int_\delta^\epsilon (\lambda_i + \lambda_j) e^{-\lambda_i t_i} dt_2 \times \frac{\lambda_j}{\lambda_i + \lambda_j} (\phi(\epsilon) - \phi(\delta)) \quad (A.13)
\end{align*}
\]

where \( \phi(\epsilon) = 1 - e^{-(\lambda_i + \lambda_j)\epsilon} \) and \( \phi(\delta) = 1 - e^{-(\lambda_i + \lambda_j)\delta} \). In this case the expected units of time it takes for this event to occur after the undercutting HFT’s arrival is:

\[
\begin{align*}
E(t_j - t_U | t_j < t_i, t_U + \delta < t_j \leq t_U + \epsilon) &= E(t_2 | t_2 < t_1, \delta < t_2 \leq \epsilon) = \int_\delta^\epsilon \int_t^\infty \frac{t_2 \lambda_i e^{-\lambda_i t_i} \lambda_j e^{-\lambda_j t_j} dt_i dt_2}{\lambda_i + \lambda_j} \times \frac{\lambda_i + \lambda_j}{\lambda_i [\phi(\epsilon) - \phi(\delta)]} \times \frac{\lambda_j}{\lambda_i + \lambda_j} [\phi(\epsilon) - \phi(\delta)] - \frac{\epsilon \lambda_j}{\lambda_i + \lambda_j} (1 - \phi(\epsilon)) + \frac{\lambda_j}{\lambda_i + \lambda_j} (\phi(\epsilon) - \phi(\delta)) = \frac{1}{\lambda_i + \lambda_j} + \epsilon - \frac{\epsilon - \delta}{\phi(\epsilon - \delta)} \quad (A.14)
\end{align*}
\]

since \( \phi(\epsilon) - \phi(\delta) = [1 - \phi(\delta)] \phi(\epsilon - \delta) \).

Similarly, with \( \text{Prob}(t_j < t_i, t_j > t_U + \epsilon | t_i, t_j > t_U) = \frac{\lambda_j}{\lambda_i + \lambda_j} [1 - \phi(\epsilon)] \) (implied from (A.9)) the risky asset’s common value jumps before the investor’s arrival but after all liquidity provision HFTs canceled their being undercut limit orders. In this case the total trading volume also depends on whether the undercutting HFT’s price improving limit order is at the same side with the asset’s value jumping or not. If it is, trading volume would be just one unit of the risky asset, because all original HFTs have canceled their being undercut limit orders. Thus, only under-cutting HFT’s limit order is subject to the sniping risk. If the under-cutting HFT’s price improving limit order is at the opposite side of the risky asset’s value jumping, then trading volume would be \( M^* \) units.
Therefore, in this case the expected trading volume for each exchange is \( \left( \frac{1}{2} M^* + \frac{1}{2} \right) / M \). We also need to calculate \( E(t_j - t_U | t_U < t_j < t_i, t_j > t_U + \epsilon) \). Still, we first calculate:

\[
E(t_i - t_U | t_U < t_i, t_i > t_U + \epsilon) = E(t_i | t_1 < t_2, t_1 > \epsilon) =
\]

\[
\int_{t_1}^{t_2} \int_{t_1}^{t_2} \frac{1}{\text{Prob}(t_i < t_2, t_1 > \epsilon)} t_i \lambda_i e^{-\lambda_i t_i} \lambda_j e^{-\lambda_j t_2} dt_2 dt_1 = \int_{t_1}^{t_2} \int_{t_1}^{t_2} \frac{t_i \lambda_i e^{-\lambda_i t_i} \lambda_j e^{-\lambda_j t_2}}{\lambda_i + \lambda_j} e^{-(\lambda_i + \lambda_j) \epsilon} dt_2 dt_1
\]

\[
= \frac{\lambda_i + \lambda_j}{\lambda_i e^{-(\lambda_i + \lambda_j) \epsilon}} \int_{t_1}^{t_2} t_i \lambda_i e^{-\lambda_i t_i} \lambda_j e^{-\lambda_j t_2} dt_2 dt_1\]

\[
= \frac{\lambda_i + \lambda_j}{\lambda_i e^{-(\lambda_i + \lambda_j) \epsilon}} \left[ -\frac{\lambda_i}{\lambda_i + \lambda_j} \int_{t_1}^{t_2} e^{-(\lambda_i + \lambda_j) t_1} \right] + \int_{t_1}^{t_2} e^{-(\lambda_i + \lambda_j) t_1} t_i \lambda_i e^{-\lambda_i t_i} \lambda_j e^{-\lambda_j t_1} dt_1\]

\[
= \frac{1}{\lambda_i + \lambda_j} e^{-(\lambda_i + \lambda_j) \epsilon} \left[ -\frac{\lambda_i}{\lambda_i + \lambda_j} \right] + \frac{1}{\lambda_i + \lambda_j} e^{-(\lambda_i + \lambda_j) \epsilon} \left[ 1 \right] = \frac{1}{\lambda_i + \lambda_j} + \epsilon \quad (A.15)
\]

Similarly, we also have:

\[
E(t_j - t_U | t_U < t_j, t_j > t_U + \epsilon) = E(t_2 | t_2 < t_1, t_2 > \epsilon) = \frac{1}{\lambda_i + \lambda_j} + \epsilon \quad (A.16)
\]

Denoting \( Q \) as each exchange’s expected per unit time trading volume, we must have:

\[
\left( \frac{1}{\Sigma \lambda} + \frac{1}{\lambda_i + \lambda_j} + \epsilon \right) Q = \frac{\lambda_i M^*}{\Sigma \lambda M} \left( \frac{1 + \frac{1}{\lambda_i + \lambda_j} + \epsilon}{M} \right) + \frac{\lambda_j M^*}{\Sigma \lambda M} \left( \frac{1 + \frac{1}{\lambda_i + \lambda_j} + \epsilon}{M} \right) + \frac{\lambda_i}{\Sigma \lambda M} \left( \frac{1 + \frac{1}{\lambda_i + \lambda_j} + \epsilon}{M} \right) + \frac{\lambda_j}{\Sigma \lambda M} \left( \frac{1 + \frac{1}{\lambda_i + \lambda_j} + \epsilon}{M} \right)
\]

\[
+ \frac{\lambda_i}{\Sigma \lambda M} \left[ \frac{1}{\lambda_i + \lambda_j} \phi(\delta) \right] + \frac{\lambda_j}{\Sigma \lambda M} \left[ \frac{1}{\lambda_i + \lambda_j} \phi(\delta) \right] + \frac{\lambda_i}{\Sigma \lambda M} \left[ \frac{1 - \phi(\delta)}{\phi(\delta)} \right] + \frac{\lambda_j}{\Sigma \lambda M} \left[ \frac{1 - \phi(\delta)}{\phi(\delta)} \right]
\]

\[
= \frac{1}{\Sigma \lambda} M^* \left[ \frac{1}{\lambda_i + \lambda_j} + \epsilon \right] + \frac{1}{\Sigma \lambda} M^* \left[ \frac{1}{\lambda_i + \lambda_j} + \epsilon \right] + \frac{1}{\Sigma \lambda} M^* \left[ \frac{1}{\lambda_i + \lambda_j} + \epsilon \right] + \frac{1}{\Sigma \lambda} M^* \left[ \frac{1}{\lambda_i + \lambda_j} + \epsilon \right]
\]

\[
+ \frac{\lambda_i}{\Sigma \lambda M} \left[ \frac{1}{\lambda_i + \lambda_j} \phi(\delta) \right] + \frac{\lambda_j}{\Sigma \lambda M} \left[ \frac{1}{\lambda_i + \lambda_j} \phi(\delta) \right] + \frac{\lambda_i}{\Sigma \lambda M} \left[ \frac{1 - \phi(\delta)}{\phi(\delta)} \right] + \frac{\lambda_j}{\Sigma \lambda M} \left[ \frac{1 - \phi(\delta)}{\phi(\delta)} \right]
\]

\[
+ \frac{\lambda_i}{\Sigma \lambda M} \left[ \frac{1 - \phi(\delta)}{\phi(\delta)} \right] + \frac{\lambda_j}{\Sigma \lambda M} \left[ \frac{1 - \phi(\delta)}{\phi(\delta)} \right]
\]

which implies that:

\[
Q^* = \lambda_i \frac{M^*}{M} + \lambda_j \frac{M^*}{M} + \frac{\lambda_i \lambda_j}{\Sigma \lambda} \phi(\delta) \frac{1}{M} - \frac{1 - \phi(\delta)}{\phi(\delta)} \frac{\lambda_i \lambda_j}{2 \Sigma \lambda} \frac{M^* - 1}{M} \quad (A.18)
\]

Now we explain equation (A.17). The left-hand side is a particular exchange’s ex ante expected trading volume in \( \frac{1}{\Sigma \lambda} + \frac{1}{\lambda_i + \lambda_j} + \epsilon \) units of time because \( Q \) is the per unit time trading volume. Starting from \( t = 0 \), with probability \( \frac{\lambda_i}{\Sigma \lambda} \) an investor arrives first. In this case it takes the particular
exchange \( \frac{1}{\lambda} \) units time to have \( \frac{1}{M} \) trading volume. Then the game will move to a new game \( G' \). The expected trading volume for the particular exchange in the remaining \( \frac{1}{\lambda_i + \lambda_f} + \epsilon \) units of time would be \((\frac{1}{\lambda_i + \lambda_f} + \epsilon)Q \). This explains the first term in the right-hand side of equation (A.17). Other terms in the right-hand side of equation (A.17) can be explained in a similar way. The result in (ii) is directly implied from equation (A.18).

Now we prove the result in (iii). Note that according to Corollary 1 (iii) when \( s^*(\delta_F) = s^*(\delta_S) \), \( M^*(\delta_F) \geq M^*(\delta_S) \). We have \( Q^*(\delta_F) - Q^*(\delta_S) = \)

\[
\frac{\lambda_j M^*(\delta_F) - M^*(\delta_S)}{M} + \frac{\lambda_U\lambda_I}{2\Sigma\lambda} [\phi(\delta_F) - \phi(\delta_S)] \frac{1}{M} - \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U\lambda_I}{\Sigma\lambda} M^*(\delta_F) - \frac{M^*(\delta_S)}{M} \tag{A.19}
\]

Thus, (A.19) < 0 if \( M^*(\delta_F) = M^*(\delta_S) \) since \( \phi(\delta_F) < \phi(\delta_S) \). If \( M^*(\delta_F) > M^*(\delta_S) \) then \( M^*(\delta_F) - M^*(\delta_S) \geq 1 \) because \( M^*(\delta_F) \) and \( M^*(\delta_S) \) are integers. Therefore, (A.19) is greater or equal to:

\[
\frac{\lambda_j}{M} + \frac{\lambda_U\lambda_I}{2\Sigma\lambda} \frac{\phi(\delta_F) - \phi(\delta_S)}{M} - \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U\lambda_I}{M\Sigma\lambda} = \frac{\lambda_I}{M} \left[ 1 + \frac{\lambda_U}{2\Sigma} \phi(\delta_F) - \frac{\lambda_U}{2\Sigma} (1 - \phi(\epsilon) + \phi(\delta_S)) \right] \tag{A.20}
\]

which is positive because \( \delta_S < \epsilon \).

**A.1.5 Proof of Lemma 1**

We will prove Lemma 1 under the assumption that uninformed investors will randomly choose one among those exchanges with best price quotes to trade with equal probability. In the remaining analysis of Section 1.3.2 some uninformed investors are smart and they will submit their market orders to fast exchanges with best price quotes. This will make fast exchanges to be more attractive for undercutting HFTs. Thus, Lemma 1 still holds in the remaining analysis of Section 1.3.2 where the portion of smart investors is positive (current proof is under the assumption that all investors are non-smart).

Suppose the undercutting HFT is a seller and denote the current bid-ask spread as \( s^* \) and \( M^* \) exchanges have these best price limit orders. Thus, the undercutting HFT is willing to sell at \( v_o + s^*/2 - d \). Suppose among those \( M^* \) exchanges there are \( K^* \) fast exchanges and \( M^* - K^* \) slow exchanges, where \( 1 \leq K^* \leq M^* - 1 \). If the undercutting HFT submits her order to a fast exchange
named exchange 1 which is in those \( M^* \) exchanges, her payoff is:

\[
\frac{1}{2} \{ \phi(\delta_F + \epsilon - \delta_S) \left[ \frac{\lambda_I}{\lambda_I + \lambda_f} M^* \left( \frac{s^*}{2} - d \right) - \frac{\lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s^*}{2} + d) \right] + [\phi(\epsilon) - \phi(\delta_F + \epsilon - \delta_S)] \left[ \frac{\lambda_I}{\lambda_I + \lambda_f} M^* \left( \frac{s^* - \delta_F + \epsilon - \delta_S}{2} - d \right) \right] \times \\
\frac{\lambda_I}{\lambda_I + \lambda_f} \left( \frac{M^* - K^* + 1}{M^*} \right) \times \\
\left[ \frac{\lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s^*}{2} + d) \right] + [1 - \phi(\epsilon)] \left[ \frac{\lambda_I}{\lambda_I + \lambda_f} \left( \frac{s^* - \delta_F + \epsilon - \delta_S}{2} - d \right) - \frac{\lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s^*}{2} + d) \right] \} \quad (A.21)
\]

There is \( \frac{1}{2} \) in the payoff function because the undercutting HFT only provide liquidity at the sell side of the limit order book. Suppose the undercutting HFT arrives at the market at time \( t \), then liquidity provision HFTs on those \( M^* - K^* \) slow exchanges will cancel their limit sell orders at \( t + \delta_F + \epsilon - \delta_S \) (see Figure 1.2). Liquidity provision HFTs on other \( K^* - 1 \) fast exchanges will cancel their limit sell orders at \( t + \delta_F + \epsilon - \delta_S = t + \epsilon \). Thus, if an uninformed investor arrives within \( t \) to \( t + \delta_F + \epsilon - \delta_S \), trade-through is possible on all remaining \( M^* - 1 \) exchanges. In this case, the undercutting HFT only has \( \frac{1}{M^*} \) probability to trade with the uninformed investor. This explains the first term in (A.21). If the uninformed investor arrives within \( t + \delta_F + \epsilon - \delta_S \) to \( t + \epsilon \), trade-through is only possible on the remaining \( K^* - 1 \) fast exchanges because all slow exchanges will reroute the uninformed investor’s order to exchange 1 (they know exchange 1 has better price). This is the second term in (A.21). Finally, if the uninformed investor arrives after \( t + \epsilon \), no trade-through is possible. All exchanges will reroute uninformed investor’s order to exchange 1. This explains the last term in (A.21).

