Why Do Stock Exchanges Compete on Speed, and How?

Xin Wang

Click here for the latest version

April 2, 2018

Abstract

This paper 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.

Keywords: Exchange Speed, High-frequency Trading, Order Protection Rule
JEL Codes: D47, G10, G14, G18

I thank my advisors Dan Bernhardt, Mao Ye and Neil Pearson for their outstanding guidance and support. I would also like to thank Christine Parlour, Craig Holden, Joshua Pollet, Alexei Tchistyi, Adam Clark-Joseph, Jiekun Huang, Julian Reif, Yufeng Wu, Mathias Kronlund, Hayden Melton, Simona Abis, Veronika Pool, Russ Wermers, Avanidhar Subrahmanyam, Daniel Andrei and participants at the University of Illinois Lunch Time Research Seminar for their valuable comments. This work also uses the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant #OCI-1053575. I thank David O'Neal of the Pittsburgh Supercomputer Center for his assistance with supercomputing, which was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program. Mailing address: University of Illinois, 214 David Kinley Hall, 1407 W Gregary Dr, Urbana, IL, 61801. Email: xinwang5@illinois.edu. Tel: 217-419-9000.

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, Jones, and Menkveld (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, Cramton, and Shim (2015, BCS) 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 undercutting 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*

[1] Nowadays, HFTs are the main liquidity providers in equity markets, as documented in Brogaard, Hendershott, and Riordan (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.

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

[3] Currently, despite the availability of high-speed microwave connectivity, stock exchanges still use fiber-optic

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 IEXs 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 exchange status, IEX attracted more trading volume in stocks with binding bid-ask spreads relative to stocks with non-binding bid-ask spreads.[5] 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

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.

[5] IEX 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.

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, Foucault, and Moinas (2015), Budish, Cramton, and Shim (2015, BCS), 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.[6]

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, Cramton, and Shim (2015, BCS) 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 competing against other faster exchanges?[7] 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 (2017b), I use the same approach to explore how newly designed

[6] Menkveld 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.

[7] In 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`.

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, Saar, and Zhong (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-though 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 2 sets up the model. Section 3 studies speed competition between exchanges. Empirical tests are presented in Section 4. Section 5 concludes. All proofs are in the Appendix.

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-though time window*. I endogenize an exchange's speed investment in Section 3.

2.1 Model Setup

Exchanges and limit order book. M exchanges use continuous limit order book to conduct trades.[8] 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

[8] Currently 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 ownd 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.

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 $\mathscr{P} = \{p^i\}_{i=-\infty}^{\infty}$ denote the discrete set of available prices for quoting and trading: the distance between any two consecutive prices in $\mathscr{P}$ 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 λ_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 γ 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 λ_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 σ. If σ 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 λ_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 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.

2.2 Exchanges Order Processing and Connection speeds

Let δ_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 δ_i indicates a faster order processing speed. At the cutting edge of technology, δ_i is about 50 microseconds. I denote the time that it takes to

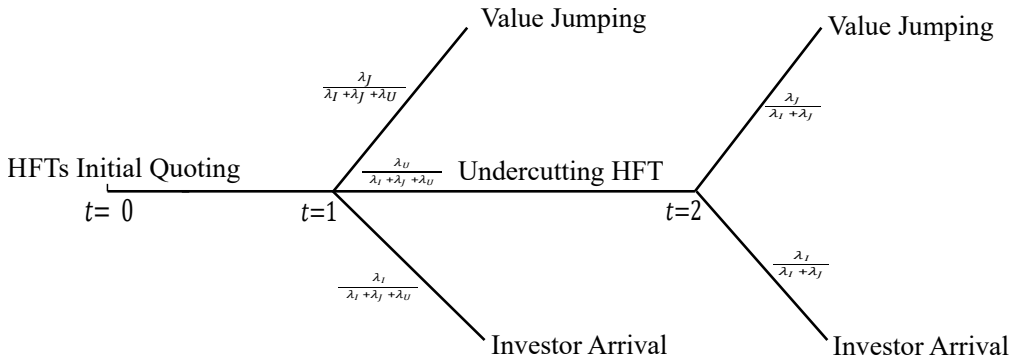

Figure 1: Event Time line of the Baseline Model

send price information between exchange i and exchange j by ϵ_{ij}, where $\epsilon_{ij} = \epsilon_{ji}$. A smaller ϵ_{ij} indicates faster connection speeds between exchanges. Currently, ϵ_{ij} is about 350 microseconds in U.S. equity market.

Figure 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 ϵ_{ij} units of time for this information to arrive at exchange j. Since exchange j needs δ_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 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

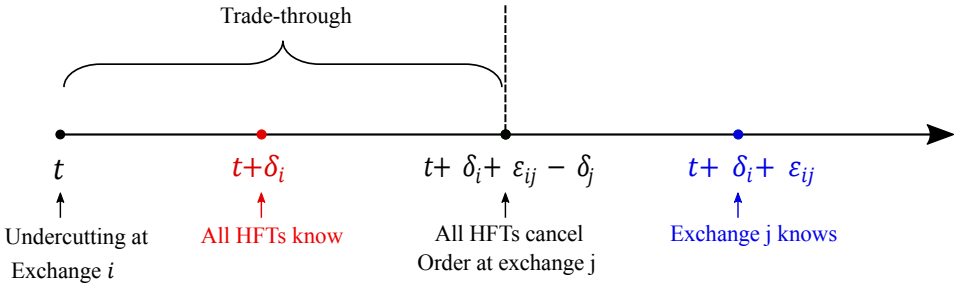

Figure 2: Potential Trade-Through Time Window

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 ζ 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 (list 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 η in Figure 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.

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.

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 $i, j \in \{1, 2, \cdots, M\}$. 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 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 $v_0 + s^*/2$ and $v_0 - s^*/2$. 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 $v_0 + \frac{s}{2}$ and limit buy orders at $v_0 - \frac{s}{2}$ 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]

$$\pi(\frac{s}{2}, X | t_I < t_J, t_U) = \frac{1}{X}\frac{s}{2} \tag{1}$$

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:

$$\pi(\frac{s}{2}, X | t_J < t_I, t_U) = -(\sigma - \frac{s}{2}) \tag{2}$$

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 δ units of time after undercutting HFT's arrival as shown in Figure 2. She will cancel her own order that has inferior price at this time and have liquidity provision profit:

$$\begin{aligned}\pi(\frac{s}{2}, X | t_U < t_I, t_J; \chi_1 = 1) = \phi(\delta)[\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{X}\frac{1}{2}\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] + \\ [1 - \phi(\delta)][\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{X}\frac{1}{2}\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}\frac{1}{2}(\sigma - \frac{s}{2})]\end{aligned} \tag{3}$$

Where $\phi(\delta) = 1 - e^{-(\lambda_I + \lambda_J)\delta}$ is the probability that either an investor or signal jumping arrives within the δ 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

[9] To simply exposition, in all the remaining analysis $t_I < t_J, t_U$ means $t_I < t_J$ and $t_I < t_U$. Other similar notations have the same meaning.

in the right hand side of equation (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 δ 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 δ 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 (3).

If the undercutting HFT arrives first but she does not send her price-improving order to exchange 1, then the liquidity-providing HFT on exchange 1 will cancel her limit order that has inferior price than undercutting HFT's order ϵ units of time after undercutting HFT's arrival as shown in Figure 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.[10] In this case, her profit from liquidity provision is similar:

$$\pi(\frac{s}{2}, X | t_U < t_I, t_J; \chi_1 = 0) = \phi(\epsilon)[\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{X}\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] + [1 - \phi(\epsilon)][\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{X}\frac{1}{2}\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}\frac{1}{2}(\sigma - \frac{s}{2})] \quad (4)$$

where $\phi(\epsilon) = 1 - e^{-(\lambda_I + \lambda_J)\epsilon}$. The first term in the right hand side of equation (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 ζ is the time for HFTs to send an information from one exchange to another exchange as drawn in Figure 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.

[10] 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 ϵ units of time after the undercutting HFT's arrival. This can be seen clearly from Figure 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 $v_0 + s/2$ has no chance to trade with an investor. As a result, liquidity-providing HFT at exchange 1 will cancel her limit sell order at $t_U + \epsilon$.

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

$$\pi(\frac{s}{2}, X) = \frac{\lambda_I}{\Sigma\lambda}\pi(\frac{s}{2}, X|t_I < t_J, t_U) + \frac{\lambda_J}{\Sigma\lambda}\pi(\frac{s}{2}, X|t_J < t_I, t_U) + \frac{\lambda_U}{\Sigma\lambda}\frac{1}{X}\pi(\frac{s}{2}, X|t_U < t_I, t_J; \chi_1 = 1) + \frac{\lambda_U}{\Sigma\lambda}\frac{X-1}{X}\pi(\frac{s}{2}, X|t_U < t_I, t_J; \chi_1 = 0) \quad (5)$$

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:*

$$s^*(\delta, \epsilon) = \begin{cases} d; & if\ \frac{\lambda_J}{\lambda_I+\lambda_J} \leq \frac{d}{2\sigma} \\ \min\{s|v_0 \pm \frac{s}{2} \in \mathscr{P}, \pi(\frac{s}{2}, 1) \geq 0\}; & if\ \frac{\lambda_J}{\lambda_I+\lambda_J} > \frac{d}{2\sigma} \end{cases} \quad (6)$$

where $\mathscr{P}$ is the available price grids set and $\pi(\frac{s}{2}, 1)$ is defined as in equation (5).