If the undercutting HFT submits her order to one slow exchange in \( M^* \), liquidity provision HFTs on other slow exchanges will cancel their limit sell orders at time \( t + \delta_S + \epsilon - \delta_S = t + \epsilon \) while liquidity provision HFTs on fast exchanges will cancel their limit sell order at \( t + \delta_S + \epsilon - \delta_F \). The undercutting HFT’s payoff would be:

\[
\frac{1}{2} \{ \phi(\epsilon) \left[ \frac{\lambda_I}{\lambda_I + \lambda_f} M^* \left( \frac{s^*}{2} - d \right) - \frac{\lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s^*}{2} + d) \right] + [\phi(\epsilon) - \phi(\delta_F + \epsilon - \delta_S)] \left[ \frac{\lambda_I}{\lambda_I + \lambda_f} M^* \left( \frac{s^* - \delta_F + \epsilon - \delta_S}{2} - d \right) \right] \times \\
\frac{\lambda_I}{\lambda_I + \lambda_f} \left( \frac{M^* - K^* + 1}{M^*} \right) \times \\
\left[ \frac{\lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s^*}{2} + d) \right] + [1 - \phi(\epsilon)] \left[ \frac{\lambda_I}{\lambda_I + \lambda_f} \left( \frac{s^* - \delta_F + \epsilon - \delta_S}{2} - d \right) - \frac{\lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s^*}{2} + d) \right] \} \quad (A.22)
\]

(A.21) - (A.22) =

\[
\frac{1}{2} \{ [\phi(\epsilon) - \phi(\delta_F + \epsilon - \delta_S)] \left[ \frac{M - K}{M} \right] + [\phi(\delta_S + \epsilon - \delta_F) - \phi(\epsilon)] \left[ \frac{K}{M} \right] \} \frac{\lambda_I}{\lambda_I + \lambda_f} \left( \frac{s^*}{2} - d \right) > 0 \quad (A.23)
\]
(A.23) is positive because \( \delta_F < \delta_S \) and \( s^*/2 > d \) (Assumption 2). (A.23) is also quite intuitive. If undercutting HFT submits her order to fast exchange not slow exchange, the potential trade-through time window on slow exchanges will be reduced from \( \epsilon \) to \( \delta_F + \epsilon - \delta_S \). Similarly, the potential trade-through time window on fast exchanges will be reduced from \( \delta_S + \epsilon - \delta_F \) to \( \epsilon \). This increases undercutting HFT’s liquidity provision revenue.

When \( M^* < M \), the undercutting HFT can also submit her order to the exchange which does not have the current best price quotes (one among the remaining \( M - M^* \) exchanges). If it is a fast exchange, the undercutting HFT’s payoff is:

\[
\frac{1}{2}\{\phi(\delta_F + \epsilon - \delta_S)\left[-\frac{\lambda_f}{\lambda_f + \lambda_j}(\sigma - \frac{s^*}{2} + d)\right] + [\phi(\epsilon) - \phi(\delta_F + \epsilon - \delta_S)]\left[-\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{M^* - K^*}{M^*})(\frac{s^*}{2} - d)\right]
\]

\[
- \frac{\lambda_f}{\lambda_f + \lambda_j}(\sigma - \frac{s^*}{2} + d)\} + [1 - \phi(\epsilon)]\left[-\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{s^*}{2} - d)\right] - \frac{\lambda_f}{\lambda_f + \lambda_j}(\sigma - \frac{s^*}{2} + d)\} \quad (A.24)
\]

If it is a slow exchange, the undercutting HFT’s payoff is:

\[
\frac{1}{2}\{\phi(\epsilon)[\frac{\lambda_f}{\lambda_f + \lambda_j}(\sigma - \frac{s^*}{2} + d)] + [\phi(\delta_S + \epsilon - \delta_F) - \phi(\epsilon)]\left[-\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{M^* - K^*}{M^*})(\frac{s^*}{2} - d)\right] - \frac{\lambda_f}{\lambda_f + \lambda_j}\times
\]

\[
(\sigma - \frac{s^*}{2} + d)\} + [1 - \phi(\delta_S + \epsilon - \delta_F)]\left[-\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{s^*}{2} - d)\right] - \frac{\lambda_f}{\lambda_f + \lambda_j}(\sigma - \frac{s^*}{2} + d)\} \quad (A.25)
\]

(A.24) and (A.25) are constructed in the same way as in (A.21) and (A.22). Either within \( \delta_F + \epsilon - \delta_S \) or \( \epsilon \) units of time after the undercutting HFT’s arrival, trade-through is possible on all \( M^* \) exchanges. Because investors only send their market orders to exchanges with current best bid and ask prices, so the undercutting HFT has no chance to trade with an uninformed investor during this trade-through time window. The reason she still has adverse selection cost is because snipers or informed traders will send their liquidity taking orders to all exchanges, which can maximize their profits because hidden orders or due to latency some limit orders can not be seen from the limit order books but are available to trade. It is obvious that (A.21) > (A.24) and (A.22) > (A.25). Therefore, it is optimal for the undercutting HFT to send her price improving limit order to a fast exchange which has the current best price quotes.
A.1.6  Proof of Proposition 4

Denote $X = X_F + X_S$ and note that $\pi_S(\frac{z}{2}, X_F, X_S)$ only depends on $X$. Thus, we can define a new function $\pi_S(\frac{z}{2}, X) = \pi_S(\frac{z}{2}, X_F, X_S)$.

I will first show that when $0 < s/2 < \sigma$ and $\gamma$ is large enough, for any $2 \leq X \leq M$ if $\pi_S(\frac{z}{2}, X) \geq 0$, then $\pi_F(\frac{z}{2}, X_F, X - X_F) \geq \pi_S(\frac{z}{2}, X)$ for any $1 \leq X_F \leq X$. Denote $E = \frac{\lambda J}{\lambda I + \lambda J} X \geq 0$ and $F = \frac{\lambda J}{\lambda I + \lambda J} (\sigma - \frac{r}{2}) > 0$, where $X = X_F + X_S$. From equation (1.16) we have:

$$
\pi_F(\frac{s}{2}, X_F, X - X_F) = \frac{\lambda_I + \lambda_J}{\Sigma \lambda X_F} [1 - \phi(\delta_F)] + \frac{1}{\Sigma \lambda X_F}[\frac{1}{\lambda I + \lambda J} F + \frac{\lambda U}{\Sigma \lambda X_F} \phi(\delta_F)\frac{1}{\lambda I + \lambda J} X_F 2 + \frac{1}{2}]
$$

$$(A.26)
$$

From equation (1.17), we have:

$$
\pi_S(\frac{s}{2}, X) = \pi_S(\frac{s}{2}, X_F, X - X_F) = \frac{\lambda_I + \lambda_J}{\Sigma \lambda} - \frac{1}{\Sigma \lambda X_F} F + \frac{1}{\Sigma \lambda X_F} \phi(\delta_F)\frac{1}{\lambda I + \lambda J} X_F 2 + \frac{1}{2}]
$$

$$(A.27)
$$

Thus, $\pi_S(\frac{s}{2}, X) \geq 0 \iff (1 - \gamma)E - F \geq 0$ (A.26) minus (A.27) generates:

$$
\pi_F(\frac{s}{2}, X_F, X - X_F) - \pi_S(\frac{s}{2}, X) = \frac{\lambda_I + \lambda_J}{\Sigma \lambda X_F} [\phi(\delta_F + \epsilon - \delta_S) - \frac{\lambda U}{\Sigma \lambda X_F} F + \frac{\lambda U}{\Sigma \lambda X_F} \phi(\delta_F + \epsilon - \delta_S)](1 - \gamma)E - F + \frac{1}{\Sigma \lambda X_F} [\frac{1}{\lambda I + \lambda J} X_F 2 + \frac{1}{2}]
$$

$$(A.28)
$$

From assumption (3), $\delta_F + \epsilon - \delta_S > \delta_F$. Thus, the second term in (A.28) is positive. Since $(1 - \gamma)E - F \geq 0$, (A.28) is non-negative if:

$$
\frac{\lambda_I + \lambda_J}{\Sigma \lambda X_F} \phi(\delta_F + \epsilon - \delta_S) - \frac{\lambda U}{\Sigma \lambda X_F} F + \frac{\lambda U}{\Sigma \lambda X_F} \phi(\delta_F + \epsilon - \delta_S)\frac{1}{\lambda I + \lambda J} X_F 2 = \frac{1}{4} \frac{\lambda I}{\lambda I + \lambda J} X_F 2 + \frac{1}{\Sigma \lambda X_F} \phi(\delta_F + \epsilon - \delta_S)\frac{1}{\lambda I + \lambda J} X_F 2
$$

$$(A.29)
$$

The reason I define this new function is because $\pi_S(\frac{s}{2}, X_F, 0)$ is not well defined. $\pi_S(\frac{s}{2}, X)$ can avoid this problem.
Note that we replaced $E = \frac{\lambda_I}{\lambda_I + \lambda_J} \cdot \frac{1}{X^2}$ in the term of (A.28). Thus, (A.28) would be non-negative if:

$$2(\lambda_I + \lambda_J) \frac{Y}{X_F} + \lambda_U \frac{X_F - 1}{X_F} \phi'(\epsilon) \frac{Y}{X_F} + \lambda_U \frac{Y}{X_F} \geq \frac{\lambda_U \phi(\delta_F + \epsilon - \delta_S)}{X} (1 - \frac{Y}{X}) \quad (A.30)$$

Since the left-hand side of (A.30) is decreasing in $X_F$ and $\phi(\delta_F + \epsilon - \delta_S) < \phi(\epsilon)$, thus for a given $X$ if:

$$2(\lambda_I + \lambda_J) \frac{Y}{X} + \lambda_U \frac{X - 1}{X} \phi'(\epsilon) \frac{Y}{X} + \lambda_U \frac{Y}{X} \geq \frac{\lambda_U \phi(\epsilon)}{X} (1 - \frac{Y}{X}) \quad (A.31)$$

(A.30) will holds. Since the left-hand side of (A.31) is increasing in $X$, (A.31) would hold for all $2 \leq X \leq M$ if $2(\lambda_I + \lambda_J) \frac{Y}{X} + \lambda_U \frac{X - 1}{X} \phi'(\epsilon) \frac{Y}{X} + \lambda_U \frac{Y}{X} \geq \lambda_U \phi(\epsilon)(1 - \frac{Y}{X})$, which is equivalent as:

$$\frac{0.5\lambda_U \phi(\epsilon)}{\lambda_I + \lambda_J + 0.5\lambda_U + 0.75\lambda_U \phi(\epsilon)} \geq \frac{Y}{X} \quad (A.32)$$

This implies that if $Y$ satisfies equation (A.32), $\pi_F(\frac{S}{2}, X_F, X - X_F) \geq \pi_S(\frac{S}{2}, X)$ for all $2 \leq X \leq M$ and all $1 \leq X_F \leq X$ if $\pi_S(\frac{S}{2}, X) > 0$. This simply means that if HFT’s provide liquidity on both fast and slow exchanges, liquidity provision profit is larger on fast exchange than on slow exchange.

If HFTs provide liquidity on only one fast exchange and other $X - 1$ slow exchanges, we also need to check whether the liquidity provision HFT on fast exchange has incentive to switch to a slow exchanges when $X < M$. This is because if she switches to a slow exchange her liquidity provision profit is $\pi(\frac{S}{2}, X|\delta = \delta_S)$, which is the profit function (1.5) (homogeneous order processing speed) evaluated at order processing speed $\delta_S$. Because if she switches, we would have HFT’s provide liquidity on $X$ slow exchanges with order processing speed $\delta_S$. Note that $\pi(\frac{S}{2}, X|\delta = \delta_S)$ is not the same as $\pi_S(\frac{S}{2}, X)$. The later is the liquidity provision profit on a slow exchange if HFTs provide liquidity on $X$ exchanges including some fast exchanges. From equation (1.5) we have:

$$\pi(\frac{S}{2}, X|\delta = \delta_S) = \frac{\lambda_I + \lambda_J}{\Sigma \lambda} E - \frac{\lambda_I + \lambda_J}{\Sigma \lambda} F + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_S)(\frac{1}{2} E - F) + \frac{\lambda_U}{\Sigma \lambda} \frac{X}{X} \frac{1}{2} \phi'(\epsilon)(E - F) \quad (A.33)$$
Evaluating (A.26) at $X_F = 1$ and minus (A.33) generates:

$$
\pi_F\left(\frac{s}{2}, 1, X - 1\right) - \pi(\frac{s}{2}, X|\delta = \delta_S) = \frac{\lambda_I Y s}{\Sigma \lambda} - Y \frac{\lambda_I + \lambda_J}{\Sigma \lambda} E + \frac{\lambda_U}{\Sigma \lambda} \left(\frac{\lambda_I Y s}{\Sigma \lambda} \frac{1}{2} - \frac{Y}{2} \right) \lambda_I + \lambda_J + \lambda_U \frac{X - 1}{\Sigma \lambda} \delta_S - \frac{\lambda_U X}{\Sigma \lambda} \frac{Y}{2} + \frac{\lambda_U Y}{\Sigma \lambda} \left[\frac{1}{X} \phi(\delta_S) + \frac{X - 1}{X} \phi(\delta) - \phi(\delta_F)\right] (A.34)
$$

The last term in (A.34) is positive since $\epsilon > \delta_S > \delta_F$. Replacing $E$ by $\frac{\lambda_I Y s}{\Sigma \lambda} \frac{1}{2}$, we have:

$$
\frac{\lambda_I}{\lambda_I + \lambda_J} \frac{1}{X} s - \frac{\lambda_I + \lambda_J}{\Sigma \lambda} Y s + \frac{\lambda_U}{\Sigma \lambda} \left(\frac{\lambda_I Y s}{\Sigma \lambda} + \frac{Y}{2} E\right) - \frac{\lambda_U X - 1}{\Sigma \lambda} \phi(\delta_S) - \frac{\lambda_U X}{\Sigma \lambda} \phi(\delta) - \frac{X - 1}{\Sigma \lambda} \phi(\delta) e^{-\lambda I} \frac{1}{X} (A.35)
$$

Since $X \geq 1$, (A.35) would be non-negative for all $X$ if $(\lambda_I + \lambda_J) Y + \lambda_U Y - \lambda_U \phi(\delta_S) \geq 0$, which is equivalent as:

$$
Y \geq \frac{0.5\lambda_U \phi(\delta)}{\lambda_I + \lambda_J + 0.5\lambda_U} (A.36)
$$

Because $\frac{0.5\phi(\delta)}{\lambda_I + \lambda_J + 0.5\lambda_U} > \frac{0.5\lambda_U \phi(\delta)}{\lambda_I + \lambda_J + 0.5\lambda_U + 0.75\lambda_U \phi(\delta)}$, thus if $Y$ satisfies (A.36) both (A.28) and (A.34) would be non-negative.

Now we verify the results in Proposition 4. From Proposition 1, $s^*(\delta_F)$ and $M^*(\delta_F)$ are the equilibrium spread and depth when all $M$ exchanges have the same order processing speed $\delta_F$. Thus, when $M^*(\delta_F) \leq K$ and suppose HFTs provide liquidity on $M^*(\delta_F)$ fast exchanges with spread $s^*(\delta_F)$, then they will earn non-negative profits. We verify that this is the unique equilibrium. First, no HFTs can earn non-negative profit by providing liquidity on other exchanges. Note that $s^*(\delta_F)$ is the smallest bid-ask spread with which a liquidity provision HFT can earn non-negative profits. Thus, if some HFTs want to provide liquidity on other exchanges, they have to provide liquidity with spread $s^*(\delta_F)$ too. If they provide liquidity on a fast exchange, there profit would be $\pi_F(s^*(\delta_F)/2, M^*(\delta_F) + 1, 0) = \pi(s^*(\delta_F)/2, M^*(\delta_F) + 1|\delta = \delta_F) < 0$ (from equation (1.7)). If they provide liquidity on a slow exchange, their profits would be $\pi_S(s^*(\delta_F)/2, M^*(\delta_F), 1) = 0$.
\[ \pi_S(s^*(\delta_F)/2, M^*(\delta_F) + 1) \leq \pi_F(s^*(\delta_F)/2, M^*(\delta_F) + 1, 0) < 0 \text{(from equation (A.28) and (1.7))}. \]

Secondly, liquidity provision HFT on one among those \( M^*(\delta_f) \) fast exchanges does not have incentive to switch to a slow exchange because \( \pi_F(s^*(\delta_F)/2, M^*(\delta_F), o) \geq \pi_S(s^*(\delta_F)/2, M^*(\delta_F) - 1, 1) \) (from (A.28)). This equilibrium is unique because competition among HFTs will drive the bid-ask spread to \( s^*(\delta_F) \) and since HFTs have larger liquidity provision profits on fast exchanges, they will run to provide liquidity on \( M^*(\delta_F) \) fast exchanges.