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

$$M^*(\delta, \epsilon) = \begin{cases} M; & if\ \frac{s^*}{2} \geq \sigma \\ \max\{X|1 \leq X \leq M, \pi(\frac{s}{2}^*, X) \geq 0\}; & if\ \frac{s^*}{2} < \sigma \end{cases} \quad (7)$$

Where the half bid-ask spread $s^*/2$ is determined in (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 (7). Although, the definition for (5) need to be adjusted for $s/2 > \sigma$ because of no sniping. But since (5) is positive for all $s/2 > \sigma$ and is strictly increasing in s, the equilibrium spread can be uniquely determined in (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 λ_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* δ *and is independent of* ϵ*;*

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

(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 δ. 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 δ and connection speed ϵ. 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:

$$\begin{aligned} TC(\delta,\epsilon) = \frac{\lambda_I}{\Sigma\lambda}\frac{s^*}{2} + \frac{\lambda_J}{\Sigma\lambda}TC(\delta,\epsilon) + \frac{\lambda_U}{\Sigma\lambda}Prob(t_I \geq t_J | t_I, t_J > t_U)TC(\delta,\epsilon) + \\ \frac{\lambda_U}{\Sigma\lambda}Prob(t_I < t_J, t_I \leq t_U + \epsilon | t_I, t_J > t_U)[\frac{1}{2}\frac{s^*}{2} + \frac{1}{2}(\frac{M^*-1}{M^*}\frac{s^*}{2} + \frac{1}{M^*}(\frac{s^*}{2} - d))] + \\ \frac{\lambda_U}{\Sigma\lambda}Prob(t_I < t_J, t_I > t_U + \epsilon | t_I, t_J > t_U)[\frac{1}{2}\frac{s^*}{2} + \frac{1}{2}(\frac{s^*}{2} - d)] \quad (8) \end{aligned}$$

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. $Prob(t_I \geq t_J | t_I, t_J > t_U)$ 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 $\frac{\lambda_I}{\Sigma\lambda}$ 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 (8).

With probability $\frac{\lambda_J}{\Sigma\lambda}$ 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 (8).

With probability $\frac{\lambda_U}{\Sigma\lambda}$ 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 (8). If the buyer arrives before the

risky asset's common value jumps and is within the ϵ 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 (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 vise 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 (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:

$$W(\delta, \epsilon) = \bar{\theta} - TC(\delta, \epsilon) \tag{9}$$

By solving equation (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:*

$$W(\delta, \epsilon) = \bar{\theta} - \{\frac{\lambda_I + \lambda_J}{\Sigma\lambda}\frac{s^*}{2} + \frac{\lambda_U}{\Sigma\lambda}[\phi(\epsilon)A + (1 - \phi(\epsilon))B]\} \tag{10}$$

where $A = \frac{1}{2}\frac{s^*}{2} + \frac{1}{2}[\frac{M^*-1}{M^*}\frac{s^*}{2} + \frac{1}{M^*}(\frac{s^*}{2} - d)]$, $B = \frac{1}{2}\frac{s^*}{2} + \frac{1}{2}(\frac{s^*}{2} - d)$, 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 ϵ if $M^* \geq 2$ and is independent of ϵ 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 δ and connection speed ϵ. 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:*

$$Q^*(\delta, \epsilon) = \lambda_I \frac{1}{M} + \lambda_J \frac{M^*(\delta, \epsilon)}{M} + \frac{\lambda_U \lambda_J}{2\Sigma\lambda} \phi(\delta) \frac{1}{M} - \frac{1 - \phi(\epsilon)}{2} \frac{\lambda_U \lambda_J}{\Sigma\lambda} \frac{M^*(\delta, \epsilon) - 1}{M} \tag{11}$$

where $M^*(\delta, \epsilon)$ is the equilibrium depth as determined in equation (7) and $\phi(\epsilon) = 1 - e^{-(\lambda_I + \lambda_J)\epsilon}$;

(ii) *$Q^*(\delta, \epsilon)$ is strictly increasing in ϵ 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 (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 ϵ 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 (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 (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 ϵ. 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.

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 ϵ in most notation for concreteness.

Exchange Trading Volume Under Speed Heterogeneity. Suppose K exchanges have the same fast order processing speed δ_F and other $M - K$ exchanges have the same slow order processing speed δ_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 δ_S or δ_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 ϵ as given in Proposition 1. I drop it for easy exposition). If HFTs provide liquidity on slow exchanges, the smallest possible spread would be $s^*(\delta_S)$.[11] 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, undercutting 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 undercutting 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 δ_F, then it is always optimal for an undercutting 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 undercutting 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 undercutting 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, undercutting HFTs strictly prefer faster exchange to submit her price-improving order. As a result, it is not reasonable to still assume that investors will randomly choose an exchange with better price to trade. In reality, some investors may consider the

[11] 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 ((6)).

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 γ 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.[12] This dichotomization 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 (5), which is HFT's liquidity provision profit when all exchanges have the same order processing speed. Still denote t_I, t_J and t_U as the arriving time of the investor, the risky asset's common value jumping and undercutting HFT. We can also use the same indicator function defined in definition 1: $\chi_1 = 1$

[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.

if undercutting HFT submits her order to exchange 1. Otherwise, $\chi_1 = 0$. If the investor arrives first, liquidity-providing HFT on exchange 1 earns profit:

$$\pi_F(\frac{s}{2}, X_F, X_S | t_I < t_J, t_U) = (\frac{\gamma}{X_F} + \frac{1-\gamma}{X_F + X_S})\frac{s}{2} \tag{12}$$

This is because if the investor is sophisticated (with probability γ), she will choose one among those X_F exchanges to trade because she knows that undercutting 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:

$$\pi_F(\frac{s}{2}, X_F, X_S | t_J < t_I, t_U) = -(\sigma - \frac{s}{2}) \tag{13}$$

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

$$\begin{aligned}\pi_F(\frac{s}{2}, X_F, X_S | t_U < t_I, t_J; \chi_1 = 1) = \phi(\delta_F)[\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}(\frac{\gamma}{X_F} + \frac{1-\gamma}{X_F + X_S})\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] + \\ [1 - \phi(\delta_F)][\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}(\frac{\gamma}{X_F} + \frac{1-\gamma}{X_F + X_S})\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}\frac{1}{2}(\sigma - \frac{s}{2})]\end{aligned} \tag{14}$$

The above profits could be explained in a similar way as in equation (3). Suppose the undercutting HFT is a seller. Thus, she is willing to sell at $v_0 + s/2 - d$. 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 undercutting 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 (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 (14). Similarly, if the undercutting HFT submits her order to other fast exchanges, liquidity-providing

HFT on exchange 1 has profits:

$$\pi_F(\frac{s}{2}, X_F, X_S | t_U < t_I, t_J; \chi_1 = 0) = \phi(\epsilon)[\frac{\lambda_I}{\lambda_I + \lambda_J}(\frac{\gamma}{X_F} + \frac{1-\gamma}{X_F + X_S})\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] + [1 - \phi(\epsilon)][\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}(\frac{\gamma}{X_F} + \frac{1-\gamma}{X_F + X_S})\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}\frac{1}{2}(\sigma - \frac{s}{2})] \quad (15)$$

If undercutting HFT submits her order to other fast exchanges, liquidity-providing HFT on exchange 1 will cancel their being undercut limit order at $t_U + \epsilon$ (see Figure 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:

$$\pi_F(\frac{s}{2}, X_F, X_S) = \frac{\lambda_I}{\Sigma\lambda}\pi_F(\frac{s}{2}, X_F, X_S | t_I < t_J, t_U) + \frac{\lambda_J}{\Sigma\lambda}\pi_F(\frac{s}{2}, X_F, X_S | t_J < t_I, t_U) + \frac{\lambda_U}{\Sigma\lambda}\frac{1}{X_F}\pi_F(\frac{s}{2}, X_F, X_S | t_U < t_I, t_J; \chi_1 = 1) + \frac{\lambda_U}{\Sigma\lambda}\frac{X_F - 1}{X_F}\pi_F(\frac{s}{2}, X_F, X_S | t_U < t_I, t_J; \chi_1 = 0) \quad (16)$$

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:

$$\pi_S(\frac{s}{2}, X_F, X_S) = \frac{\lambda_I}{\Sigma\lambda}\frac{1-\gamma}{X_F + X_S}\frac{s}{2} - \frac{\lambda_J}{\Sigma\lambda}(\sigma - \frac{s}{2}) + \frac{\lambda_U}{\Sigma\lambda}\phi(\delta_F + \epsilon - \delta_S)[\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1-\gamma}{X_F + X_S}\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}(\sigma - \frac{s}{2})] + \frac{\lambda_U}{\Sigma\lambda}[1 - \phi(\delta_F + \epsilon - \delta_S)][\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}\frac{1-\gamma}{X_F + X_S}\frac{s}{2} - \frac{\lambda_J}{\lambda_I + \lambda_J}\frac{1}{2}(\sigma - \frac{s}{2})] \quad (17)$$

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 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 (17).