For similar reason when \( M^*(\delta_F) > K \), HFTs will provide liquidity on all \( K \) fast exchanges with bid-ask spread \( s^*(\delta_F) \). They will also provide liquidity on slow exchanges until their profits get to negative. Equation (1.18) determines the equilibrium depth on slow exchanges (similar to equation (1.7)).

A.1.7 Proof of Proposition 5

When \( M^*(\delta_F) \leq K \) or \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) < 0 \), the results have been explained in the explanation of equation (1.19). Thus, we only need to prove the results when \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) \geq 0 \). In this case, HFTs provide liquidity on \( M_F^*(K) = K \) fast exchanges and \( M_S^*(K) \) slow exchanges. I will construct trading volume for each exchange in the same way as in Proposition 3.

Starting from \( t = 0 \) the baseline trading game will end and restart when a trade occurs. The idea is to look at each potential path the game will restart from \( t = 0 \). For each path, I will calculate the average time the path will takes (the length of the path), the probability of this path to occur and each exchange’s expected trading volume on that path. Because the trading game is stationary, each exchange’s per unit time trading volume would be the average trading volume among all these paths adjusted by the length of each path.

Specifically, starting from \( t = 0 \) three events may occur: investor arrival, the risky asset’s value jump and undercutting HFT’s arrival. In average it takes \( \frac{1}{\lambda_I} \) time for these events to occur. With probability \( \frac{\lambda_I}{\lambda_I + \gamma} \) an investor arrives first, if she is smart (with probability \( \gamma \)) she will send her market order to a fast exchange. Otherwise, she will randomly choose one among those \( M^*(K) = M_F^*(K) + M_S^*(K) = K + M_S^*(K) \) exchanges to trade with equal probability. Thus, a fast exchange has expected trading volume \( \frac{\gamma}{K} + \frac{1-\gamma}{M_F^*(K)} \) and a slow exchanges have expected trading
volume \((1 - \gamma)\frac{M_s^c(K)}{M-K} \frac{1}{M(K)}\).

With probability \(\frac{\lambda I}{\lambda I + \lambda J}\), the risky asset’s common value jumps first. All stale limit orders on those \(M^*(K)\) exchanges will be taken by either the informed trader or sniping HFTs depends on the jumping is publicly observable or not. Each fast exchange’s expected trading volume is 1. Since slow exchanges have probability \(\frac{M_s^c(K)}{M-K}\) to be chosen by HFTs to provide liquidity and the informed trader or snipers trade on all those \(M^*(K)\) exchanges, thus a slow exchange’s expected trading volume is \(\frac{M_s^c(K)}{M-K}\).

With probability \(\frac{\lambda I}{\lambda I + \lambda J}\) the undercutting HFT arrives first. The baseline game will end and restart either when an investor arrives or the risky asset’s common value jumps. Denoting \(t_U, t_I\) and \(t_J\) as the arriving time of the undercutting HFT, investor and the risky asset’s common value jumping. Since the undercutting HFT will submit her pricing improving order to a fast exchange (Lemma 1), liquidity provision HFT on the fast exchange which is chosen by the undercutting HFT will cancel her being undercut limit order at \(t_U + \delta_F\).

Liquidity provision HFTs on other \(K - 1\) fast exchanges will cancel their being undercut limit orders at \(t_U + \epsilon\) while liquidity provision HFTs on those \(M_s^c(K)\) slow exchanges will cancel their being undercut limit orders at \(t_U + \delta_F + \epsilon - \delta_S\) (see Figure 1.2). We first look at the case when the investor arrives before the risky asset’s common value jumps. That is the case when \(t_I < t_J\). Denoting \(e' = \delta_F + \epsilon - \delta_S\), each exchange’s expected trading volume depends on whether there is trade-through or not.

Trade-through is possible on both fast and slow exchanges if \(t_I < t_U + \epsilon\). From equation (A.9) and (A.11) we have \(\text{Prob}(t_I < t_J, t_I < t_U + \epsilon' | t_I, t_J > t_U) = \frac{\lambda I}{\lambda I + \lambda J} \phi(e')\) and \(E(t_I - t_U | t_U < t_I < t_J < t_U + \epsilon') = \frac{1}{\lambda I + \lambda J} - \frac{1 - \phi(e')}{\phi(e')} \epsilon'\), where \(\phi(e') = 1 - e^{-(\lambda I + \lambda J) \epsilon'}\). In this case, a fast exchange’s expected trading volume is still \(\frac{1}{K} + \frac{1 - \gamma}{M(K)}\) and a slow exchange still has trading volume \((1 - \gamma)\frac{M_s^c(K)}{M-K} \frac{1}{M(K)}\).

Each exchange has exactly the same expected trading volume as in the case when an investor arrives before the undercutting HFT and the risky asset’s common value jumping for exactly the

\[e' = \delta_F + \epsilon - \delta_S\]

\[\phi(e') = 1 - e^{-(\lambda I + \lambda J) \epsilon'}\]

\[\text{Prob}(t_I < t_J, t_I < t_U + \epsilon' | t_I, t_J > t_U) = \frac{\lambda I}{\lambda I + \lambda J} \phi(e')\]

\[E(t_I - t_U | t_U < t_I < t_J < t_U + \epsilon') = \frac{1}{\lambda I + \lambda J} - \frac{1 - \phi(e')}{\phi(e')} \epsilon'\]

\[E(t_I - t_U | t_U < t_I < t_J < t_U + \epsilon') = \frac{1}{\lambda I + \lambda J} - \frac{1 - \phi(e')}{\phi(e')} \epsilon'\]

\[\phi(e') = 1 - e^{-(\lambda I + \lambda J) \epsilon'}\]

\[\phi(e') = 1 - e^{-(\lambda I + \lambda J) \epsilon'}\]

\[\text{Prob}(t_I < t_J, t_I < t_U + \epsilon' | t_I, t_J > t_U) = \frac{\lambda I}{\lambda I + \lambda J} \phi(e')\]

\[E(t_I - t_U | t_U < t_I < t_J < t_U + \epsilon') = \frac{1}{\lambda I + \lambda J} - \frac{1 - \phi(e')}{\phi(e')} \epsilon'\]

\[\phi(e') = 1 - e^{-(\lambda I + \lambda J) \epsilon'}\]
same arguments.

When \( t_U + \varepsilon' \leq t_I \leq t_U + \varepsilon \), liquidity provision HFTs on all \( M_\gamma^*(K) \) slow exchanges have already canceled their being undercut limit orders. A fast exchange has expected trading volume \( \frac{r}{K} + \frac{1-\gamma}{M'(K)} + \frac{1-\gamma}{2} \frac{M_\gamma^*(K)}{M'(K)} = \frac{1}{2} \left( \frac{1+\gamma}{K} + \frac{1-\gamma}{M'(K)} \right) \) (here I use the fact that \( M'(K) = M_\gamma^*(K) + M_\gamma^*(K) = K + M_\gamma^*(K) \)). Comparing to the above case where \( t_I < t_U + \varepsilon' \), fast exchange has additional trading volume \( \frac{1}{2} \frac{1-\gamma}{M'(K)} \). This is because when the investor is at the opposite side of the undercutting HFT (i.e. the investor is a buyer while the undercutting HFT is a seller and vise versa), there is no trade-through on slow exchanges. Slow exchanges will reroute their market orders to fast exchange with better price. With probability \( \frac{1}{2} \) the investor is at the opposite side of the undercutting HFT and only non-smart investor (with probability \( 1 - \gamma \)) will send their orders to a slow exchange. There are total \( M_\gamma^*(K) \) slow exchanges which will reroute their market orders to the fast exchange chosen by the undercutting HFT. Each fast exchange has probability \( \frac{1}{K} \) to be chosen by the undercutting HFT. This explains the additional trading volume for a fast exchange. Similarly, trade occurs on a slow exchange only when the investor is non-smart (with probability \( 1 - \gamma \)) and at the same side of the undercutting HFT (with probability \( \frac{1}{2} \)), the exchange is chosen by HFTs to provide liquidity (with probability \( \frac{M_\gamma^*(K)}{M - K} \)) and is chosen by the investor to trade (with probability \( \frac{1}{M'(K)} \)). Thus, a slow exchange has expected trading volume \( \frac{1-\gamma}{2} \frac{M_\gamma^*(K)}{M - K} \frac{1}{M'(K)} \).

When \( t_I > t_U + \varepsilon \), liquidity provision HFTs on those remaining \( K - 1 \) fast exchanges have canceled their stale limit orders. Thus, no trade-through is possible on any exchange. When the investor is at the opposite side of the undercutting HFT, those remaining \( K - 1 \) fast exchanges have to reroute their market orders to the fast exchange chosen by the undercutting HFT. But this will not affect a fast exchange’s ex ante expected trading volume comparing with the case when \( t_U + \varepsilon' \leq t_I \leq t_U + \varepsilon \) because each fast exchange has equal probability to be chosen by the undercutting HFT. Slow exchanges expected trading volume would not be affected either because trade through is impossible on those \( M_\gamma^*(K) \) slow exchanges as long as \( t_I \geq t_U + \varepsilon' \).

Therefore, we can combine the above two cases and conclude that when \( t_I \geq t_U + \varepsilon' \), a fast and slow exchange has expected trading volume \( \frac{1}{2} \left( \frac{1+\gamma}{K} + \frac{1-\gamma}{M'(K)} \right) \) and \( \frac{1-\gamma}{2} \frac{M_\gamma^*(K)}{M - K} \frac{1}{M'(K)} \) respectively. We have \( \text{Prob}(t_I < t_j, t_I \geq t_U + \varepsilon'|t_I, t_j > t_U) = \frac{\lambda_I}{\lambda_I + \lambda_j} [1 - \phi(\varepsilon')] \) (similar to (A.9)) and \( E(t_I - t_U|t_U < \)

\(^6\)Alternatively, we can directly calculate a fast exchange’s expected trading volume when \( t_I > t_U + \varepsilon \) as \( \frac{1}{2} \left( \frac{1+\gamma}{K} + \frac{1-\gamma}{2} \frac{M_\gamma^*(K)}{M'(K)} \right) \).
\[ t_l < t_j, t_l \geq t_U + \epsilon') = \frac{1}{\lambda_I + \lambda_J} + \epsilon' \] (similar to (A.15)).

The only case left is that the risky asset’s common value jumps before investor’s arrival, that is \( t_U < t_j < t_l \). If \( t_j < t_U + \delta_F \), none liquidity provision HFT has canceled their being undercut limit orders. With probability \( \frac{1}{2} \) the undercutting HFT’s order is also stale (i.e. the undercutting HFT is a seller while the asset’s common value jumps up by \( \sigma \) and vise versa), in this case the trading volume on the fast exchange chosen by the undercutting HFT is 2. If the undercutting HFT’s order is not stale, trading volume on the fast exchange chosen by undercutting HFT is 1.

For all remaining \( K - 1 \) fast exchanges trading volume is always 1. Thus, the expected trading volume on a fast exchange is \( \frac{1}{K} (\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2) + \frac{K-1}{K} \cdot 1 = 1 + \frac{1}{2K} \) when \( t_j < t_U + \delta_F \). For a slow exchange, if it is one among those \( M_s^*(K) \) exchanges having best price quotes, the trading volume would be 1 because the stale limit orders would be taken by either the informed trader or sniper HFTs. Since each slow exchange has probability \( \frac{M_s^*(K)}{M-K} \) to be chosen by HFTs to provide liquidity, thus a slow exchange has expected trading volume \( \frac{M_s^*(K)}{M-K} \) when \( t_j < t_U + \delta_F \). Similar to (A.9) and (A.11) we have \( \text{Prob}(t_j < t_l, t_j < t_U + \delta_F | t_l \text{ t_j > t_U}) = \frac{\lambda_J}{\lambda_I + \lambda_J} \phi(\delta_F) \) and \( E(t_j - t_U | t_U < t_l, t_j < t_U + \delta_F) = \frac{1}{\lambda_I + \lambda_J} - \frac{1}{\phi(\delta_F)} \delta_F \).

If \( t_U + \delta_F \leq t_j \leq t_U + \epsilon' \), only the liquidity provision HFT on the fast exchange chosen by the undercutting HFT 1 has canceled her being undercut limit order. Thus, on all those \( M^*(K) \) exchanges, there is one limit order on both side of the limit order book. Therefore, a fast exchange’s expected trading volume is 1 and a slow exchange has expected trading volume \( \frac{M_s^*(K)}{M-K} \). Similar to (A.13) and (A.14), we have \( \text{Prob}(t_j < t_l, t_U + \epsilon' \leq t_j \leq t_U + \epsilon | t_l \text{ t_j > t_U}) = \frac{\lambda_I}{\lambda_I + \lambda_J} \phi(\epsilon') - \phi(\delta_F) \) and \( E(t_j - t_U | t_U < t_l, t_U + \epsilon' \leq t_j \leq t_U + \epsilon') = \frac{1}{\lambda_I + \lambda_J} + \epsilon' - \frac{\epsilon' - \delta_F}{\phi(\epsilon' - \delta_F)} \).

If \( t_U + \epsilon' < t_j \leq t_U + \epsilon \), only those liquidity provision HFTs on slow exchanges have canceled their stale limit orders. Thus, the expected trading volume on a fast exchange is still 1 while on a slow exchange expected trading volume is only \( \frac{M_s^*(K)}{M-K} \cdot \frac{1}{2} = \frac{M_s^*(K)}{2(M-K)} \) (limit order at the same side of the undercutting HFT on slow exchanges have been canceled). Similarly we have \( \text{Prob}(t_U < t_l, t_U + \epsilon' < t_j \leq t_U + \epsilon | t_l \text{ t_j > t_U}) = \frac{\lambda_I}{\lambda_I + \lambda_J} \phi(\epsilon) - \phi(\epsilon') \) and \( E(t_j - t_U | t_U < t_l, t_U + \epsilon' < t_j \leq t_U + \epsilon) = \frac{1}{\lambda_I + \lambda_J} + \epsilon - \frac{\epsilon - \epsilon'}{\phi(\epsilon - \epsilon')} \).

Finally, when \( t_j > t_U + \epsilon \) all liquidity provision HFTs have canceled their being undercut limit orders. On the fast exchange chosen by the undercutting HFT, there is one limit order on both side of the limit order book because undercutting HFT posts her limit order on this exchange.
too. On all remaining $K - 1$ fast exchanges HFTs only provide liquidity at the opposite side of the undercutting HFT. For example, if the undercutting HFT is a seller, HFTs only provide liquidity on the buy (bid) side of the limit order book. Therefore, a fast exchange’s expected trading volume is $\frac{1}{K} \cdot 1 + \frac{K-1}{K} \cdot \frac{1}{2} = \frac{K+1}{2K}$. On slow exchanges, HFTs only provide liquidity on one side of the limit order book too. Thus, a slow exchange has expected trading volume $\frac{M_s^0(K)}{2(M-K)}$. Similar to the calculation in (A.9) and (A.15), we have $\text{Prob}(t_j < t_i, t_j > t_u + \epsilon | t_i, t_j > t_u) = \frac{\lambda_j}{\lambda_i + \lambda_j} [1 - \phi(\epsilon)]$ and $E(t_j - t_u | t_j < t_i, t_j > t_u + \epsilon) = \frac{1}{\lambda_i + \lambda_j} + \epsilon$.

Now we can calculate each exchange’s expected per unit time trading volume. For a fast exchange, similar to (A.17) we have:

\[
\frac{1}{\Sigma \lambda} + \frac{1}{\lambda_i + \lambda_j} + \epsilon) Q_F(K) = \lambda_l \left[ \frac{1 - \phi(\epsilon)}{\Sigma \lambda} \left[ \frac{1}{K} + \frac{1}{M^*(K)} + \frac{1}{\lambda_i + \lambda_j} \right] \right] + \lambda_r \left[ \frac{1}{\Sigma \lambda} \left[ \frac{1}{1 - \phi(\epsilon)} + \frac{1}{\lambda_i + \lambda_j} \right] Q_F(K) \right] + \frac{\lambda_u}{\Sigma \lambda} \left[ \frac{1}{\lambda_i + \lambda_j} + \epsilon \right) Q_F(K) \right]
\]

Which implies that:

\[
Q_F^*(K) = \lambda_l \left[ \frac{1 - \phi(\epsilon)}{\Sigma \lambda} \left[ \frac{1}{K} + \frac{1}{M^*(K)} \right] \right] + \lambda_r \left[ \frac{1}{\lambda_i + \lambda_j} \left[ \frac{1}{1 - \phi(\epsilon)} \right] \right] + \frac{\lambda_u}{\Sigma \lambda} \left[ \frac{1}{\lambda_i + \lambda_j} + \epsilon \right) Q_F(K) \right]
\]

Similarly, for a slow exchange, we have:

\[
\frac{1}{\Sigma \lambda} + \frac{1}{\lambda_i + \lambda_j} + \epsilon) Q_S(K) = \lambda_l \left[ \frac{M_s^0(K)}{\Sigma \lambda} \left[ \frac{1}{M-K} + 1 - \phi(\epsilon) \right] \right] + \lambda_r \left[ \frac{Q_S(K)}{\lambda_i + \lambda_j} \left[ \frac{1}{M-K} + 1 - \phi(\epsilon) \right] \right] + \frac{\lambda_u}{\Sigma \lambda} \left[ \frac{M_s^0(K)}{M-K} \right] + \frac{\lambda_u}{\Sigma \lambda} \left[ \frac{1}{\lambda_i + \lambda_j} + \epsilon \right)
\]

\[
\text{(A.39)}
\]

120
which implies that:

\[
Q_\ast^\ast(K) = \frac{M'_\ast(K)}{M-K} \left[ \frac{\lambda I(1-\gamma)}{M'(K)} + \lambda J \left[ 1 - \frac{\lambda U}{2\Sigma \lambda} \left[ 1 - \phi(\varepsilon') \right] \right] \right]
\]

(A.40)

(A.37) and (A.39) hold for the same reason as explained in (A.17).

A.1.8 Proof of Corollary 2

(i) When \( M'(\delta_F) \leq K \) or \( M'(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) < 0 \), \( Q_\ast^\ast(K) = 0 \). Thus, in this case we only need to prove that \( Q_\ast^\ast(K) > Q^*(\delta_F) \):

\[
Q_\ast^\ast(K) - Q^*(\delta_F) = \lambda J \left[ \frac{1}{K} + \frac{M'_\ast(K)}{M-K} + \frac{\lambda U \lambda J}{2\Sigma \lambda} \phi(\delta_F) \right] \left( \lambda I + \frac{1 - \phi(e)}{2} \lambda U \lambda J M'_\ast(K) - \frac{1}{M} \right) - \lambda J \left[ \frac{1}{M} - \lambda J \frac{M'(\delta_F)}{M} \right]
\]

\[
= \lambda J \left[ \frac{1}{K} + \frac{1 - \phi(e)}{2} \lambda U \lambda J \left( \frac{1}{M} - \frac{1}{M} \right) + \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{2\Sigma \lambda} \left[ M'(\delta_F) - \frac{M'_\ast(K)}{K} \right] > \lambda J \left[ \frac{M'_\ast(K)}{K} - \frac{M'(\delta_F)}{M} \right] + \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{2\Sigma \lambda} \left[ M'(\delta_F) - \frac{M'_\ast(K)}{K} \right]
\]

\[
\geq 0 \quad (A.41)
\]

The first inequality holds because \( K < M \). The second inequality holds because \( \frac{M'_\ast(K)}{K} \geq M'(\delta_F) \) (recall that \( M'_\ast(K) = \min\{M'(\delta_F), K\} \) and \( 1 \leq M'(\delta_F) \leq M \)). When \( M'(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) \geq 0 \), we first show that \( Q^*(\delta_F) \) is increasing in \( M'(\delta_F) \) and \( Q_\ast^\ast(K) \) is increasing in \( M'(K) \) for a given \( K \):

\[
\frac{dQ^*(\delta_F)}{dM'(\delta_F)} = \left[ 1 - \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{\Sigma \lambda} \right] \frac{\lambda I}{M} > 0
\]

(A.42)

From equation (1.21):

\[
Q_\ast^\ast(K) = \left[ \frac{\lambda I(1-\gamma)}{M-K} \right] \left( 1 - \frac{K}{M'(K)} \right) + \frac{M'(K) - K}{M-K} \lambda J \left[ 1 - \left[ \frac{\lambda U}{2\Sigma \lambda} \left( 1 - \phi(\varepsilon') \right) \right] \right]
\]

(A.43)

Thus, \( Q_\ast^\ast(K) \) is increasing in \( M'(K) \) for a given \( K \). From equation (1.19) and (1.11), we have:

\[
Q_\ast^\ast(K) - Q^*(\delta_F) = \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{M'(K)} \left[ (1-K)(1-\phi(e)) \right] \right] + \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{2\Sigma \lambda} \left[ M'(\delta_F) - \frac{M'_\ast(K)}{K} \right] \right] > \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{M'(K)} \right] + \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{2\Sigma \lambda} \right] \phi(\delta_F) +
\]

\[
\frac{\lambda I}{M} - \lambda J \frac{M'(\delta_F)}{M} - \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{M'(K)} \right] + \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{M'(K)} \right] > \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{M'(K)} \right] + \lambda J \left[ \frac{1 - \phi(e)}{2} \frac{\lambda U \lambda J}{M'(K)} \right] \phi(\delta_F) +
\]

121
For any one, \( T_hus, \) from (A.\( \lambda \)) we have:

\[
(1 - \lambda) (1 - \phi(\lambda)) = \frac{\lambda I}{M} - \lambda J \frac{M}{M} + \frac{\lambda_U \lambda_I \phi(\delta_F)}{2 \Sigma \lambda} \frac{1}{M} - \frac{1 - \phi(\lambda) \lambda_U \lambda_I \phi(\delta_F)}{2 \Sigma \lambda} \frac{1}{M} > \frac{\lambda_U \lambda_I \phi(\delta_F) + (1 - \lambda)(1 - \phi(\lambda))}{M} - \frac{\lambda_U \lambda_I \phi(\delta_F)}{2 \Sigma \lambda} \frac{1}{M} + (1 - \phi(\lambda)) \left( \frac{M - 1}{M} \right) \]

The first inequality holds because \( Q^*(\delta_F) \) is increasing in \( M^*(\delta_F) \) and the last term in \( Q^*_F(K) \) is positive. The second inequality holds because \( \frac{\gamma}{K} + \frac{1 - \gamma}{M^*(K)} > \frac{1}{M} \).

Before comparing \( Q^*(\delta_F) \) and \( Q^*_S(K) \), we first compare \( M^*(\delta_F) \) and \( M^*(K) = K + M^*_S(K) \) when \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) \geq 0 \). In the proof of Proposition 4 we have shown that when \( o < s/2 < \sigma \) and \( \gamma \geq \tilde{\gamma} \):

\[
\pi_F(\frac{s}{2}, X_F, X - X_F) \geq \pi_S(\frac{s}{2}, X) \quad (A.45)
\]

For any \( 1 \leq X_F \leq X \) if \( \pi_S(\frac{s}{2}, X) \geq 0 \), where \( X = X_F + X_S \) and \( \pi_S(\frac{s}{2}, X, X_S) = \pi_S(\frac{s}{2}, X_F, X_S) \) because the later only depends on the sum of \( X_F + X_S \). From equation (1.18), if \( s^*(\delta_F)/2 < \sigma \) we have:

\[
\pi_S(\frac{s^*(\delta_F)}{2}, K, M^*_S(K)) = \pi_S(\frac{s^*(\delta_F)}{2}, K, M^*_S(K)) \geq 0 \quad (A.46)
\]

Thus, from (A.45) we have:

\[
\pi(\frac{s^*(\delta_F)}{2}, K, M^*_S(K) \mid \delta = \delta_F) = \pi_F(\frac{s^*(\delta_F)}{2}, K, M^*_S(K), o) \geq \pi_S(\frac{s^*(\delta_F)}{2}, K, M^*_S(K)) \geq 0 \quad (A.47)
\]

Therefore, from equation (1.7) we conclude that \( M^*(\delta_F) \geq M^*(K) = K + M^*_S(K) \). From equation (1.11) and (1.21) we have:

\[
Q^*(\delta_F) - Q^*_S(K) = \lambda_I \frac{1}{M} + \lambda_J \frac{M^*(\delta_F)}{M} + \lambda_U \lambda_I \frac{\phi(\delta_F)}{2 \Sigma \lambda} \frac{1}{M} - \frac{1 - \phi(\lambda)}{2 \Sigma \lambda} \frac{\lambda_I \lambda_J M^*(\delta_F) - 1}{M} - \frac{M^*(K) - K}{M - K} \frac{\lambda_I (1 - \gamma)}{M^*_F(K)}
\]

\[
+ \lambda_J \left[ \frac{1 - \lambda_U}{2 \Sigma \lambda} \frac{1 - \phi(\lambda')}{1 - \phi(\lambda')} \right] \geq \lambda_I \frac{1}{M} + \lambda_J \frac{M^*(\delta_F)}{M} + \lambda_U \lambda_I \frac{\phi(\delta_F)}{2 \Sigma \lambda} \frac{1}{M} - \frac{1 - \phi(\lambda)}{2 \Sigma \lambda} \frac{\lambda_I \lambda_J M^*(\delta_F) - 1}{M} - \frac{M^*(\delta_F) - K}{M - K} \times
\]

\[
\left[ \frac{\lambda_I \phi(\delta_F)}{M^*(\delta_F)} + \lambda_J \left[ \frac{1 - \lambda_U}{2 \Sigma \lambda} \frac{1 - \phi(\lambda)}{1 - \phi(\lambda')} \right] \right] \times \left[ \frac{\lambda_I \phi(\delta_F)}{M^*(\delta_F)} + \lambda_J \left[ \frac{1 - \lambda_U}{2 \Sigma \lambda} \frac{1 - \phi(\lambda)}{1 - \phi(\lambda')} \right] \right] \geq \lambda_J \left( \frac{M^*(\delta_F) - M^*(\delta_F) - K}{M - K} \right) - \frac{1 - \phi(\lambda)}{2 \Sigma \lambda} \frac{\lambda_I \lambda_J M^*(\delta_F) - 1}{M} - \frac{M^*(\delta_F) - K}{M - K} \times
\]

\[
\left[ \frac{\lambda_I \phi(\delta_F)}{M^*(\delta_F)} + \lambda_J \left[ \frac{1 - \lambda_U}{2 \Sigma \lambda} \frac{1 - \phi(\lambda)}{1 - \phi(\lambda')} \right] \right] \times \left[ \frac{\lambda_I \phi(\delta_F)}{M^*(\delta_F)} + \lambda_J \left[ \frac{1 - \lambda_U}{2 \Sigma \lambda} \frac{1 - \phi(\lambda)}{1 - \phi(\lambda')} \right] \right] \geq 0 \quad (A.48)
\]

The first inequality holds because \( Q^*_S(K) \) is increasing in \( M^*(K) \) and \( M^*(K) \leq M^*(\delta_F) \). The second inequality holds because \( \frac{1}{M} \geq \frac{M^*(\delta_F) - K}{M - K} \). The third inequality holds because \( \phi(\lambda') < \phi(\lambda) \).
(ii) When $M^*(\delta_F) \leq K$ or $M^*(\delta_F) > K$ and $\pi_S(s^*(\delta_F)/2, K, 1) < 0$, then $M'_F(1) = \min\{M^*(\delta_F), 1\} = 1$. When $M^*(\delta_S) < M$, from equation (1.19) and (1.11), we have:

$$Q^*_F(1) - Q^*(\delta_S) = \lambda_I + \lambda_J + \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_F) - \lambda_I M - \lambda_J \frac{M^*(\delta_S)}{M} + \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U \lambda_J M^*(\delta_S) - 1}{M} \geq \lambda_I + \lambda_J$$

$$+ \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_F) - \frac{\lambda_I}{M} \lambda_J M - \lambda_J M - 1 - \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_S) + \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U \lambda_J M - 2}{M} > \frac{\lambda_J}{M} \left[ 1 - \frac{\lambda_U}{2 \Sigma \lambda} \phi(\delta_S) \right] > 0$$  (A.49)

The first inequality holds because $Q^*(\delta_S)$ is increasing in $M^*(\delta_S)$ and $M^*(\delta_S) \leq M - 1$. When $M^*(\delta_S) = M$, from (A.49) we have:

$$Q^*_F(1) - Q^*(\delta_S) = \lambda_I + \lambda_J + \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_F) - \lambda_I \frac{M}{M} \lambda_J \frac{M^*(\delta_S)}{M} + \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U \lambda_J M - 1}{M} > \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \left[ \phi(\delta_F) - \frac{\phi(\delta_S)}{M} + \frac{(1 - \phi(\epsilon))}{M} \right] \geq 0$$  (A.50)

If:

$$\frac{\phi(\delta_F)}{M} - \frac{\phi(\delta_S)}{M} + \frac{(1 - \phi(\epsilon))}{M} \frac{M - 1}{M} \geq 0 \iff \phi(\delta_S) \leq M \phi(\delta_F) + (M - 1)[1 - \phi(\epsilon)]$$  (A.51)

Similarly, when $M^*(\delta_F) > K$, $\pi_S(s^*(\delta_F)/2, K, 1) \geq 0$ and $M^*(\delta_S) < M$ from equation (1.19) and (1.11) we have:

$$Q^*_F(1) - Q^*(\delta_S) = \lambda_I \left[ 1 - \frac{\pi_S(s^*(\delta_F)/2, K, 1)}{M^*(\delta_F)} \right] + \lambda_J + \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_F) + \frac{\lambda_U \lambda_I}{2 \Sigma \lambda} \left[ 1 - \phi(\epsilon') \right] \left[ 1 - \gamma \right] M^*(\delta_S) - 1 \right) - \frac{\lambda_I}{M} \frac{M^*(\delta_S)}{M}$$

$$- \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_S) + \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U \lambda_J \phi(\delta_S)}{M} > \frac{\lambda_J}{M} \left[ M - 1 \right] - \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_S) \geq 0$$  (A.52)