Note that when $X_S = 0$, $\pi_F(\frac{s}{2}, X_F, 0) = \pi(\frac{s}{2}, X_F|\delta = \delta_F)$. The later is the liquidity provision profit in equation (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 3.1 would apply. If HFTs only provide liquidity on one exchange, then their liquidity provision profit on fast exchange is always lager than on slow exchange for the same bid-ask spread ($\pi(\frac{s}{2}, 1|\delta = \delta_F) > \pi(\frac{s}{2}, 1|\delta = \delta_S)$, see equation (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, $\pi_F(\frac{s}{2}, X_F, X_S)$ is not always larger than $\pi_S(\frac{s}{2}, X_F, X_S)$. 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 γ is large enough it is possible that $\pi_F(\frac{s}{2}, X_F, X_S) \geq \pi_S(\frac{s}{2}, X_F, X_S)$ 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 γ is large.

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

$$M_S^*(K) = \max\{X_S|\pi_S(\frac{s^*(\delta_F)}{2}, K, X_S) \geq 0, 1 \leq X_S \leq M - K\} \tag{18}$$

Note that $M^*(\delta_F)$ is the equilibrium depth when all exchanges have the same fast order processing speed (determined in equation (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 (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 (18) while the depth on fast exchanges is always K.

In the Appendix B, I construct the equilibrium when $\gamma < \bar{\gamma}$. For small γ 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 γ could be very high. For conciseness, in the remaining 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

[13] When $\gamma \geq \bar{\gamma}$, 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.

[14] Miller and Shorter (2016) estimates that HFTs account around 55% trading volume in U.S. equity market. HFTs are just a subset of algorithm traders.

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:

$$Q_F^*(K) = \lambda_I \frac{1}{K} + \lambda_J \frac{M_F^*(K)}{K} + \frac{\lambda_U \lambda_J}{2\Sigma\lambda} \phi(\delta_F) \frac{1}{K} - \frac{1-\phi(\epsilon)}{2} \frac{\lambda_U \lambda_J}{\Sigma\lambda} \frac{M_F^*(K) - 1}{K} \tag{19}$$

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 (11) by simply replacing total number of exchanges M with K, equilibrium depth M^* with $M_F^*(K)$ and order processing speed δ with δ_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 δ_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 δ_F and $M - K$ slow exchanges with order processing speed δ_S,*

where $1 \leq K \leq M-1$ *and* $\delta_F < \delta_S$. *When* $\gamma \geq \bar{\gamma}$ *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* (19) *if* $M^*(\delta_F) \leq K$ *or* $M^*(\delta_F) > K$ *and* $\pi_S(s^*(\delta_F)/2, K, 1) < 0$. *Otherwise,*

$$Q_F^*(K) = \lambda_I[\frac{\gamma}{K} + \frac{1-\gamma}{M^*(K)}] + \lambda_J + \frac{\lambda_U\lambda_J[\phi(\delta_F) + (1-K)(1-\phi(\epsilon))]}{2K\Sigma\lambda} + \frac{\lambda_U\lambda_I[1-\phi(\epsilon')](1-\gamma)M_S^*(K)}{2KM^*(K)\Sigma\lambda} \quad (20)$$

(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,*

$$Q_S^*(K) = \frac{M_S^*(K)}{M-K}[\frac{\lambda_I(1-\gamma)}{M^*(K)} + \lambda_J]\{1 - \frac{\lambda_U}{2\Sigma\lambda}[1-\phi(\epsilon')]\} \quad (21)$$

where $\phi(\delta_F) = 1 - e^{-(\lambda_I+\lambda_J)\delta_F}$, $\phi(\epsilon') = \phi(\delta_F + \epsilon - \delta_S) = 1 - e^{-(\lambda_I+\lambda_J)(\delta_F+\epsilon-\delta_S)}$ *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 δ_S, denote $Q^*(\delta_S)$ as each exchange's per unit time trading volume which is determined in equation (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 (19) or (20). All remaining $M-1$ exchanges with slow order processing speed δ_S have expected per unit time trading volume $Q_S^*(1)$ determined in equation (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 large 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 δ_F units time after undercutting HFT's arrival. If δ_F is too smaller than δ_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 δ_F is too smaller than δ_S. But as long as δ_F is close to δ_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 $Q_F^*(1) > Q^*(\delta_S)$.[15]

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, δ_F is not too smaller than δ_S when trading is already fast enough. It is natural to think that exchange may not invest in speed technology anymore because it can not significantly enhance its trading speed. Surprisingly, Corollary 2 points out that exactly when δ_F is not too smaller than δ_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

[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 $\Delta_{speed} \leq \delta_S - \phi^{-1}[\frac{\phi(\delta_S)-(M-1)(1-\phi(\epsilon))}{M}]$ where $\Delta_{speed} = \delta_S - \delta_F$.

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 δ_S to δ_F at per unit time cost C_{speed}, where $\delta_F < \delta_S$.[16] 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, Yao, and Ye (2017) to define $\bar{f}_m$ and $\bar{f}_t$ as maker and taker fee. The total maker taker fee is defined as $\bar{f} = \bar{f}_m + \bar{f}_t > 0$.[17] The rebate is paid only when transaction occurs, and exchanges earn $\bar{f}$ 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 can also attract more listings due to the positive externalities of liquidity. As a result, it is safe to conclude that an exchange's

[16] I use the same notations as in BCS for the cost of speed investment. While in BCS C_{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.

[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 $\bar{f} = \$0.001$, $f_m = -\$0.002$ and $f_t = \$0.003$. Note that, $\bar{f}_m$ 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.

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 $\bar{f}_m = 0$ and $\bar{f} = \bar{f}_t > 0$, 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:

$$\pi_F(K) = (\bar{f}_m + \bar{f}_t)Q_F^*(K) - C_{speed}; \qquad \pi_S(K) = (\bar{f}_m + \bar{f}_t)Q_S^*(K)$$

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 $\gamma \geq \bar{\gamma}$, $s^*(\delta_S)/2 + \bar{f}_t < \sigma$ and $\phi(\delta_S) \leq M\phi(\delta_F) + (M-1)[1-\phi(\epsilon)]$, then for given $\bar{f} = \bar{f}_t > 0$:*
(i) *If $\frac{C_{speed}}{\bar{f}} < min\{Q_F^*(1) - Q^*(\delta_S), Q^*(\delta_F) - Q_S^*(1)\}$, investing in the fast speed technology is a dominant strategy for all exchanges;*
(ii) *If $\frac{C_{speed}}{\bar{f}} > Q^*(\delta_F) - Q^*(\delta_S)$, 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 (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

[18] Investor's welfare defined in equation (10) is the one without maker-taker fee. Since investors always take liquidity in my model, under maker-taker fee investor's welfare is $W(Buy|\delta_i = \delta_S, \epsilon_{ij} = \epsilon) - \bar{f}_t$.

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, Yao, and Ye (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.

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.

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 program 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 Jun 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 IEXs 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 605 data.

SEC 605 data is well-known for comparing execution quality such as quoted spread and effective spread.[19] 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.

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 IEXs 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:

$$y_{it} = \beta(Post_t \times Pilot_i) + X'_{it}\delta + \text{Stock FE}_i + \text{Time FE}_t + \epsilon_{it} \tag{22}$$

[19] 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, Shkilko, Van Ness, and Van Ness (2008) for their study on market quality using the SEC 605 filing.

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

$$y_{it} = \beta(Post_t \times NonBinding_i) + X'_{it}\delta + \text{Stock FE}_i + \text{Time FE}_t + \epsilon_{it} \tag{23}$$

The dependent variable y_{it} equals to IEXs 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.

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 increase 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.

[20] IEXs 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.

Table 1: Summary Statistics for Tick Size Pilot Test

VARIABLES	N	Mean	SD	Min	Max
Panel A: Control Group (1168 Stocks)					
IEX Market Share	94,414	1.78	3.00	0	1
Closing Price	94,398	23.82	28.22	0.365	485.6
Share Turnover	94,398	231,171	556,946	1	51,220,000
Market Cap	94,398	720.5	759.9	3.861	4129
Panel B: Treatment Group 1 (393 Stocks)					
IEX Market Share	31,765	1.89	2.91	0	1
Closing Price	31,760	23.90	23.71	1.050	172.6
Share Turnover	31,760	237,763	560,753	7	50,260,000
Market Cap	31,760	711.2	756.1	4.048	3776
Panel C: Treatment Group 2 (396 Stocks)					
IEX Market Share	31,602	1.90	2.96	0	1
Closing Price	31,599	23.56	23.26	1.250	203.8
Share Turnover	31,599	220,863	529,772	1	47,750,000
Market Cap	31,599	699.6	729	5.471	4019
Panel D: Treatment Group 3 (390 Stocks)					
IEX Market Share	31,470	1.95	2.82	0	1
Closing Price	31,468	24.77	39.19	1.100	542.0
Share Turnover	31,468	247,085	622,077	2	34,880,000
Market Cap	31,468	732	772.5	5.276	4127

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 2: Impact of Tick Size on IEX's Market Share of Trading Volume

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

Note: the above table reports estimations results from $y_{it} = \beta(Post_t \times Pilot_i) + X'_{it}\delta +$ Stock FE_i + 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. X'_{it} 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 3: Impact of Switching to Public Exchange on IEX's Market Share of Total Trading Volume

VARIABLES	(1) IEX Market Share	(2) IEX Market Share	(3) IEX Market Share
NonBinding×Post	-0.16***	-0.17***	-0.17***
	(0.02)	(0.02)	(0.02)
ln (Share Turnover)		-0.25***	0.36***
		(0.06)	(0.07)
Inverse of Share Price		-0.13***	-0.07**
		(0.02)	(0.03)
ln (Market Cap)		0.10***	0.04
		(0.002)	(0.02)
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 $y_{it} = \beta(NonBinding_i \times Post_t) + X'_{it}\delta +$ Stock $\text{FE}_i +$ Time $\text{FE}_t + \epsilon_{it}$, 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. $Post_t$ and $NonBinding_i$ are two dummy variables. $Post_t = 1$ for September, October and November of 2016, when IEX is a public exchange. $NonBinding_i = 1$ 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.

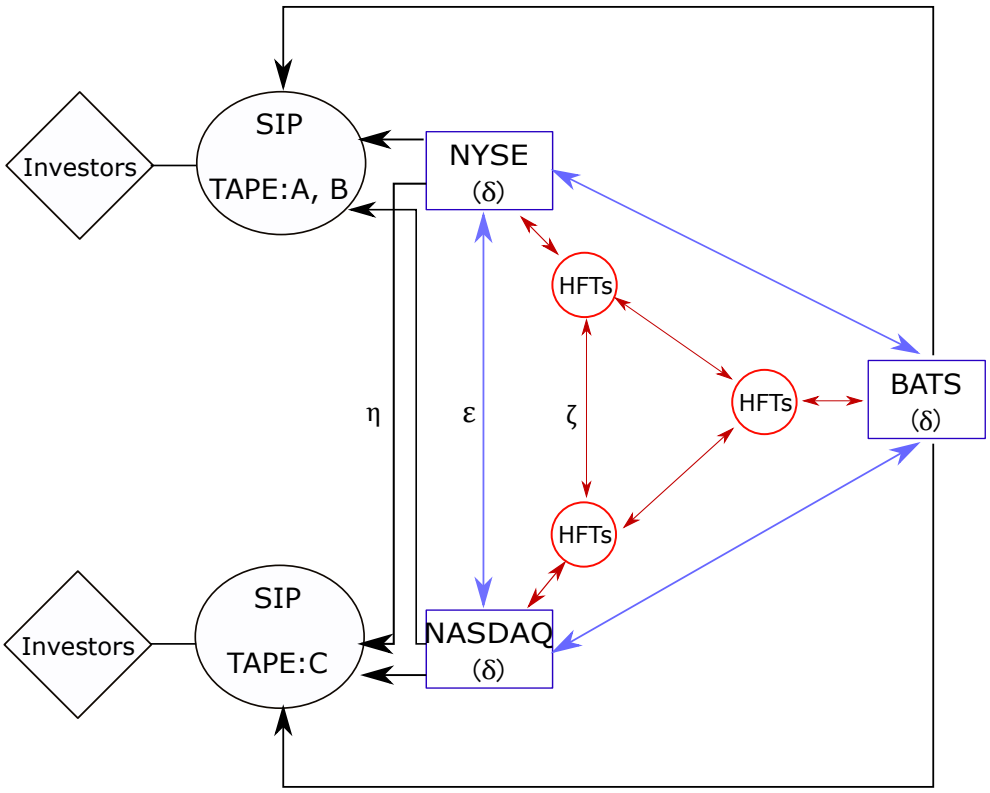

Figure 3: Latency Map. δ denotes exchange's order processing speed. ζ is the time it takes for a HFT to send information from one exchange to another exchange. ϵ is the time it takes for an exchange to send its order updating information or route orders to other exchange. η is the time it takes for an investor to receive any updates from exchanges through securities information processor (SIP). These latencies dependent 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 δ, ζ, ϵ and η are around 50, 100, 300 and 1000 microseconds.

Appendix A Proofs

A.1 Proof of Proposition 1

I first show that when $s/2 < \sigma$, $\pi(\frac{s}{2}, X)$ is strictly decreasing in X. Insert equation (1), (2), (3) and (4) to (5), we have:

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

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

$$\pi(\frac{s}{2}, \frac{1}{x}) = \frac{\lambda_I}{\Sigma\lambda}x\frac{s}{2} - \frac{\lambda_J}{\Sigma\lambda}(\sigma - \frac{s}{2}) + \frac{\lambda_U}{\Sigma\lambda}x\phi(\delta)(\frac{1}{2}xC - D) + \frac{\lambda_U}{\Sigma\lambda}x[1 - \phi(\delta)](\frac{1}{2}xC - \frac{1}{2}D) + \frac{\lambda_U}{\Sigma\lambda}(1-x)\phi(\epsilon)\times (xC - D) + \frac{\lambda_U}{\Sigma\lambda}(1-x)[1 - \phi(\epsilon)](\frac{1}{2}xC - \frac{1}{2}D) \quad (A.2)$$

Now we take derivative with respect to x:

$$\frac{d\pi(\frac{s}{2}, \frac{1}{x})}{dx} = \frac{\lambda_I}{\Sigma\lambda}\frac{s}{2} + \frac{\lambda_U}{\Sigma\lambda}\phi(\delta)(xC - D) + \frac{\lambda_U}{\Sigma\lambda}[1 - \phi(\delta)](xC - \frac{1}{2}D) + \frac{\lambda_U}{\Sigma\lambda}\phi(\epsilon)(C + D - 2xC) + \frac{\lambda_U}{\Sigma\lambda}[1 - \phi(\epsilon)]\times (\frac{1}{2}C + \frac{1}{2}D - xC) = \frac{\lambda_I}{\Sigma\lambda}\frac{s}{2} + \frac{\lambda_U}{\Sigma\lambda}[\phi(\epsilon) - \phi(\delta)]\frac{1}{2}B + \frac{\lambda_U}{\Sigma\lambda}C[\frac{1}{2} + \frac{1}{2}\phi(\epsilon) - \phi(\epsilon)x] \geq \frac{\lambda_I}{\Sigma\lambda}\frac{s}{2} + \frac{\lambda_U}{\Sigma\lambda}[\phi(\epsilon) - \phi(\delta)]\times \frac{1}{2}D + \frac{\lambda_U}{\Sigma\lambda}C[\frac{1}{2}\phi(\epsilon) + \frac{1}{2}\phi(\epsilon) - \phi(\epsilon)x] = \frac{\lambda_I}{\Sigma\lambda}\frac{s}{2} + \frac{\lambda_U}{\Sigma\lambda}[\phi(\epsilon) - \phi(\delta)]\frac{1}{2}D + \frac{\lambda_U}{\Sigma\lambda}C\phi(\epsilon)(1-x) > 0 \quad (A.3)$$

since $x = \frac{1}{X} \leq 1$ and $\delta < \epsilon$ (under assumption that $\delta + \zeta < \epsilon$). Therefore, the profit function $\pi(\frac{s}{2}, \frac{1}{x})$ 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 $\pi(\frac{s}{2}, X)$ is strictly decreasing in X, the equilibrium bid-ask spread s^* is determined when $X = 1$. This is the result in equation (6). Intuitively, if $\pi(\frac{s}{2}, 1)$ 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 $v_0 + d/2$ and limit buy at $v_0 - d/2$ 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 $\frac{\lambda_J}{\lambda_I+\lambda_J} \leq \frac{d}{2\sigma}$. 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 (5) when $X = 1$. So the equilibrium bid-ask spread would be the minimum available price in the price grids such that $\pi(\frac{s}{2}, 1)$ is nonnegative. This proves the result in equation (6).

If equilibrium half spread $s^*/2$ is larger or equal to the risky asset's common value jumping size σ, 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 $v_0 + s^*/2 - d$ to one exchange. If she sends her order to one among those M^* exchanges, her payoff is:

$$\frac{1}{2}\{\phi(\epsilon)[\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{M^*}(\frac{s^*}{2}-d)-\frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma-\frac{s^*}{2}+d)]+[1-\phi(\epsilon)][\frac{\lambda_I}{\lambda_I+\lambda_J}(\frac{s^*}{2}-d)-\frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma-\frac{s^*}{2}+d)]\} \tag{A.4}$$

If she submits her pricing improving order to one exchange among those $M - M^*$ exchanges, her payoff would be:

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

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.2 Proof of Corollary 1

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

$$\frac{d\pi(\frac{s}{2},1)}{d\delta} = -\frac{1}{2}\phi'(\delta)\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}(\sigma-\frac{s}{2}) < 0 \tag{A.6}$$

where $\phi'(\delta) = \frac{1}{\lambda_I+\lambda_J}e^{-\frac{1}{\lambda_I+\lambda_J}\delta} > 0$. Since $\pi(\frac{s}{2}, 1)$ is strictly increasing in $\frac{s}{2}$, equation (6) implies that s^* is weakly increasing in δ. Since $\pi(\frac{s}{2}, 1)$ is independent of ϵ so as the equilibrium bid-ask spread s^*.