The first inequality holds because $Q^*(\delta_S)$ is increasing in $M^*(\delta_S)$ and $M^*(\delta_S) \leq M - 1$. When $M^*(\delta_S) = M$, from (A.52) we have:

$$Q^*_F(1) - Q^*(\delta_S) = \lambda_I \left[ 1 - \frac{\pi_S(s^*(\delta_F)/2, K, 1)}{M^*(\delta_F)} \right] + \lambda_J + \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_F) + \frac{\lambda_U \lambda_I}{2 \Sigma \lambda} \left[ 1 - \phi(\epsilon') \right] \left[ 1 - \gamma \right] M^*(\delta_S) - 1 \right) - \frac{\lambda_I}{M} \frac{M^*(\delta_S)}{M}$$

$$- \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_S) + \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U \lambda_J \phi(\delta_S)}{M} > \frac{\lambda_J}{M} \left[ M - 1 \right] - \frac{\lambda_U \lambda_J}{2 \Sigma \lambda} \phi(\delta_S) \geq 0$$  (A.53)

if (A.51) holds.
A.1.9 Proof of Proposition 6

(i) The results are directly implied from Corollary 2. Considering exchange \( i \), when all remaining exchanges have slow order processing speed, speeding up is profitable for exchange \( i \) if:

\[
\tilde{f} Q^*(\delta_S) < \tilde{f} Q^*_F(1) - C_{\text{speed}} \iff \frac{C_{\text{speed}}}{\tilde{f}} < Q^*_F(1) - Q^*(\delta_S) \tag{A.54}
\]

If there are \( 1 \leq K \leq M - 1 \) exchanges among the remaining \( M - 1 \) exchanges have fast order processing speed, then it is profitable for exchange \( i \) to speed up if:

\[
\tilde{f} Q^*_S(K) < \tilde{f} Q^*(\delta_F) - C_{\text{speed}} \iff \frac{C_{\text{speed}}}{\tilde{f}} < Q^*(\delta_F) - Q^*_S(K) \tag{A.55}
\]

This is because \( \tilde{f} Q^*(\delta_F) - C_{\text{speed}} \leq \tilde{f} Q^*_F(K + 1) - C_{\text{speed}} \) from Corollary 2 (i).\(^7\) The later is exchange \( i \)'s per unit time profit if it speeds up. Now we show that \( Q^*_S(K) \) is decreasing in \( K \). From Proposition 5, when \( M^*(\delta_F) \leq K \) or \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) < 0 \), \( Q^*_S(K) = 0 \). Thus when \( K \) increases, \( Q^*_S(K) \) is more likely to be zero. When \( M^*(\delta_F) > K \) and \( \pi_S(s^*(\delta_F)/2, K, 1) > 0 \), \( Q^*_S(K) \) is determined in equation (1.21). Note that the maximum \( X_F + X_S \) such that equation (1.17) is non-negative at \( s^*(\delta_F) \) is independent of \( K \) and denote this maximum depth as \( \hat{M} \). Thus, \( M^*_S(K) = \hat{M} - K \). Equation (1.21) can be rewritten as:

\[
Q^*_S(K) = \frac{\hat{M} - K}{\hat{M} - \lambda_f(1 - \gamma)M} + \lambda_f \{ 1 - \frac{\lambda_f}{2\Sigma\lambda}[1 - \phi(\epsilon')]} \tag{A.56}
\]

which is clearly decreasing in \( K \). Therefore, as long as \( C_{\text{speed}}/\tilde{f} < \min\{Q^*_S(1) - Q^*(\delta_S), Q^*(\delta_F) - Q^*_S(1)\} \) (the right-hand side is positive because of Corollary 2) exchange \( i \) will always speed up no matter how many other exchanges have increased their order processing speed. Thus, investing in the new speed technology is a dominant strategy for all exchanges.

(ii) When all exchanges speed up, each exchange’s expected per unit time profit decreases if:

\[
\tilde{f} Q^*(\delta_F) - C_{\text{speed}} < \tilde{f} Q^*(\delta_S) \iff \frac{C_{\text{speed}}}{\tilde{f}} > Q^*(\delta_F) - Q^*(\delta_S) \tag{A.57}
\]

\(^7\)Note that Corollary 2 (i) hold for any \( 1 \leq K \leq M - 1 \). When the total number of fast exchanges is \( M \), thus all exchanges have the same order processing speed \( \delta_F \). We have \( Q^*_F(M) = Q^*(\delta_F) \).
(iii) is the same result as in Proposition 2 (iii) by simply minus the taker fee from investor’s welfare defined in equation (1.10). Thus, it is the same proof of Proposition 2 (iii).

A.2 Equilibrium Analysis with Exchange Speed Heterogeneity When \( \gamma < \tilde{\gamma} \)

When \( \gamma < \tilde{\gamma} \) it is possible that \( \pi_F\left(\frac{s}{2}, X_F, X_S\right) < \pi_S\left(\frac{s}{2}, X_F, X_S\right) \) for some \( \frac{s}{2}, X_F \) and \( X_S \) because undercutting HFT always submit her order to fast exchanges. This will cause problems when constructing the equilibrium. For example, when \( K = 1 \) without loss of generality, we can assume exchange 1 is the only fast exchange. Since exchange 1 has faster order processing speed than other exchanges, liquidity-providing HFTs can potentially provide liquidity with the smallest bid-ask spread on exchange 1. Thus, a natural equilibrium one could think would be that HFTs provide liquidity on exchange 1 at the bid-ask spread \( s^* (\delta_F) \) and other \( M^* (1) \) (determined in equation (1.18)) slow exchanges. But if \( \pi_F\left(\frac{s^* (\delta_F)}{2}, 1, M^* (1)\right) < \pi_S\left(\frac{s^* (\delta_F)}{2}, 1, M^* (1)\right) \), it is possible that \( \pi_F\left(\frac{s^* (\delta_F)}{2}, 1, M^* (1)\right) < 0 \). Then the above results could not be in equilibrium if the liquidity-providing HFT on exchange 1 only provides liquidity on that exchange because she earns negative profits. But if the liquidity-providing HFT on exchange 1 also provides liquidity on some slow exchanges, her total profit might be positive and thus potentially could be an equilibrium. Therefore, when \( \gamma < \tilde{\gamma} \), to construct the equilibrium we should allow a single HFT to provide liquidity on multiple exchanges. General speaking, there could be two kinds of equilibrium depending on whether HFTs provide liquidity on fast exchanges or not. Specifically, for a given integer \( 0 \leq J \leq M - K \) (remember that \( K \) is the total number of fast exchanges) denote \( X^*_F(s/2, K, J) \) and \( X^*_S(s/2, K, J) \) as the solution for the following problem:

\[
\max_{\frac{s}{2} \leq X_F \leq K; \frac{s}{2} \leq X_S \leq M - K} X_F \pi_F(s/2, X_F, X_S) + (X_S - J)\pi_S(s/2, X_F, X_S) \tag{A.58}
\]

s.t. \( \pi_S(s/2, X_F, X_S + 1) < 0 \) if \( X_S + 1 \leq M - K \);

\( \pi_F(s/2, X_F + 1, X_S) < 0 \) if \( X_F + 1 \leq K \)
what problem (A.58) solves is the maximum liquidity provision profit a single HFT can earn if she provides liquidity on some exchanges (at least one fast exchange) with bid-ask spread \( s \) and other HFTs have already provide liquidity on \( J \) slow exchanges with the same spread \( s \). Moreover, no further HFTs can enter the market to make non-negative liquidity provision profits with the same spread. Note that the solution for problem (A.58) always exists because \( X_F = K \) and \( X_S = M - K \) satisfy all conditions in problem (A.58). Further, we denote:

\[
J^*(s/2, K) = \max \{1 \leq J \leq M - K \text{ s.t. } \pi(s/2, J|\delta = \delta_S) \geq 0\} \tag{A.59}
\]

where \( \pi(s/2, J(K)|\delta = \delta_S) \) is the liquidity provision profit (on one exchange) when HFTs provide liquidity on \( J(K) \) slow exchanges with bid-ask spread \( s/2 \) (this profit function is defined in equation (1.5)). Define:

\[
TP_1(s/2, K, J) = X_F^* \pi_F(s/2, X_F^*, X_S^*) + [X_S^* - J] \pi_S(s/2, X_F^*, X_S^*) \tag{A.60}
\]

Where \( X_F^* = X_F^*(s/2, K, J) \) and \( X_S^* = X_S^*(s/2, K, J) \) are the solutions for problem (A.58). And:

\[
TP_2(s/2, K) = J^*(s/2, K)\pi(s/2, J^*(s/2, K)|\delta = \delta_S) \tag{A.61}
\]

We will discuss three potential equilibriums. Denoting \( E_1 \) as the case when a single HFT provides liquidity on \( X_F^*(s^*(\delta_F)/2, K, o) \) fast exchanges and \( X_S^*(s^*(\delta_F)/2, K, o) \) slow exchanges with bid-ask spread \( s^*(\delta_F) \); \( E_2 \) as the case when a single HFT provides liquidity on \( X_F^*(s^*(\delta_S)/2, K, o) \) fast exchanges and \( X_S^*(s^*(\delta_S)/2, K, o) \) slow exchanges with bid-ask spread \( s^*(\delta_S) \) and \( E_3 \) as the case when a single HFT provides liquidity on \( J^*(s^*(\delta_S)/2, K) \) slow exchanges with bid-ask spread \( s^*(\delta_S) \). We have the following results:

**Proposition 13** (Equilibrium with Exchange Speed Heterogeneity When \( \gamma < \bar{\gamma} \)) If there are \( K \) fast exchanges with order processing speed \( \delta_F \) and \( M - K \) slow exchanges with order processing speed \( \delta_S \), where \( 1 \leq K \leq M - 1 \) and \( \delta_F < \delta_S \). When \( \gamma < \bar{\gamma} \), then:

(i) If \( s^*(\delta_F) < s^*(\delta_S) \) and \( TP_1(s^*(\delta_F)/2, K, o) \geq 0 \), \( E_1 \) is an equilibrium;

(ii) If \( s^*(\delta_F) = s^*(\delta_S) \) or \( TP_1(s^*(\delta_F)/2, K, o) < 0 \), either \( E_2 \) or \( E_3 \) could be an equilibrium: 1) if
\[ TP_1(s^*(\delta_S)/2, K, o) > TP_2(s^*(\delta_S)/2, K) \], \text{ } E_2 \text{ is an equilibrium; } 2) \text{ if } TP_1(s^*(\delta_S)/2, K, o) < 0, E_3 \text{ is an equilibrium; } 3) \text{ if } o \leq TP_1(s^*(\delta_S)/2, K, o) \leq TP_2(s^*(\delta_S)/2, K), \text{ then } E_2 \text{ is an equilibrium if: }

\[ TP_1(s^*(\delta_S)/2, K, J^*(s^*(\delta_S)/2, K)) \geq 0 \quad \text{(A.62)} \]

Otherwise, \( E_3 \) is an equilibrium.

**Proof.** In (i) a single HFT can make non-negative profits by providing liquidity on \( X^*_F(s^*(\delta_F)/2, K, o) \) fast exchanges and \( X^*_S(s^*(\delta_F)/2, K, o) \) slow exchanges with bid-ask spread \( s^*(\delta_F) \). Since \( X^*_F(s^*(\delta_F)/2, K, o) \) and \( X^*_S(s^*(\delta_F)/2, K, o) \) are solutions for problem (A.58), thus no additional HFTs can enter the market to make non-negative profit with bid-ask spread \( s^*(\delta_F) \), which is the smallest spread. Therefore, we only need to check whether the single HFT has incentive to change her current quotes. If she cancels her quotes on all fast exchanges and only provide liquidity on some slow exchanges, she has to quote with bid-ask spread \( s^*(\delta_S) \). But other HFTs can enter the market to provide liquidity on \( X^*_F(s^*(\delta_F)/2, K, o) \) fast exchanges and \( X^*_S(s^*(\delta_F)/2, K, o) \) slow exchanges with bid-ask spread \( s^*(\delta_F) \). Then the original single HFT will loss her liquidity provision profits. If she still keep some quotes on fast exchanges, other HFTs will response the quotes changes until the conditions in problem (A.58) hold. The original single HFT’s profit would be smaller than the profits by providing liquidity on \( X^*_F(s^*(\delta_F)/2, K, o) \) fast exchanges and \( X^*_S(s^*(\delta_F)/2, K, o) \) slow exchanges with bid-ask spread \( s^*(\delta_F) \) because the later maximizes the total liquidity provision profits (solving problem (A.58) with \( J = o \)). Therefore, the original single HFT has no incentive to change her quotes.

In (ii), HFTs will provide liquidity with bid-ask spread \( s^*(\delta_S) \). When \( TP_1(s^*(\delta_S)/2, K, o) > TP_2(s^*(\delta_S)/2, K) \), the single HFT provides liquidity on \( X^*_F(s^*(\delta_S)/2, K, o) \) fast exchanges and \( X^*_S(s^*(\delta_S)/2, K, o) \) slow exchanges attaining the maximum liquidity provision profits. Thus she has no incentive to change her quotes. For the same reason, other HFTs can not enter the market to provide liquidity. Therefore, \( E_2 \) is an equilibrium. When \( TP_1(s^*(\delta_S)/2, K, o) < 0 \), no HFT can make non-negative profit by providing liquidity on some fast exchanges because as long as other HFTs response to the quotes and satisfies the conditions in problem (A.58), the total profit is negative. Thus, the original HFT who provides liquidity on some fast exchanges will have negative profits. In this case, \( E_3 \) is an equilibrium, in which no HFTs provide liquidity on fast exchanges.
When $0 \leq TP_1(s^*(\delta_S)/2, K, o) \leq TP_2(s^*(\delta_S)/2, K)$, a single HFT has larger liquidity provision profits in $E_3$ than in $E_2$. $E_3$ is an equilibrium only when no other HFTs can enter the market and earn non-negative liquidity provision profits. Since there is already a HFT proving liquidity on $J^*(s^*(\delta_S)/2, K)$ slow exchanges with bid-ask spread $s^*(\delta_S)$, the largest liquidity provision profit an entrant HFT could earn is $TP_1(s^*(\delta_S)/2, K, J^*(s^*(\delta_S)/2, K))$. As long as this is negative, $E_3$ would be an equilibrium. The reason why $E_2$ is an equilibrium when $TP_1(s^*(\delta_S)/2, K, J^*(s^*(\delta_S)/2, K)) \geq 0$ is because the single HFT who provides liquidity on $X^*_f(s^*(\delta_S)/2, K, o)$ fast exchanges and $X^*_S(s^*(\delta_S)/2, K, o)$ slow exchanges with bid-ask spread $s^*(\delta_S)$ has no incentive to change her quotes. Since $TP_1(s^*(\delta_S)/2, K, o) \leq TP_2(s^*(\delta_S)/2, K)$, one would think that this single HFT has incentive to provide liquidity only on slow exchanges as in $E_3$. But unfortunately, only providing liquidity on slow exchanges is not an equilibrium because other HFTs can enter the market to earn $TP_1(s^*(\delta_S)/2, K, J^*(s^*(\delta_S)/2, K))$. Therefore, in this case HFTs always provide liquidity on fast exchanges and the largest liquidity provision profit a single HFT can earn is $TP_1(s^*(\delta_S)/2, K, o)$. Thus, the original single HFT has no incentive to change her current quotes. So $E_2$ is an equilibrium.