(ii) Take derivative of $\pi(\frac{s}{2}^*, X)$ with respect to ϵ, we have:

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

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 $\pi(\frac{s}{2}^*, X) < 0$ for all $X \geq 1$. So $\frac{d\pi(\frac{s}{2}^*,X)}{d\epsilon} > 0$. Equation (7) implies that M^* is weakly increasing in ϵ.

(iii) From equation (A.1), we have $\pi(\frac{s}{2}^*, X|\delta = \delta_F) - \pi(\frac{s}{2}^*, X|\delta = \delta_S) =$

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

According to equation (7), we have $M^*(\delta_F) \geq M^*(\delta_S)$.

A.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 $Prob(t_I \geq t_J|t_I, t_J > t_U) = Prob(t_1 \geq t_2|t_1, t_2 > 0) = Prob(t_1 \geq t_2)$ 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 $Prob(t_1 \geq t_2) = \frac{\lambda_J}{\lambda_I+\lambda_J}$. Now we show that:

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

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 (8), we would have:

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

Denoting $A = \frac{1}{2}\frac{s}{2}^* + \frac{1}{2}[\frac{M^*-1}{M^*}\frac{s}{2}^* + \frac{1}{M^*}(\frac{s}{2}^* - d)]$ and $B = \frac{1}{2}\frac{s}{2}^* + \frac{1}{2}(\frac{s}{2}^* - d)$ 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 ϵ because M^* is weakly increasing in ϵ according to Corollary 1. Also from Corollary 1, $M^*(\delta_F) \geq M^*(\delta_S)$. Thus $W(Buy|\delta_F = \delta_1, \epsilon_{ij} = \epsilon) \leq W(Buy|\delta_i = \delta_S, \epsilon_{ij} = \epsilon)$.

A.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 $\frac{1}{\Sigma\lambda}$ units time for one of these events to occur. With probability $\frac{\lambda_I}{\Sigma\lambda}$ 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.[21] Thus the ex ante expected trading volume for each exchange is $1/M$. With probability $\frac{\lambda_J}{\Sigma\lambda}$ 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 $\frac{\lambda_U}{\Sigma\lambda}$ 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 $Prob(t_J < t_I, t_J \leq t_U + \delta | t_I, t_J > t_U) = \frac{\lambda_J}{\lambda_I+\lambda_J}\phi(\delta)$ 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 risky asset's common value jumps up by σ, 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 comon value jumps conditional on it happens before investor arrival and liquidity provision HFTs have not canceled their orders. This is denoted as $E(t_J - t_U | t_U < t_J < t_I, t_J \leq t_U + \delta)$. As in the proof of Proposition 2, we define $t_1 = t_I - t_U$ and $t_2 = t_J - t_U$, thus $t_1 \sim Exp(\lambda_I)$ and $t_2 \sim Exp(\lambda_J)$. In order to calculate $E(t_J - t_U | t_U < t_J < t_I, t_J \leq t_U + \delta)$ we first calculate $E(t_I - t_U | t_U < t_I < t_J, t_I \leq t_U + \delta)$. In this way we can directly use the result in equation (A.9). Specifically,

$$E(t_I - t_U | t_U < t_I < t_J, t_I \leq t_U + \delta) = E(t_1 | t_1 < t_2, t_1 \leq \delta) =$$

[21] 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.

$$\int_0^{\delta}\int_{t_1}^{\infty}\frac{1}{Prob(t_1<t_2,t_1\leq\delta)}t_1\lambda_I e^{-\lambda_I t_1}\lambda_J e^{-\lambda_J t_2}dt_2dt_1=\int_0^{\delta}\int_{t_1}^{\infty}\frac{t_1\lambda_I e^{-\lambda_I t_1}\lambda_J e^{-\lambda_J t_2}}{\frac{\lambda_I}{\lambda_I+\lambda_J}[1-e^{-(\lambda_I+\lambda_J)\delta}]}dt_2dt_1$$

$$=\frac{\lambda_I+\lambda_J}{\lambda_I[1-e^{-(\lambda_I+\lambda_J)\delta}]}\int_0^{\delta}t_1\lambda_I e^{-\lambda_I t_1}\int_{t_1}^{\infty}\lambda_J e^{-\lambda_J t_2}dt_2dt_1=\frac{\lambda_I+\lambda_J}{\lambda_I[1-e^{-(\lambda_I+\lambda_J)\delta}]}\int_0^{\delta}t_1\lambda_I e^{-\lambda_I t_1}e^{-\lambda_J t_1}dt_1$$

$$=\frac{\lambda_I+\lambda_J}{\lambda_I[1-e^{-(\lambda_I+\lambda_J)\delta}]}[-\frac{\lambda_I}{\lambda_I+\lambda_J}t_1e^{-(\lambda_I+\lambda_J)t_1}|_0^{\delta}+\int_0^{\delta}e^{-(\lambda_I+\lambda_J)t_1}\frac{\lambda_I}{\lambda_I+\lambda_J}dt_1]=\frac{\lambda_I+\lambda_J}{\lambda_I[1-e^{-(\lambda_I+\lambda_J)\delta}]}\times$$

$$[-\frac{\lambda_I}{\lambda_I+\lambda_J}\delta e^{-(\lambda_I+\lambda_J)\delta}-\frac{\lambda_I}{(\lambda_I+\lambda_J)^2}e^{-(\lambda_I+\lambda_J)\delta}+\frac{\lambda_I}{(\lambda_I+\lambda_J)^2}]=\frac{1}{1-e^{-(\lambda_I+\lambda_J)\delta}}[-\delta e^{-(\lambda_I+\lambda_J)\delta}$$

$$-\frac{1}{\lambda_I+\lambda_J}e^{-(\lambda_I+\lambda_J)\delta}+\frac{1}{\lambda_I+\lambda_J}]=\frac{1}{\lambda_I+\lambda_J}-\frac{e^{-(\lambda_I+\lambda_J)\delta}}{1-e^{-(\lambda_I+\lambda_J)\delta}}\delta=\frac{1}{\lambda_I+\lambda_J}-\frac{1-\phi(\delta)}{\phi(\delta)}\delta \quad (A.11)$$

Therefore,

$$E(t_J-t_U|t_U<t_J<t_I,t_J\leq t_U+\delta)=E(t_2|t_2<t_1,t_2\leq\delta)=\frac{1}{\lambda_I+\lambda_J}-\frac{1-\phi(\delta)}{\phi(\delta)}\delta \quad (A.12)$$

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

With $Prob(t_J<t_I,t_U+\delta<t_J\leq t_U+\epsilon|t_I,t_J>t_U)$ the risky asset's common value jumps after δ units of time but within ϵ 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 $\frac{M^*}{M}$. We also need to calculate the probability of this case similarly to (A.9):

$$Prob(t_J<t_I,t_U+\delta<t_J\leq t_U+\epsilon|t_I,t_J>t_U)=Prob(t_2<t_1,\delta<t_2\leq\epsilon)=$$

$$\int_{\delta}^{\epsilon}\int_{t_2}^{\infty}\lambda_I e^{-\lambda_I t_1}\lambda_J e^{-\lambda_J t_2}dt_1dt_2=\int_{\delta}^{\epsilon}e^{-\lambda_I t_2}\lambda_J e^{-\lambda_J t_2}dt_2=\frac{\lambda_J}{\lambda_I+\lambda_J}\int_{\delta}^{\epsilon}(\lambda_I+\lambda_J)e^{-(\lambda_I+\lambda_J)t_2}dt_2$$

$$=\frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon)-\phi(\delta)] \quad (A.13)$$

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:

$$E(t_J-t_U|t_J<t_I,t_U+\delta<t_J\leq t_U+\epsilon)=E(t_2|t_2<t_1,\delta<t_2\leq\epsilon)=$$

$$\int_{\delta}^{\epsilon}\int_{t_2}^{\infty}\frac{t_2\lambda_I e^{-\lambda_I t_1}\lambda_J e^{-\lambda_J t_2}dt_1dt_2}{\frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon)-\phi(\delta)]}=\frac{\lambda_I+\lambda_J}{\lambda_J[\phi(\epsilon)-\phi(\delta)]}\int_{\delta}^{\epsilon}t_2\lambda_J e^{-\lambda_J t_2}e^{-\lambda_I t_2}dt_2=\frac{\lambda_I+\lambda_J}{\lambda_J[\phi(\epsilon)-\phi(\delta)]}\times$$

$$[-\frac{t_2\lambda_J}{\lambda_I+\lambda_J}e^{-(\lambda_I+\lambda_J)t_2}|_{\delta}^{\epsilon}+\int_{\delta}^{\epsilon}\frac{\lambda_J}{\lambda_I+\lambda_J}e^{-(\lambda_I+\lambda_J)t_2}dt_2]=\frac{\lambda_I+\lambda_J}{\lambda_J[\phi(\epsilon)-\phi(\delta)]}[\frac{\delta\lambda_J}{\lambda_I+\lambda_J}(1-\phi(\delta))-$$

$$\frac{\epsilon\lambda_J}{\lambda_I+\lambda_J}(1-\phi(\epsilon))+\frac{\lambda_J}{(\lambda_I+\lambda_J)^2}(\phi(\epsilon)-\phi(\delta))]=\frac{1}{\lambda_I+\lambda_J}+\epsilon-\frac{\epsilon-\delta}{\phi(\epsilon-\delta)} \quad (A.14)$$

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

Similarly, with $Prob(t_J<t_I,t_J>t_U+\epsilon|t_I,t_J>t_U)=\frac{\lambda_J}{\lambda_I+\lambda_J}[1-\phi(\epsilon)]$ (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 undercutting HFT's limit order is subject to the sniping risk. If the undercutting 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 $(\frac{1}{2}M^*+\frac{1}{2})/M$. 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:

$$\begin{aligned}
&E(t_I-t_U|t_U<t_I<t_J,t_I>t_U+\epsilon)=E(t_1|t_1<t_2,t_1>\epsilon)=\\
&\int_\epsilon^\infty\int_{t_1}^\infty\frac{1}{Prob(t_1<t_2,t_1>\epsilon)}t_1\lambda_Ie^{-\lambda_It_1}\lambda_Je^{-\lambda_Jt_2}dt_2dt_1=\int_\epsilon^\infty\int_{t_1}^\infty\frac{t_1\lambda_Ie^{-\lambda_It_1}\lambda_Je^{-\lambda_Jt_2}}{\frac{\lambda_I}{\lambda_I+\lambda_J}e^{-(\lambda_I+\lambda_J)\epsilon}}dt_2dt_1\\
&=\frac{\lambda_I+\lambda_J}{\lambda_Ie^{-(\lambda_I+\lambda_J)\epsilon}}\int_\epsilon^\infty t_1\lambda_Ie^{-\lambda_It_1}\int_{t_1}^\infty\lambda_Je^{-\lambda_Jt_2}dt_2dt_1=\frac{\lambda_I+\lambda_J}{\lambda_Ie^{-(\lambda_I+\lambda_J)\epsilon}}\int_\epsilon^\infty t_1\lambda_Ie^{-\lambda_It_1}e^{-\lambda_Jt_1}dt_1\\
&=\frac{\lambda_I+\lambda_J}{\lambda_Ie^{-(\lambda_I+\lambda_J)\epsilon}}[-\frac{\lambda_I}{\lambda_I+\lambda_J}t_1e^{-(\lambda_I+\lambda_J)t_1}|_\epsilon^\infty+\int_\epsilon^\infty e^{-(\lambda_I+\lambda_J)t_1}\frac{\lambda_I}{\lambda_I+\lambda_J}dt_1]=\frac{\lambda_I+\lambda_J}{\lambda_Ie^{-(\lambda_I+\lambda_J)\epsilon}}\times\\
&[\frac{\lambda_I}{\lambda_I+\lambda_J}\epsilon e^{-(\lambda_I+\lambda_J)\epsilon}+\frac{\lambda_I}{(\lambda_I+\lambda_J)^2}e^{-(\lambda_I+\lambda_J)\epsilon}]=\frac{1}{e^{-(\lambda_I+\lambda_J)\epsilon}}[\epsilon e^{-(\lambda_I+\lambda_J)\epsilon}\\
&+\frac{1}{\lambda_I+\lambda_J}e^{-(\lambda_I+\lambda_J)\epsilon}]=\frac{1}{\lambda_I+\lambda_J}+\epsilon \quad (A.15)
\end{aligned}$$

Similarly, we also have:

$$E(t_J-t_U|t_U<t_J<t_I,t_J>t_U+\epsilon)=E(t_2|t_2<t_1,t_2>\epsilon)=\frac{1}{\lambda_I+\lambda_J}+\epsilon \quad (A.16)$$

Denoting Q as each exchange's expected per unit time trading volume, we must have:

$$\begin{aligned}
&(\frac{1}{\Sigma\lambda}+\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q=\frac{\lambda_I}{\Sigma\lambda}[\frac{1}{M}+(\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q]+\frac{\lambda_J}{\Sigma\lambda}[\frac{M^*}{M}+(\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}(\frac{1}{M}+\epsilon Q)\\
&+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}\phi(\delta)[\frac{\frac{1}{2}(M^*+1)+\frac{1}{2}M^*}{M}+(\epsilon+\frac{1-\phi(\delta)}{\phi(\delta)}\delta)Q]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon)-\phi(\delta)][\frac{M^*}{M}+\frac{\epsilon-\delta}{\phi(\epsilon-\delta)}Q]\\
&+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[1-\phi(\epsilon)]\frac{\frac{1}{2}M^*+\frac{1}{2}}{M} \quad (A.17)
\end{aligned}$$

which implies that:

$$Q^*=\lambda_I\frac{1}{M}+\lambda_J\frac{M^*}{M}+\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\phi(\delta)\frac{1}{M}-\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M^*-1}{M} \quad (A.18)$$

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 $\dfrac{\lambda_I}{\Sigma\lambda}$ an investor arrives first. In this case it takes the particular exchange $\frac{1}{\Sigma\lambda}$ units time to have $\frac{1}{M}$ trading volume. Then the game will moves to a new game G'. The expected trading volume for the particular exchange in the remaining $\frac{1}{\lambda_I+\lambda_J} + \epsilon$ units of time would be $(\frac{1}{\lambda_I+\lambda_J} + \epsilon)Q$. 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 $Q^*(\delta_F) - Q^*(\delta_S) =$

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

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:

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

which is positive because $\delta_S < \epsilon$.

A.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 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 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 $v_0 + s^*/2 - d$. 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:

$$\begin{aligned}\frac{1}{2}\{\phi(\delta_F+\epsilon-\delta_S)[\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{M^*}(\frac{s^*}{2}-d) - \frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma-\frac{s^*}{2}+d)] + [\phi(\epsilon)-\phi(\delta_F+\epsilon-\delta_S)][\frac{\lambda_I}{\lambda_I+\lambda_J}(\frac{M^*-K^*+1}{M^*})\times \\ (\frac{s^*}{2} - d) - \frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma - \frac{s^*}{2} + d)] + [1-\phi(\epsilon)][\frac{\lambda_I}{\lambda_I+\lambda_J}(\frac{s^*}{2} - d) - \frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma - \frac{s^*}{2} + d)]\}\end{aligned} \tag{A.21}$$

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 2). Liquidity provision HFTs on other $K^* - 1$ fast exchanges will cancel their limit sell orders at $t + \delta_F + \epsilon - \delta_F = t + \epsilon$. 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:

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

(A.21) − (A.22) =

$$\frac{1}{2}\{[\phi(\epsilon)-\phi(\delta_F+\epsilon-\delta_S)]\frac{M-K}{M}+[\phi(\delta_S+\epsilon-\delta_F)-\phi(\epsilon)]\frac{K}{M}\}\frac{\lambda_I}{\lambda_I+\lambda_J}(\frac{s^*}{2}-d)>0 \quad \text{(A.23)}$$

(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 ϵ 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 ϵ. 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:

$$\frac{1}{2}\{\phi(\delta_F+\epsilon-\delta_S)[-\frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma-\frac{s^*}{2}+d)]+[\phi(\epsilon)-\phi(\delta_F+\epsilon-\delta_S)][\frac{\lambda_I}{\lambda_I+\lambda_J}(\frac{M^*-K^*}{M^*})(\frac{s^*}{2}-d)$$
$$-\frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma-\frac{s^*}{2}+d)]+[1-\phi(\epsilon)][\frac{\lambda_I}{\lambda_I+\lambda_J}(\frac{s^*}{2}-d)-\frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma-\frac{s^*}{2}+d)]\} \quad \text{(A.24)}$$

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

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

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

(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 ϵ 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.6 Proof of Proposition 4