Note that when $\gamma < \bar{\gamma}$, the above equilibriums in Proposition 13 are not necessarily unique. In $E_1$, $E_2$ and $E_3$, I allow a single HFT to provide liquidity on all possible exchanges. Other equilibrium may exists when HFTs provide liquidity on some exchanges not all possible exchanges. I ignore this analysis because the equilibrium would depends on other parameters such as $\lambda_I$, $\lambda_J$ and $\lambda_U$. But for most cases, the equilibrium features the same spread and depth as stated in Proposition 13.
Appendix B

Appendix for Chapter 2

B.1 Proofs

B.1.1 Proof of Proposition 7

Denoting $\pi_{\text{LOB}}(\frac{s}{2})$ and $\pi_{\text{FBA}}(\frac{s}{2}, \tau)$ as the profit for a HFT who provides liquidity on the LOB or the FBA exchange (with trading frequency $\tau$) at bid-ask spread $s$. $\pi_{\text{LOB}}(\frac{s}{2})$ is determined in equation (1.5) when $X = 1$, thus $\pi_{\text{LOB}}(\frac{s}{2}) = \pi(\frac{s}{2}, 1)$. Denote $\Phi(\tau) = 1 - e^{-\Delta \tau}$, $\phi(\tau - x) = 1 - e^{-(\lambda_I + \lambda_f)(\tau - x)}$ and $f(\tau) = \int_0^\tau \phi(\tau - x)d\psi(\tau)$, then:

$$
\pi_{\text{FBA}}(\frac{s}{2}, \tau) = \Phi(\tau) \frac{\lambda_I}{\sum \lambda} \frac{s}{2} - \Phi(\tau) \frac{\mu \lambda_f}{\sum \lambda}(\sigma - \frac{s}{2}) + \Phi(\tau) \frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} \frac{s}{2} - \frac{\mu \lambda_f}{\lambda_I + \lambda_f}(\sigma - \frac{s}{2}) 
$$

(B.1)

Note that $\Phi(\tau)$ is the probability of an event (the arrival of investor, common value jump or undercutting HFT) happening within one auction round. $f(\tau)$ is the probability of either an investor or the asset’s value jumps within the same auction round after undercutting HFT’s arrival. No sniping on FBA exchange, thus only private news can cause adverse selection cost. Other terms in (B.1) have similar explanations as in equation (1.5) when $X = 1$. Similar to equation (1.6):

$$
\frac{s^*_{\text{LOB}}}{2} = \min\{\frac{s}{2} | \nu_0 \pm \frac{s}{2} \in \mathcal{P}, \pi_{\text{LOB}}(\frac{s}{2}) \geq 0\} 
$$

(B.2)

$$
\frac{s^*_{\text{FBA}}(\tau)}{2} = \min\{\frac{s}{2} | \nu_0 \pm \frac{s}{2} \in \mathcal{P}, \pi_{\text{FBA}}(\frac{s}{2}, \tau) \geq 0\} = \min\{\frac{s}{2} | \nu_0 \pm \frac{s}{2} \in \mathcal{P}, \pi_{\text{FBA}}(\frac{s}{2}, \tau) \geq 0\} 
$$

(B.3)

The second equality holds in (B.3) because $\Phi(\tau) > 0$.

$$
\pi_{\text{LOB}}(\frac{s}{2}) - \frac{\pi_{\text{FBA}}(\frac{s}{2}, \tau)}{\Phi(\tau)} = \frac{(1 - \mu) \lambda_f}{\sum \lambda} (\sigma - \frac{s}{2}) + \frac{\lambda_I}{\sum \lambda} \frac{1}{2} \frac{s}{2} - \phi(\delta) - \frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} (\sigma - \frac{s}{2}) + \frac{\lambda_I}{\sum \lambda} (1 - f(\tau)) \frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} \frac{s}{2} 
$$

129
\[
\frac{\lambda_U \lambda_f}{\Sigma \lambda \lambda_I + \lambda_f} (\sigma - \frac{s}{2}) + \frac{\lambda_U f(\tau)}{\Sigma \lambda} \frac{\mu \lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s}{2}) \tag{B.4}
\]

(B.4) > 0 if \( \mu \geq \frac{\lambda_I + \lambda_f + \lambda_U [1 - \frac{\lambda_I}{\lambda_f + \lambda_f} f(\tau)]}{\lambda_I + \lambda_f + \lambda_U f(\tau)} \). Thus under this condition, \( s_{LOB}^* \leq s_{FBA}^*(\tau) \).

\[
\frac{\pi_{FBA}(\frac{s}{2}, \tau_1)}{\Phi(\tau_1)} - \frac{\pi_{FBA}(\frac{s}{2}, \tau_2)}{\Phi(\tau_2)} = \frac{\pi_{FBA}(\frac{s}{2}, \tau_1)}{\Phi(\tau_1)} \frac{\pi_{FBA}(\frac{s}{2}, \tau_2)}{\Phi(\tau_2)} = \frac{\lambda_U}{\Sigma \lambda} [f(\tau_1) - f(\tau_2)][\frac{\lambda_I}{\lambda_I + \lambda_f} + \frac{1}{2} s - \frac{\mu \lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s}{2})] \tag{B.5}
\]

If \( \frac{\lambda_I}{\lambda_I + \lambda_f} \frac{s_{FBA}(\tau_1)}{2} - \frac{\mu \lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s_{FBA}(\tau_2)}{2}) \geq 0, \) then from (B.1) \( \pi_{FBA}(\frac{s_{FBA}(\tau_2)}{2}, \tau_1) \geq 0. \) Thus, \( s_{FBA}^*(\tau_1) \leq s_{FBA}^*(\tau_2) \).

If \( \frac{\lambda_I}{\lambda_I + \lambda_f} \frac{s_{FBA}(\tau_1)}{2} - \frac{\mu \lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s_{FBA}(\tau_2)}{2}) < 0 \) and since \( f(\tau) \) is increasing in \( \tau \) (\( f'(0) = 0 \) and \( f''(\tau) > 0 \)), \( \frac{\pi_{FBA}(\frac{s}{2}, \tau_1)}{\Phi(\tau_1)} > \frac{\pi_{FBA}(\frac{s}{2}, \tau_2)}{\Phi(\tau_2)} \), from (B.5) when \( \tau_1 < \tau_2 \). Thus, in either case we always have \( s_{FBA}^*(\tau_1) \leq s_{FBA}^*(\tau_2) \).

B.1.2 Proof of Proposition 8

(i) is straightforward. In (ii), we need to show when \( \gamma \geq \bar{\gamma} \) all HFTs prefer to provide liquidity on the LTD exchange. As in the proof of Proposition 4, denote \( E = \frac{\lambda_I}{\lambda_I + \lambda_f} \frac{s}{2} > 0, F = \frac{\lambda_I}{\lambda_I + \lambda_f} (\sigma - \frac{s}{2}) > 0 \) and \( F' = \frac{\mu \lambda_f}{\lambda_I + \lambda_f} (\sigma - \frac{s}{2}) > 0 \). Further denote \( \pi_{LTD}(\frac{s}{2}, X) \) as the HFT’s liquidity provision profit on the LTD exchange at bid-ask spread \( s \) and HFTs also provide liquidity on other \( X - 1 \) LOB exchanges at the same spread, where \( X \geq 2 \). Similar to (A.26), we have:

\[
\pi_{LTD}(\frac{s}{2}, X) = \frac{\lambda_I}{\Sigma \lambda} \gamma s + (1 - \gamma) \frac{\lambda_I + \lambda_f}{\Sigma \lambda} E - \frac{\lambda_I + \lambda_f}{\Sigma \lambda} F + \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_F)(\frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} s + \frac{1}{2} - \frac{1}{2} E - F') + \frac{\lambda_U}{\Sigma \lambda} [1 - \phi(\delta_F)] \\
\times (\frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} s + \frac{1}{2} - \frac{1}{2} E - F') \tag{B.6}
\]

From equation (2.3), we have:

\[
\pi_{LOB}(\frac{s}{2}, X) = (1 - \gamma) \frac{\lambda_I + \lambda_f}{\Sigma \lambda} E - \frac{\lambda_I + \lambda_f}{\Sigma \lambda} F + \frac{\lambda_U}{\Sigma \lambda} \phi(e)[(1 - \gamma)E - F] + \frac{\lambda_U}{\Sigma \lambda} [1 - \phi(e)][(1 - \gamma)E - F] \tag{B.7}
\]

Therefore,

\[
\pi_{LTD}(\frac{s}{2}, X) - \pi_{LOB}(\frac{s}{2}, X) = \frac{\lambda_I}{\Sigma \lambda} \gamma s + \frac{\lambda_I + \lambda_f}{\Sigma \lambda} (1 - \mu) F - \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_F) \frac{F'}{2} + \frac{\lambda_U}{\Sigma \lambda} \phi(e) \frac{(1 - \gamma)E - F}{2} + \frac{\lambda_U}{\Sigma \lambda} (\frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} s + \frac{1}{2} - \frac{1}{2} E - F') \times (\frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} s + \frac{1}{2} - \frac{1}{2} E - F') \geq \frac{\lambda_I}{\Sigma \lambda} \gamma s - \frac{\lambda_U}{\Sigma \lambda} \phi(e) \frac{(1 - \gamma)E}{2} + \frac{\lambda_U}{\Sigma \lambda} \frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} s + \frac{1}{2} \frac{1}{2} E - F' + \frac{\lambda_U}{\Sigma \lambda} (\frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} s + \frac{1}{2} - \frac{1}{2} E - F') \times (\frac{\lambda_I}{\lambda_I + \lambda_f} \frac{1}{2} s + \frac{1}{2} - \frac{1}{2} E - F')
\]
\[ 0.5\lambda_{UY} \geq \frac{s}{2\Sigma\lambda} \frac{\lambda_I}{\lambda_I + \lambda_J} ((\lambda_I + \lambda_J)\gamma - \lambda_U\phi(\epsilon)\frac{1 - \gamma}{4} + 0.5\lambda_{UY}) \]  

(B.8)

The last inequality holds because \( X \geq 2 \). Therefore, when

\[ (\lambda_I + \lambda_J)\gamma - \lambda_U\phi(\epsilon)\frac{1 - \gamma}{4} + 0.5\lambda_{UY} \geq 0 \Leftrightarrow \gamma \geq \frac{0.25\lambda_U\phi(\epsilon)}{\lambda_I + \lambda_J + 0.5\lambda_U + 0.25\lambda_U\phi(\epsilon)} \]  

(B.9)

(B.8) is non-negative. Note that \( \frac{0.25\lambda_U\phi(\epsilon)}{\lambda_I + \lambda_J + 0.5\lambda_U + 0.25\lambda_U\phi(\epsilon)} < \tilde{\gamma} \). Thus when \( \gamma \geq \tilde{\gamma} \), \( \pi_{LTD}(\frac{\epsilon}{2}, X) \geq \pi_{LOB}(\frac{\epsilon}{2}, X) \). Similar to (A.34) we also have:

\[ \pi_{LTD}(\frac{s}{2}, X) - \pi(\frac{s}{2}, X|\delta = \delta_F) = \frac{\lambda_I}{\Sigma\lambda} s - \gamma \frac{\lambda_I + \lambda_J}{\Sigma\lambda} E + \frac{\lambda_I + \lambda_J}{\Sigma\lambda} (1 - \mu)F - \frac{\lambda_U\phi(\delta_F)}{\Sigma\lambda} E - \frac{\lambda_U}{\Sigma\lambda} F \]

\[ \times \frac{X - 1}{X} \phi(\epsilon) + \frac{\lambda_U}{\Sigma\lambda} \left( \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{\gamma s}{2} - \frac{\gamma}{2} E + \frac{1 - \mu}{2} F \right) \geq \frac{\lambda_I}{\Sigma\lambda} s - \gamma \frac{\lambda_I + \lambda_J}{\Sigma\lambda} E + \frac{\lambda_I + \lambda_J}{\Sigma\lambda} \frac{\lambda_I}{\lambda_I + \lambda_J} \frac{\gamma s}{2} - \frac{\gamma}{2} E - \frac{\lambda_U}{\Sigma\lambda} F \]

\[ \times \frac{X - 1}{X} \phi(\epsilon) + \frac{\lambda_U}{\Sigma\lambda} \left( \frac{1}{X} \phi(\delta_F) + \frac{X - 1}{X} \phi(\epsilon) - \phi(\delta_F) \right) \]  

(B.10)

Since \( \frac{1}{X} \phi(\delta_F) + \frac{X - 1}{X} \phi(\epsilon) - \phi(\delta_F) \geq 0 \) (because \( \delta_F < \epsilon \)), (B.10) is non-negative when \( \gamma \geq \tilde{\gamma} \) for exactly the same argument as in (A.35). Regarding of undercutting HFT, the result in Lemma 1 directly applies to here. Thus when LTD is implemented, HFTs will provide liquidity on the LTD exchange first with equilibrium bid-ask spread \( s_{LTD}^* \) determined in equation (2.2). HFTs will provide liquidity on other LOB exchanges as long as they can earn non-negative liquidity provision profit. The equilibrium depth \( M_{LTD}^* \) is determined in equation (2.4). The LTD exchange’s expected per unit time trading volume could be calculated similarly to the proof of Proposition 5.

Alternatively, we could also directly use the result in equation (1.19) with \( K = 1 \) and by replacing \( M^*(K) \) with \( M_{LTD}^* \), \( \lambda_I \) with \( \mu \lambda_J \) (no sniping on the LTD exchange), \( \epsilon \) with \( \epsilon \) (HFTs on the LOB exchange will cancel their stale limit orders \( \delta_F + \epsilon - \delta_F = \epsilon \) units of time after the undercutting HFT’s arrival) and \( M_{LTD}^*(K) = M_{LTD}^* - 1 \). This will generate the result in equation (2.5).  

(iii) The ND exchange will attain its maximum per unit time trading volume only when all HFTs prefer to provide liquidity on the ND exchange. Therefore, our goal is to check under which conditions ND exchange will be preferred by all HFTs. Similar to (B.8) we have:

\[ \text{[Equation]} \]
\[
\pi_{ND}(s/2, X) - \pi'_{LOB}(s/2, X) = \frac{\lambda_I}{\Sigma \lambda} s + \frac{\lambda_I + \lambda_J}{2} (1 - \mu) F - \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_S) E - \frac{\lambda_U}{\Sigma \lambda} \phi(\epsilon'') \left( \frac{1 - \gamma}{2} E - \frac{\lambda_U}{\Sigma \lambda} \phi\left( \frac{\lambda_I}{\lambda_I + \lambda_J} \right) \right) \left( \frac{1 - \mu}{2} F \right) - \frac{\lambda_U}{\Sigma \lambda} \phi\left( \frac{\lambda_I}{\lambda_I + \lambda_J} \right) \phi\left( \frac{s}{2} \right) \phi(\epsilon'') \frac{1}{2} \left( \frac{1 - \mu}{2} F \right) \right)
\]

where \( \epsilon'' = \delta_S + \epsilon - \delta_F \). Thus for the same argument as in (B.9), when \( \gamma \geq \frac{0.25\lambda_U \phi(\epsilon'')}{\lambda_I + \lambda_J + 0.5\lambda_U + 0.25\lambda_U \phi(\epsilon'')} \), (B.11) is non-negative. Note that \( \frac{\lambda_I + \lambda_J + 0.5\lambda_U + 0.25\lambda_U \phi(\epsilon'')}{\lambda_I + \lambda_J + 0.5\lambda_U} < \frac{1}{\lambda_I + \lambda_J + 0.5\lambda_U} \) because \( \phi(\epsilon'') < 2\phi(\epsilon) \) \((\phi(\cdot) \text{ is concave}, \phi(0) = 0 \text{ and } \delta_F < \delta_S < \epsilon) \). Therefore, \( \pi_{ND}(s/2, X) \geq \pi'_{LOB}(s/2, X) \) when \( \gamma \geq \bar{\gamma} \).