Denote $X = X_F + X_S$ and note that $\pi_S(\frac{s}{2}, X_F, X_S)$ only depends on X. Thus, we can define a new function $\pi_S(\frac{s}{2}, X) = \pi_S(\frac{s}{2}, X_F, X_S)$.[22] I will first show that when $0 < s/2 < \sigma$ and γ is large enough, for any $2 \leq X \leq M$ if $\pi_S(\frac{s}{2}, X) \geq 0$, then $\pi_F(\frac{s}{2}, X_F, X - X_F) \geq \pi_S(\frac{s}{2}, X)$ for any $1 \leq X_F \leq X$. Denote $E = \frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{X}\frac{s}{2} > 0$ and $F = \frac{\lambda_J}{\lambda_I+\lambda_J}(\sigma - \frac{s}{2}) > 0$, where $X = X_F + X_S$. From equation (16) we have:

$$\pi_F(\frac{s}{2}, X_F, X - X_F) = \frac{\lambda_I}{\Sigma\lambda}\frac{\gamma}{X_F}\frac{s}{2} + (1-\gamma)\frac{\lambda_I + \lambda_J}{\Sigma\lambda}E - \frac{\lambda_I + \lambda_J}{\Sigma\lambda}F + \frac{\lambda_U}{\Sigma\lambda}\frac{1}{X_F}\phi(\delta_F)[\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}\frac{\gamma}{X_F}\frac{s}{2} + \frac{1-\gamma}{2}E$$
$$- F] + \frac{\lambda_U}{\Sigma\lambda}\frac{1}{X_F}[1 - \phi(\delta_F)][\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}\frac{\gamma}{X_F}\frac{s}{2} + \frac{1-\gamma}{2}E - \frac{F}{2}] + \frac{\lambda_U}{\Sigma\lambda}\frac{X_F - 1}{X_F}\phi(\epsilon)[\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{\gamma}{X_F}\frac{s}{2} + (1-\gamma)E -$$
$$F] + \frac{\lambda_U}{\Sigma\lambda}\frac{X_F - 1}{X_F}[1 - \phi(\epsilon)][\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}\frac{\gamma}{X_F}\frac{s}{2} + \frac{1-\gamma}{2}E - \frac{F}{2}] \quad \text{(A.26)}$$

From equation (17), we have:

$$\pi_S(\frac{s}{2}, X) = \pi_S(\frac{s}{2}, X_F, X - X_F) = (1-\gamma)\frac{\lambda_I + \lambda_J}{\Sigma\lambda}E - \frac{\lambda_I + \lambda_J}{\Sigma\lambda}F + \frac{\lambda_U}{\Sigma\lambda}\phi(\delta_F + \epsilon - \delta_S)[(1-\gamma)E - F] +$$
$$\frac{\lambda_U}{\Sigma\lambda}[1 - \phi(\delta_F + \epsilon - \delta_S)][\frac{(1-\gamma)E}{2} - \frac{F}{2}] \quad \text{(A.27)}$$

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

$$\pi_F(\frac{s}{2}, X_F, X - X_F) - \pi_S(\frac{s}{2}, X) = \frac{\lambda_I}{\Sigma\lambda}\frac{\gamma}{X_F}\frac{s}{2} + \frac{\lambda_U}{\Sigma\lambda}[\phi(\delta_F + \epsilon - \delta_S) - \frac{\phi(\delta_F)}{X_F}]\frac{F}{2} + \frac{\lambda_U}{\Sigma\lambda}\frac{X_F - 1}{X_F}\phi(\epsilon)[\frac{\lambda_I}{\lambda_I + \lambda_J} \times$$
$$\frac{1}{2}\frac{\gamma}{X_F}\frac{s}{2} + \frac{1-\gamma}{2}E - \frac{F}{2}] + \frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I + \lambda_J}\frac{1}{2}\frac{\gamma}{X_F}\frac{s}{2} - \frac{\lambda_U}{\Sigma\lambda}\phi(\delta_F + \epsilon - \delta_S)\frac{1-\gamma}{2}E \quad \text{(A.28)}$$

[22] 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.

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:

$$\frac{\lambda_I}{\Sigma\lambda}\frac{\gamma}{X_F}\frac{s}{2} + \frac{\lambda_U}{\Sigma\lambda}\frac{X_F-1}{X_F}\phi(\epsilon)\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{2}\frac{\gamma}{X_F}\frac{s}{2} + \frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{2}\frac{\gamma}{X_F}\frac{s}{2} - \frac{\lambda_U}{\Sigma\lambda}\phi(\delta_F+\epsilon-\delta_S)\frac{1-\gamma}{2}\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{X}\frac{s}{2} =$$
$$\frac{s}{4\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}[2(\lambda_I+\lambda_J)\frac{\gamma}{X_F} + \lambda_U\frac{X_F-1}{X_F}\phi(\epsilon)\frac{\gamma}{X_F} + \lambda_U\frac{\gamma}{X_F} - \lambda_U\phi(\delta_F+\epsilon-\delta_S)\frac{1-\gamma}{X}] \geq 0 \quad \text{(A.29)}$$

Note that we replaced $E = \frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{X}\frac{s}{2}$ in the term of (A.28). Thus, (A.28) would be non-negative if:

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

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:

$$2(\lambda_I+\lambda_J)\frac{\gamma}{X} + \lambda_U\frac{X-1}{X}\phi(\epsilon)\frac{\gamma}{X} + \lambda_U\frac{\gamma}{X} \geq \lambda_U\phi(\epsilon)\frac{1-\gamma}{X}$$

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

(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 $2(\lambda_I+\lambda_J)\gamma + \lambda_U\frac{2-1}{2}\phi(\epsilon)\gamma + \lambda_U\gamma \geq \lambda_U\phi(\epsilon)(1-\gamma)$, which is equivalent as:

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

This implies that if γ 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 (5) (homogeneous order processing speed) evaluated at order processing speed δ_S. Because if she switches, we would have HFT's provide liquidity on X slow exchanges with order processing speed δ_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 (5) we have:

$$\pi(\frac{s}{2}, X|\delta=\delta_S) = \frac{\lambda_I+\lambda_J}{\Sigma\lambda}E - \frac{\lambda_I+\lambda_J}{\Sigma\lambda}F + \frac{\lambda_U}{\Sigma\lambda}\frac{1}{X}\phi(\delta_S)(\frac{1}{2}E-F) + \frac{\lambda_U}{\Sigma\lambda}\frac{1}{X}\times$$
$$[1-\phi(\delta_S)](\frac{1}{2}E-\frac{1}{2}F) + \frac{\lambda_U}{\Sigma\lambda}\frac{X-1}{X}\phi(\epsilon)(E-F) + \frac{\lambda_U}{\Sigma\lambda}\frac{X-1}{X}[1-\phi(\epsilon)](\frac{1}{2}E-\frac{1}{2}F) \quad \text{(A.33)}$$

Evaluating (A.26) at $X_F = 1$ and minus (A.33) generates:

$$\pi_F(\frac{s}{2}, 1, X-1) - \pi(\frac{s}{2}, X|\delta=\delta_S) = \frac{\lambda_I}{\Sigma\lambda}\gamma\frac{s}{2} - \gamma\frac{\lambda_I+\lambda_J}{\Sigma\lambda}E +$$

$$\frac{\lambda_U}{\Sigma\lambda}(\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{\gamma}{2}\frac{s}{2}-\frac{\gamma}{2}E)-\frac{\lambda_U}{\Sigma\lambda}\frac{X-1}{X}\phi(\epsilon)\frac{E}{2}+\frac{\lambda_U}{\Sigma\lambda}\frac{F}{2}[\frac{1}{X}\phi(\delta_S)+\frac{X-1}{X}\phi(\epsilon)-\phi(\delta_F)] \quad (A.34)$$

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

$$\begin{aligned}\frac{\lambda_I}{\Sigma\lambda}\gamma\frac{s}{2}-\gamma\frac{\lambda_I+\lambda_J}{\Sigma\lambda}E+\frac{\lambda_U}{\Sigma\lambda}(\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{\gamma}{2}\frac{s}{2}-\frac{\gamma}{2}E)-\frac{\lambda_U}{\Sigma\lambda}\frac{X-1}{X}\phi(\epsilon)\frac{E}{2}=\frac{\lambda_I}{\Sigma\lambda}\gamma\frac{s}{2}-\gamma\frac{\lambda_I+\lambda_J}{\Sigma\lambda}\times\\ \frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{X}\frac{s}{2}+\frac{\lambda_U}{\Sigma\lambda}(\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{\gamma}{2}\frac{s}{2}-\frac{\gamma}{2}\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{X}\frac{s}{2})-\frac{\lambda_U}{\Sigma\lambda}\frac{X-1}{X}\phi(\epsilon)\frac{1}{2}\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{1}{X}\frac{s}{2}\\ =\frac{X-1}{X}\frac{1}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}\frac{s}{2}[(\lambda_I+\lambda_J)\gamma+\lambda_U\frac{\gamma}{2}-\lambda_U\phi(\epsilon)\frac{1}{2X}] \quad (A.35)\end{aligned}$$

Since $X \geq 1$, (A.35) would be non-negative for all X if $(\lambda_I+\lambda_J)\gamma+\lambda_U\frac{\gamma}{2}-\lambda_U\phi(\epsilon)\frac{1}{2} \geq 0$, which is equivalent as:

$$\gamma \geq \frac{0.5\lambda_U\phi(\epsilon)}{\lambda_I+\lambda_J+0.5\lambda_U} \quad (A.36)$$

Because $\frac{0.5\phi(\epsilon)\lambda_U}{\lambda_I+\lambda_J+0.5\lambda_U} > \frac{0.5\lambda_U\phi(\epsilon)}{\lambda_I+\lambda_J+0.5\lambda_U+0.75\lambda_U\phi(\epsilon)}$, thus if γ 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 δ_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.[23] 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 (7)). If they provide liquidity on a slow exchange, their profits would be $\pi_S(s^*(\delta_F)/2, M^*(\delta_F), 1) = \pi_S(s^*(\delta_F)/2, M^*(\delta_F)+1) \leq \pi_F(s^*(\delta_F)/2, M^*(\delta_F)+1, 0) < 0$(from equation (A.28) and (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 $\pi_F(s^*(\delta_F)/2, M^*(\delta_F), 0) \geq \pi_S(s^*(\delta_F)/2, M^*(\delta_F)-1, 1)$ (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 (18) determines the equilibrium depth on slow exchanges (similar to equation (7)).