Similar to (B.10),

\[
\pi_{ND}(s/2, X) - \pi(s/2, X) | \delta = \delta_F = \frac{\lambda_I}{\Sigma \lambda} s + \frac{\lambda_I + \lambda_J}{2} (1 - \mu) F - \frac{\lambda_U}{\Sigma \lambda} \phi(\delta_S) E - \frac{\lambda_U}{\Sigma \lambda} \phi(\epsilon'') \left( \frac{1 - \gamma}{2} E - \frac{\lambda_U}{\Sigma \lambda} \phi(\epsilon'') \left( \frac{1 - \mu}{2} F \right) \right) - \frac{\lambda_U}{\Sigma \lambda} \phi\left( \frac{\lambda_I}{\lambda_I + \lambda_J} \right) \phi\left( \frac{s}{2} \right) \phi(\epsilon'') \left( \frac{1 - \mu}{2} F \right) \right)
\]

\[
X - 1 \quad \phi(\epsilon) \left( \frac{1 - \mu}{2} F \right)
\]

The last part in (B.12) is increasing in \( X \). Therefore, when \( X = 1 \) and:

\[
\phi(\delta_F) - \mu \phi(\delta_S) + (1 - \mu) + \frac{\lambda_I + \lambda_J}{\lambda_U} (1 - \mu) = 0 \Leftrightarrow \mu \phi(\delta_S) \leq \phi(\delta_F) + (1 + 2 \frac{\lambda_I + \lambda_J}{\lambda_U})(1 - \mu)
\]  

(B.13)

the last part in (B.12) is non-negative for all \( X \). Note that for any given \( \mu > 0 \) as long as \( \delta_S \) is not too larger than \( \delta_F \), (B.13) will hold. The remaining part in (B.12) is non-negative when \( \gamma \geq \bar{\gamma} \) (see (A.35)). Thus, \( \pi_{ND}(s/2, X) \geq \pi(\delta_F) | \delta = \delta_F \) when (B.13) holds and \( \gamma \geq \bar{\gamma} \).

Now we check under which conditions undercutting HFT prefers to provide liquidity on the ND exchange. Lemma 1 does not directly apply to here because the ND exchange takes \( \delta_S \) units of time to process undercutting HFT’s order while LOB exchange only takes \( \delta_F \) units of time. Intuitively, if delaying time \( \Delta_{speed} = \delta_S - \delta_F \) is not too large, undercutting HFT will prefer to provide liquidity on the ND exchange because no sniping on ND exchange. If all HFTs (including undercutting HFT) prefer to provide liquidity on the ND exchange, then the equilibrium spread and depth would be \( \delta_{ND}^*\text{ and } \delta_{ND}^* \). Similar to the proof of Lemma 1, suppose the undercutting HFT submits her order to the ND exchange. Since no sniping on ND exchange, the undercutting

\( ^2 \)When we compare \( \pi_{ND}(s/2, X) \) and \( \pi'_{LOB}(s/2, X) \) the minimum of \( X \) is 2 because \( \pi'_{LOB}(s/2, X) \) is only defined for \( X \geq 2 \). Since \( \pi(s/2, X) | \delta = \delta_F \) is well defined for \( X = 1 \), when compare \( \pi_{ND}(s/2, X) \) and \( \pi(s/2, X) | \delta = \delta_F \), we choose the minimum of \( X \) to be 1.
HFT’s payoff would be:

\[
\frac{1}{2} \{ \phi(e'') \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \left( \frac{s_{ND}^*}{2} - d \right) - \frac{\mu \lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s_{ND}^*}{2} + d) \right] + [1 - \phi(e'')] \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \left( \frac{s_{ND}^*}{2} - d \right) - \frac{\mu \lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s_{ND}^*}{2} + d) \right] \}
\]

where \( e'' = \delta_S + \epsilon - \delta_F \) (HFTs on LOB exchanges will cancel their stale limit order \( e'' \) units of time after undercutting HFT’s arrival). If the undercutting HFT submits her order to one LOB exchange (among those \( M_{ND}^* - 1 \) LOB exchanges), her payoff is:

\[
\frac{1}{2} \{ \phi(e') \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \left( \frac{s_{ND}^*}{2} - d \right) - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s_{ND}^*}{2} + d) \right] + \phi(\epsilon) - \phi(e') \left[ \frac{\lambda_I}{\lambda_I + \lambda_J} \left( \frac{s_{ND}^*}{2} - d \right) - \frac{\lambda_J}{\lambda_I + \lambda_J} (\sigma - \frac{s_{ND}^*}{2} + d) \right] \}
\]

where \( e' = \delta_F + \epsilon - \delta_S \) (HFT on the ND and other LOB exchanges will cancel their stale limit orders \( e' \) and \( \epsilon \) units of time after undercutting HFT’s arrival respectively). (B.14) – (B.15) \( \geq 0 \) is equivalent as:

\[
(1 - \mu) \lambda_J (\sigma - \frac{s_{ND}^*}{2} + d) - \frac{\lambda_I}{M_{ND}^*} \left( \frac{s_{ND}^*}{2} - d \right) [(M_{ND}^* - 1)(\phi(e'') - \phi(\epsilon)) + \phi(\epsilon) - \phi(e')] \geq 0 \]

(B.16)

Therefore, when \( \gamma \geq \bar{\gamma} \), (B.13) and (B.16) hold, all HFTs prefer to provide liquidity on the ND exchange, which generates maximum trading volume for ND exchange. When \( s_{ND}^* < s^*(\delta_F) \), the reason we do not need conditions in (B.13) and (B.16) is because \( M_{ND}^* = 1 \) when \( s_{ND}^* < s^*(\delta_F) \). HFTs can not provide liquidity on other LOB exchanges with non-negative profit. Thus, undercutting HFT will prefer the ND exchange too because it is the unique exchange with current best price quotes. The ND exchange’s maximum trading volume is calculated exactly in the same way as LTD case. It could be directly calculated as in the proof of Proposition 5 or adjusted from equation (1.19).

---

3Suppose \( M_{ND}^* \geq 2 \) when \( s_{ND}^* < s^*(\delta_F) \), then \( \pi_{LOB}^* (\frac{s_{ND}^*}{2}, M_{ND}^*) \geq 0 \), which implies that \((1 - \gamma)E - F \) evaluated at \( s_{ND}^* \) is non-negative. Thus, \( \pi_0 (\frac{s_{ND}^*}{2}, M_{ND}^*) \geq 0 \) (defined in (A.27)). From the proof of Proposition 4, we can conclude that HFTs can earn non-negative liquidity provision profit by providing liquidity on \( M_{ND}^* \) LOB exchanges with spread \( s_{ND}^* \). This contradicts to the fact that the minimum spread on LOB exchange is \( s^*(\delta_F) > s_{ND}^* \).
B.1.3 Proof of Corollary 3

From equation (1.11), \(Q^*(\delta_F)\) is increasing in \(M^*(\delta_F)\). Therefore, \(\min Q^*(\delta_F) = \lambda I \frac{1}{M} + \lambda J \frac{1}{M} + \frac{\lambda U \lambda J}{2 \Sigma \lambda} \phi(\delta) \frac{1}{M}\), which is the volume when \(M^*(\delta_F) = 1\). From equation (2.5) and (2.6), both \(Q^*_{LTD}\) and \(Q^*_{ND}\) are decreasing in the equilibrium depth. They reach their maximum when the depth is one. Thus, \(\max\{Q^*_{LTD}, Q^*_{ND}\} \leq Q^* (\delta_F)\) if:

\[
\max \{\lambda I (y + \frac{1 - y}{1}) + \mu \lambda J + \frac{\lambda U \lambda J}{2 \Sigma \lambda} \mu \phi(\delta_F), \lambda I (y + \frac{1 - y}{1}) + \mu \lambda J + \frac{\lambda U \lambda J}{2 \Sigma \lambda} \mu \phi(\delta_S)\} \leq \lambda I \frac{1}{M} + \lambda J \frac{1}{M} + \frac{\lambda U \lambda J}{2 \Sigma \lambda} \phi(\delta) \frac{1}{M}
\]

(B.17)

which generates the condition on \(\mu\) in Corollary 3.
Appendix C

Appendix for Chapter 3

C.1 Proofs

C.1.1 Proof of Lemma 2

For the $Q^{th}$ share in the queue at the half bid-ask spread $s/2$, we define its value for the liquidity supplier as $LP_{s/2}(Q)$ and its value for each sniper as $SN_{s/2}(Q)$. In all proofs, we drop the subscript if $s/2 = d/2$. HFTs race to supply liquidity for the first share at $\pm d/2$ iff $LP(1) > SN(1)$.

We consider the first share on the ask side in the proof, and the race on the bid side follows symmetrically. When tick size in binding, both BATs and non-algo traders demand liquidity, so we use non-HFTs to refer to both in the proofs of Lemma 2 and Proposition 9. A non-HFT seller does not change the state of the LOB; an non-HFT buyer, who arrives with probability $\frac{1}{2} \lambda_I$, provides a profit of $d/2$ to HFT liquidity supplier; fundamental value jumps up with probability $\frac{1}{2} \frac{\lambda_I}{\lambda_I + \lambda_J}$ and costs an HFT firm $\frac{d N - 1}{2 N}$; fundamental value jumps down with probability $\frac{1}{2} \frac{\lambda_J}{\lambda_I + \lambda_J}$, which reduces the value of the current queue position to $0$. Therefore:

$$LP(1) = \frac{\frac{1}{2} \lambda_I}{\lambda_I + \lambda_J} \frac{d}{2} + \frac{\frac{1}{2} \lambda_I}{\lambda_I + \lambda_J} LP(1) = \frac{\frac{1}{2} \lambda_J}{\lambda_I + \lambda_J} \frac{d N - 1}{2 N} + \frac{\frac{1}{2} \lambda_J}{\lambda_I + \lambda_J} \cdot 0$$

Each sniper has a probability of $1/N$ to snipe the stale quote after an upward value jump. A successful sniping leads to a profit of $d/2$, so:

$$SN(1) = \frac{\lambda_J}{\lambda_I + 2 \lambda_J} \frac{d}{2 N}$$
LP(1) > SN(1) ⇐ \frac{\lambda_I}{\lambda_I + 2\lambda_J} \cdot \frac{d}{2} - \frac{\lambda_J}{\lambda_I + 2\lambda_J} \cdot \frac{d}{2} \cdot \frac{N - 1}{N} > \frac{\lambda_J}{\lambda_I + 2\lambda_J} \cdot \frac{d}{2} \cdot \frac{1}{N}

\frac{\lambda_I}{\lambda_J} > 1

Therefore, the tick size is binding at \( d/2 \) if \( \frac{\lambda_I}{\lambda_J} > 1 \).

C.1.2 Proof of Lemma 3

We prove Lemma 3 using mathematical induction.

1. From the proof for Lemma 2,

\[
LP(1) = \frac{\lambda_I}{\lambda_I + 2\lambda_J} \cdot \frac{d}{2} - \frac{1}{2} \left[ 1 - \frac{\lambda_I}{\lambda_I + 2\lambda_J} \right] \cdot \frac{d}{2} \cdot \frac{N - 1}{N}
\]

which satisfies Equation (3.3).

2. Suppose that Equation (3.3) holds for some \( Q \in \mathbb{N}^+ \). The following proof shows that it holds for \( Q + 1 \in \mathbb{N}^+ \) as well.

\[
LP(Q + 1) = \frac{\lambda_I}{\lambda_I + 2\lambda_J} \cdot LP(Q) - \frac{\lambda_J}{\lambda_I + 2\lambda_J} \cdot \frac{d}{2} \cdot \frac{N - 1}{N} = (\frac{\lambda_I}{\lambda_I + 2\lambda_J})^{Q+1} \cdot \frac{d}{2} \cdot \frac{1}{2} \cdot \frac{\lambda_I}{\lambda_I + 2\lambda_J} - (\frac{\lambda_I}{\lambda_I + 2\lambda_J})^{Q+1} \cdot \frac{d}{2} \cdot \frac{1}{N}
\]

Thus, Equation (3.3) holds with \( Q \) replaced by \( Q + 1 \). Hence Equation (3.3) holds for all \( Q \in \mathbb{N}^+ \).

C.1.3 Proof of Proposition 10

BATs use flash limit orders when regular limit orders are more costly. We start the proof by finding the boundary between the flash equilibrium and the undercutting equilibrium.

In an undercutting equilibrium, a BAT submits a limit order to an empty LOB \((0, 0)\) and changes the state to \((1, 0)\); a BAT submits a limit order to \((0, 1)\) and changes the state to \((1, 1)\). We denote the cost for the first case as \(C(1, 0)\) and the cost for the second case as \(C(1, 1)\). Then
\[
\begin{aligned}
&C(1, 0) = p_1 C(1, 1) + p_1 C(1, 0) + p_2 \left( -\frac{d}{6} \right) + p_2 C(1, 0) + p_3 \frac{5d}{6} + p_3 C(1, 0) \\
&C(1, 1) = p_1 \left( -\frac{d}{6} \right) + p_1 C(1, 0) + p_2 \left( -\frac{d}{6} \right) + p_2 C(1, 0) + p_3 \frac{5d}{6} + p_3 C(1, 0)
\end{aligned}
\]  \tag{C.1}

In Equation (C.1) and Figure 3.8, we describe six event types that can change the LOB in an undercutting equilibrium. Consider $C(1, 0)$ on the ask side. A BAT buyer and a BAT seller each arrive each with probability $p_1$. A BAT buyer posts a limit order on the bid side and changes the state to $C(1, 1)$; a BAT seller uses a flash limit order so the state remains at $C(1, 0)$. A non-algo buyer and a non-algo seller arrive each with probability $p_2$. The BAT seller enjoys a negative transaction cost of $-d/6$ when the non-algo buyer takes his liquidity; the non-algo seller hits a HFT’s quote on the bid side and does not change the state on the ask side. Upward and downward value jumps occur with probability $p_3$. An upward jump leads to a sniping cost of $5d/6$, whereas a downward jump does not change the state of the LOB.\(^1\) $C(1, 1)$ differs in two ways from $C(1, 0)$. First, the arrival of a BAT buyer leads to execution of a sell limit order from a BAT.\(^2\) Second, a downward jump under $C(1, 1)$ leads to sniping on the opposite side of the LOB and changes the state to $C(1, 0)$.

If an undercutting order gets immediate execution, the cost $-d/6$. $C(1, 1)$ must be greater than $-d/6$ because of the cost of being sniped. Therefore, $C(1, 0) - C(1, 1) = p_1(C(1, 1) + d/6) > 0$. Intuitively, if a BAT chooses to post a sell limit order at $v_t + d/6$ on an empty LOB, he must post a sell limit order when the bid side has a limit order posed by a BAT, because the existence of a limit order on the bid side increases the execution probability for a limit order on the ask side. Note that our model starts with no limit orders from BATs, so $C(1, 0) < d/6$ is needed to jumpstart the undercutting equilibrium.

The solution for equation Equation (C.1) is:

\[
C(1, 1) = \frac{(-2 + \beta)\lambda I + 10\lambda J d}{(2 - \beta)\lambda I + 2\lambda J} = \frac{(-2 + \beta)R + 10 d}{(2 - \beta)R + 2 d}
\]

\(^1\)Here we assume that BATs position their order one tick above the new fundamental value. BATs are able to reposition their orders because they face no competition from other BATs in a short time period.

\(^2\)The execution of this order results from our assumption that BATs do not queue after another limit order at the same price, but the intuition that a longer queue on the bid side increases the execution probability on the ask side holds true generally (Parlour (1998)).
\[ C(1, 0) = \frac{d}{6} \left[ \frac{\beta R \left( (-2 + \beta)R + 10 \right)}{R + 1} + \frac{5 - (1 - \beta)R}{(2 - \beta)R + 2} \right] \]

\[ C(1, 0) < \frac{d}{6} \text{ iff } \frac{\beta R \left( (-2 + \beta)R + 10 \right)}{R + 1} + \frac{5 - (1 - \beta)R}{(2 - \beta)R + 2} < 1 \]

i.e.,

\[ (2 - \beta)R^2 + (-2 - 4\beta)R - 4 > 0 \]

Equation \((2 - \beta)R^2 + (-2 - 4\beta)R - 4 = 0\) has two roots: \(R_{1,2} = \frac{1 + 2\beta \pm \sqrt{4\beta^2 + 9}}{2 - \beta}\),

\[ R_2 < 0, \; R_1 = \frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta} \]

So BATs choose to undercut when \(R > R_1\), because \(C(1, 0) < d/6\); BATs choose to flash when \(R > R_2\).

Above is the boundary between undercutting equilibrium and flash equilibrium. On both sides of the boundary, we let a BAT buyer (seller) use limit order to respond to the other side’s limit order. Such a response is both rational and necessary. It is rational because \(C(1, 1) < C(1, 0) = d/6\), thus a limit order response, which costs \(C(1, 1)\), is strictly better than flash order. It is necessary because otherwise all BATs buyers (sellers) will still use flash orders when off-equilibrium sell (buy) order is present. The off-equilibrium sell (buy) order will have an execution cost as follows:

\[ C(1, 0) = p_1\left(-\frac{d}{6}\right) + p_2C(1, 0) + p_3\left(-\frac{d}{6}\right) + p_2C(1, 0) + p_3\frac{5d}{6} + p_3C(1, 0) \]

\[ C(1, 0) = \frac{d}{6} \frac{5 - R}{1 + R} \]

\[ C(1, 0) < \frac{d}{6} \iff R > 2 \]

Thus, undercutting is an optimal deviation when \(\frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta} > R > 2\). The existence

\[ ^3\text{In flash equilibrium, any BAT’s undercutting limit order is off-equilibrium.} \]
of deviation proves that, in the $R > 2$ region of flash equilibrium, BATs should use limit orders to respond to the other side’s off-equilibrium undercutting order, otherwise the off-equilibrium undercutting order will become a profitable deviation.

However, in the $R < 2$ region of flash equilibrium, BATs should use flash orders to respond to the other side’s off-equilibrium undercutting order, because the cost of limit order response, $C(1, 1)$, is larger than $d/6$. On the other hand, even if other BATs use flash orders, the deviator is still not profiting.

In other words, regardless of whether $R > 2$ or $R < 2$, the equilibrium outcome is the same, but BATs need to use different rational strategies in off-equilibrium paths to eliminate profitable deviations, thus these deviations will never appear under equilibrium.

To sum up, the complete strategy (including the optimal response to off-equilibrium paths) of a BATs seller under flash equilibrium is:

1. If $R < 2 < \frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta}$, use limit order under off-equilibrium path:
   i. If there is no order at $-d/6$, submit a limit sell order at $-d/6$.
   ii. Else, submit a limit sell order at $d/6$.

2. If $R < 2$, use flash order under off-equilibrium path:
   i. Submit a limit sell order at $-d/6$ regardless of state of the book.

BATs buyer’s strategy is symmetric. These strategies will generate the equilibrium outcome sketched in Proposition 10. Predictions on depth and HFT participation follow the proof of Proposition 9.

C.1.4 Proof of Proposition 11

1. In Proposition 10, we address the boundary between the flash equilibrium and the undercutting equilibrium.

2. The solution for HFT depth follows from Figure 3.5 and Equation (3.14). The depth decreases because the revenue from liquidity supply for HFTs decreases. BATs never take HFTs’ liquidity at $d/2$, and BATs can also supply liquidity to non-algo traders. The decreased revenue for HFTs also reduces their entry.

3. Equation (3.14) can be solved for any $R$ and $\beta$. Here we give an example for $R = 4$ and
$\beta = 0.1$. First, we assume that all $D^{(i,j)}(1) > 0$. Thus we solve:

\begin{align*}
D^{(0,0)}(1) &= p_1D^{(0,1)}(1) + p_1D^{(1,0)}(1) + p_2\frac{d}{2} + p_3(-\frac{d}{2}) + p_3 \cdot 0 \\
D^{(1,0)}(1) &= p_1D^{(1,1)}(1) + p_1D^{(1,0)}(1) + p_2D^{(0,0)}(1) + p_2d^{(1,0)}(1) + p_3(-\frac{d}{2}) + p_3 \cdot 0 \\
D^{(0,1)}(1) &= p_1D^{(0,1)}(1) + p_1D^{(1,1)}(1) + p_2\frac{d}{2} + p_2D^{(0,0)}(1) + p_3(-\frac{d}{2}) + p_3 \cdot 0 \\
D^{(1,1)}(1) &= p_1D^{(0,1)}(1) + p_1D^{(1,0)}(1) + p_2D^{(0,0)}(1) + p_3(-\frac{d}{2}) + p_3 \cdot 0
\end{align*}

We then obtain:

\begin{align*}
D^{(0,0)}(1) &= 8 + 12R + 12\beta R - 4R^2 + 24\beta^2 R^2 + 2\beta^2 R^2 - 12R^3 + 21\beta R^3 - 2\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4 \\
&= 0.2202 \\
D^{(1,0)}(1) &= 8 + 24R + 20R^2 + 6\beta R^2 - 4\beta^2 R^2 + 12\beta R^3 - 5\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4 \\
&= 0.0527 \\
D^{(0,1)}(1) &= 8 + 12R + 12\beta R - 4R^2 + 24\beta R^2 + 2\beta^2 R^2 - 12R^3 + 21\beta R^3 - 5\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4 \\
&= 0.2205 \\
D^{(1,1)}(1) &= 8 + 24R + 20R^2 + 2\beta^2 R^2 + 6\beta R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4 \\
&= 0.0593
\end{align*}

$D^{(i,j)}(1) > 0$ is satisfied. Therefore, the depth is at least one share in any state of the LOB.

Then we assume all $D^{(i,j)}(2) > 0$. Thus, we solve:

\begin{align*}
D^{(0,0)}(2) &= p_1D^{(0,1)}(2) + p_1D^{(1,0)}(2) + p_2D^{(0,0)}(2) + p_2D^{(0,0)}(2) + p_3(-\frac{d}{2}) + p_3 \cdot 0
\end{align*}
\[D^{(1,0)}(2) = p_1 D^{(1,1)}(2) + p_1 D^{(1,0)}(2) + p_2 D^{(0,0)}(2) + p_2 D^{(1,0)}(2) + p_3 \left(\frac{d}{2}\right) + p_3 \cdot o\]

\[D^{(0,1)}(2) = p_1 D^{(0,1)}(2) + p_1 D^{(1,1)}(2) + p_2 D^{(0,0)}(1) + p_2 D^{(1,0)}(2) + p_3 \left(\frac{d}{2}\right) + p_3 \cdot o\]

\[D^{(1,1)}(2) = p_1 D^{(0,1)}(2) + p_1 D^{(1,0)}(2) + p_2 D^{(0,1)}(2) + p_2 D^{(1,0)}(2) + p_3 \left(\frac{d}{2}\right) + p_3 \cdot o\]

We get:\(^{4}\)

\[D^{(0,0)}(2) = 0.0448\]

\[D^{(1,0)}(2) = -0.0602 < 0\]

\[D^{(0,1)}(2) = 0.0451\]

\[D^{(1,1)}(2) = -0.0561 < 0\]

We reject the assumption that all \(D(2) > 0\). Therefore, under certain states of the LOB, HFTs would not supply the second share of liquidity. We start from the worst state for liquidity suppliers, \((1, 0)\), in which a BAT undercuts HFTs on the same side of the LOB, but no BAT undercuts HFTs on the other side of LOB.\(^{5}\) Therefore, \(D^{(1,0)}(2) = 0\) and all other \(D^{(i,j)}(2) > 0\). Thus we solve:

\[D^{(0,0)}(2) = p_1 D^{(0,1)}(2) + p_2 D^{(0,0)}(1) + p_2 D^{(0,0)}(2) + p_3 \left(\frac{d}{2}\right) + p_3 \cdot o\]

\[D^{(0,1)}(2) = p_1 D^{(0,1)}(2) + p_1 D^{(1,1)}(2) + p_2 D^{(0,1)}(1) + p_2 D^{(0,0)}(2) + p_3 \left(\frac{d}{2}\right) + p_3 \cdot o\]

\[D^{(1,1)}(2) = p_1 D^{(0,1)}(2) + p_2 D^{(0,1)}(2) + p_2 D^{(1,0)}(2) + p_3 \left(\frac{d}{2}\right) + p_3 \cdot o\]

We obtain:

\[D^{(0,0)}(2) = 0.0475\]

\[D^{(0,1)}(2) = 0.0487\]

\[D^{(1,1)}(2) = -0.0310 < 0\]

\(^{4}\)For briefness, the closed-form solution is not presented, but it is available upon request.

\(^{5}\)In this state, an HFT liquidity supplier on the ask side cannot trade with the next non-HFT buyer, because a BAT buyer chooses to supply liquidity and changes the state to \((1, 1)\), and a non-algo buyer chooses to take the BAT seller’s liquidity and changes the state to \((0, 0)\).
However, $D^{(1,1)}(2)$ is still smaller than 0. We further assume that $D^{(1,1)}(2)$ is also 0, i.e., HFTs cancel the second order when BATs submit limit orders on both sides. Therefore,

$$D^{(0,0)}(2) = p_1 D^{(0,1)}(2) + p_2 D^{(0,0)}(1) + p_2 (-\frac{d}{2}) + p_3 \cdot 0$$

$$D^{(0,1)}(2) = p_1 D^{(0,1)}(2) + p_2 D^{(0,1)}(1) + p_2 D^{(0,0)}(2) + p_3 (-\frac{d}{2}) + p_3 \cdot 0$$

We obtain:

$$D^{(0,0)}(2) = 0.0488$$

$$D^{(0,1)}(2) = 0.0489$$

Further calculation shows $D^{(0,0)}(3) = 0, D^{(0,1)}(3) = 0$. We then conclude that $Q^{(0,0)} = Q^{(0,1)} = 2$ and $Q^{(1,0)} = Q^{(1,1)} = 1$ is the solution for Equation (3.14) under $R = 4$ and $\beta = 0.1$.

C.1.5 Proof of Proposition 12

HFTs do not compete to supply liquidity at $5d/6$ when:

$$LP_{sL}^{0d} (1) < SN_{sL}^{0d} (1)$$

$$LP_{sL}^{0d} (1) = p_1 \cdot LP_{sL}^{0d} (1) + p_1 \cdot 0 + p_2 \cdot \frac{5d}{6} + p_2 \cdot LP_{sL}^{0d} (1) - p_3 \frac{d N - 1}{N} + p_3 \cdot 0$$

$$LP_{sL}^{0d} (1) = \frac{(1 - \beta) \lambda_l}{\lambda_l + 2 \lambda_j} \frac{5d}{6} - \frac{\lambda_j}{\lambda_l + 2 \lambda_j} \frac{d N - 1}{N}$$

$$SN_{sL}^{0d} (1) = \frac{\lambda_j}{\lambda_l + 2 \lambda_j} \frac{d}{6 N}$$

$$\frac{(1 - \beta) \lambda_l}{\lambda_l + 2 \lambda_j} \frac{5d}{6} - \frac{\lambda_j}{\lambda_l + 2 \lambda_j} \frac{d N - 1}{N} < \frac{\lambda_j}{\lambda_l + 2 \lambda_j} \frac{d}{6 N}$$

$$R < \frac{1}{5(1 - \beta)}$$

Thus, HFTs supply liquidity at $7d/6$. WOLOG, we consider a BATs seller’s strategy. The complete strategy (including the optimal response to off-equilibrium paths, see proof of Proposition...
On page 10) of a BAT seller is:

1. If there is no limit sell order on $d/6$, $d/2$, and $5d/6$, submit a limit sell order at $5d/6$.
2. Else, if there is no limit buy order on $-d/6$, submit a limit sell order at $-d/6$.
3. Else, there is a limit buy order on $-d/6$ (this is an off-equilibrium path, there are two possible responses, same intuition as the proof of Proposition 10).
   i. If $R > 2$, submit a limit sell order at $d/6$, costs $C(1, 1)$.
   ii. Else, submit a limit sell order at $-d/6$, costs $d/6$.

If all BATs follow this strategy, no limit sell (buy) order will be present at $d/2$ ($-d/2$) or $d/6$ ($-d/6$). We show that a deviator will suffer a higher execution cost.

Firstly, a BAT seller will not post a limit sell order at $d/2$, because only a non-algo buy order will trade with this seller. The seller’s execution cost is:

$$C = p_1 \cdot C + p_1 \cdot C + p_2 \cdot \left(-\frac{d}{2}\right) + p_2 \cdot C + p_3 \cdot \frac{d}{2} + p_3 \cdot C$$

$$C = \frac{d}{2} \cdot \frac{-\left(1 - \beta\right)R + 1}{\left(1 - \beta\right)R + 1}$$

Since in flash crash equilibrium $R(1 - \beta) < \frac{1}{5}$, the BAT’s cost is at least $\frac{d}{2} \cdot \frac{4/5}{6/5} = \frac{d}{3} > \frac{d}{6} = Cost of flash order. Thus, it is never optimal to submit a limit order at $d/2$.

Secondly, the BAT seller will not post a limit sell order at $d/6$. In this case, non-algo traders and other BAT buyers might trade with the seller: the non-algo trader will execute a buy order and a BAT will execute a flash buy order (when he cannot or finds not optimal to post a limit buy order at $-d/6$). The intuition is similar with Equation (C.1) and Figure 3.8, but in the flash crash equilibrium, the BAT seller faces equal or higher costs than in an undercutting equilibrium: The BATs buyer does not have to post a limit buy order in a flash crash equilibrium. The solution of formula Equation (C.1) is:

$$R_1 = \frac{1 + 2\beta + \sqrt{4\beta^2 + 9}}{2 - \beta}$$

However, there is no combination of $(R, \beta)$ in the flash crash equilibrium that satisfies $R > R_1$.

Finally, the BAT seller will post a sell limit order at $5d/6$. Her cost is:
\[ C = p_1 \cdot C + p_1 \cdot C + p_2 \left( \frac{-5d}{6} \right) + p_2 \cdot C + p_3 \cdot \frac{d}{6} + p_3 \cdot C \]

\[ C = \frac{d - 5(1 - \beta)R + 1}{6} < \frac{d}{6} \]