[23] If they provide liquidity with lager bid-ask spread, they have no chance to trade with uninformed investors. Because the later only trades one unit and prefer an exchange with lower bid-ask spread.

A.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 (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 $\frac{1}{\Sigma\lambda}$ time for these events to occur. With probability $\frac{\lambda_I}{\Sigma\lambda}$ an investor arrives first, if she is smart (with probability γ) 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 $\frac{\gamma}{K} + \frac{1-\gamma}{M^*(K)}$ and a slow exchanges have expected trading volume $(1-\gamma)\frac{M_S^*(K)}{M-K}\frac{1}{M^*(K)}$.[24]

With probability $\frac{\lambda_J}{\Sigma\lambda}$ 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 a slow exchanges have probability $\frac{M_S^*(K)}{M-K}$ 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 $\frac{M_S^*(K)}{M-K}$.

With probability $\frac{\lambda_U}{\Sigma\lambda}$ 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 $t_U + \delta_F$[25]. 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 $M_S^*(K)$

[24] According to Proposition 4 and since $\gamma \geq \bar{\gamma}$, HFTs provide liquidity on K fast exchanges and $M_S^*(K)$ slow exchanges. Thus, for a slow exchange it has probability $\frac{M_S^*(K)}{M-K}$ to be chosen by HFTs to provide liquidity. Only when the investor is non-smart (with probability $1-\gamma$), she will randomly choose one among those $M^*(K)$ exchanges to trade. Therefore, A slow has expected trading volume $(1-\gamma)\frac{M_S^*(K)}{M-K}\frac{1}{M^*(K)}$ when the uninformed investor arrives first.

[25] For example, if exchange 1 is a fast exchange and it is chosen by the undercutting HFT to submit price improving limit order. At $t_U + \delta_F$, exchange 1 processed this new order. And thus all co-located HFTs know the existence of this new order at $t_U + \delta_F$. Liquidity provision HFT on exchange 1 will immediately cancel her being undercut limit order at $t_U + \delta_F$ because this order has no chance to trade with an investor.

slow exchanges will cancel their being undercut limit orders at $t_U + \delta_F + \epsilon - \delta_S$ (see Figure 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 $\epsilon' = \delta_F + \epsilon - \delta_S$, each exchange's expected trading volume depends on whether there is trade-through or not.

Trade-through is possible on both fast ans slow exchanges if $t_I < t_U + \epsilon'$. From equation (A.9) and (A.11) we have $Prob(t_I < t_J, t_I < t_U + \epsilon' | t_I, t_J > t_U) = \frac{\lambda_I}{\lambda_I + \lambda_J}\phi(\epsilon')$ and $E(t_I - t_U | t_U < t_I < t_J, t_I < t_U + \epsilon') = \frac{1}{\lambda_I + \lambda_J} - \frac{1-\phi(\epsilon')}{\phi(\epsilon')}\epsilon'$, where $\phi(\epsilon') = 1 - e^{-(\lambda_I + \lambda_J)\epsilon'}$. In this case, a fast exchange's expected trading volume is still $\frac{\gamma}{K} + \frac{1-\gamma}{M^*(K)}$ and a slow exchange still has trading volume $(1-\gamma)\frac{M_S^*(K)}{M-K}\frac{1}{M^*(K)}$. 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 same arguments.

When $t_U + \epsilon' \leq t_I \leq t_U + \epsilon$, liquidity provision HFTs on all $M_S^*(K)$ slow exchanges have already canceled their being undercut limit orders. A fast exchange has expected trading volume $\frac{\gamma}{K} + \frac{1-\gamma}{M^*(K)} + \frac{1}{K}\frac{1-\gamma}{2}\frac{M_S^*(K)}{M^*(K)} = \frac{1}{2}(\frac{1+\gamma}{K} + \frac{1-\gamma}{M^*(K)})$ (here I use the fact that $M^*(K) = M_F^*(K) + M_S^*(K) = K + M_S^*(K)$). Comparing to the above case where $t_I < t_U + \epsilon'$, fast exchange has additional trading volume $\frac{1}{K}\frac{1-\gamma}{2}\frac{M_S^*(K)}{M^*(K)}$. 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 $M_S^*(K)$ 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 $\frac{M_S^*(K)}{M-K}$) and is chosen by the investor to trade (with probability $\frac{1}{M^*(K)}$). Thus, a slow exchange has expected trading volume $\frac{1-\gamma}{2}\frac{M_S^*(K)}{M-K}\frac{1}{M^*(K)}$.

When $t_I > t_U + \epsilon$, 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 $t_U + \epsilon' \leq t_I \leq t_U + \epsilon$ because each fast exchange has equal probability to be chosen by the undercutting HFT.[26] Slow exchanges expected trading volume would not be affected either because trade through is impossible on those $M_S^*(K)$ slow exchanges as long as $t_I \geq t_U + \epsilon'$. Therefore, we can combine the above two cases and conclude that when $t_I \geq t_U + \epsilon'$, a fast and slow exchange has expected trading volume $\frac{1}{2}(\frac{1+\gamma}{K} + \frac{1-\gamma}{M^*(K)})$ and $\frac{1-\gamma}{2}\frac{M_S^*(K)}{M-K}\frac{1}{M^*(K)}$ respectively. We

[26] Alternatively, we can directly calculate a fast exchange's expected trading volume when $t_I > t_U + \epsilon$ as $\frac{1}{K}(\frac{1}{2} + \frac{1}{2}\frac{\gamma}{K} + \frac{1}{2}\frac{1-\gamma}{M^*(K)}) + \frac{K-1}{K}(\frac{1}{2}\frac{\gamma}{K} + \frac{1}{2}\frac{1-\gamma}{M^*(K)}) = \frac{1}{2}(\frac{1+\gamma}{K} + \frac{1-\gamma}{M^*(K)})$.

have $Prob(t_I < t_J, t_I \geq t_U + \epsilon' | t_I, t_J > t_U) = \frac{\lambda_I}{\lambda_I+\lambda_J}[1-\phi(\epsilon')]$ (similar to (A.9)) and $E(t_I - t_U | t_U < t_I < t_J, t_I \geq t_U + \epsilon') = \frac{1}{\lambda_I+\lambda_J} + \epsilon'$ (similar to (A.15)).

The only case left is that the risky asset's common value jumps before investor's arrival, that is $t_U < t_J < t_I$. 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 σ 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 $Prob(t_J < t_I, t_J < t_U + \delta_F | t_I, t_J > t_U) = \frac{\lambda_J}{\lambda_I+\lambda_J}\phi(\delta_F)$ and $E(t_J - t_U | t_U < t_J < t_I, t_J < t_U + \delta_F) = \frac{1}{\lambda_I+\lambda_J} - \frac{1-\phi(\delta_F)}{\phi(\delta_F)}\delta_F$.

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 $Prob(t_J < t_I, t_U + \delta_F \leq t_J \leq t_U + \epsilon' | t_I, t_J > t_U) = \frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon') - \phi(\delta_F)]$ and $E(t_J - t_U | t_J < t_I, t_U + \delta_F \leq t_J \leq t_U + \epsilon') = \frac{1}{\lambda_I+\lambda_J} + \epsilon' - \frac{\epsilon'-\delta_F}{\phi(\epsilon'-\delta_F)}$.

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 $Prob(t_J < t_I, t_U + \epsilon' < t_J \leq t_U + \epsilon | t_I, t_J > t_U) = \frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon) - \phi(\epsilon')]$ and $E(t_J - t_U | t_J < t_I, t_U + \epsilon' < t_J \leq t_U + \epsilon) = \frac{1}{\lambda_I+\lambda_J} + \epsilon - \frac{\epsilon-\epsilon'}{\phi(\epsilon-\epsilon')}$.

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 $\frac{M_S^*(K)}{2(M-K)}$. Similar to the calculation in (A.9) and (A.15), we have $Prob(t_J < t_I, t_J > t_U + \epsilon | t_I, t_J > t_U) = \frac{\lambda_J}{\lambda_I+\lambda_J}[1-\phi(\epsilon)]$ 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:

$$(\frac{1}{\Sigma\lambda}+\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q_F(K)=\frac{\lambda_I}{\Sigma\lambda}[\frac{\gamma}{K}+\frac{1-\gamma}{M^*(K)}+(\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q_F(K)]+\frac{\lambda_J}{\Sigma\lambda}[1+(\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q_F(K)]+\frac{\lambda_U}{\Sigma\lambda}\times$$
$$\frac{\lambda_I}{\lambda_I+\lambda_J}\phi(\epsilon')[\frac{\gamma}{K}+\frac{1-\gamma}{M^*(K)}+(\epsilon+\frac{1-\phi(\epsilon')}{\phi(\epsilon')}\epsilon')Q_F(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}[1-\phi(\epsilon')][\frac{1}{2}(\frac{1+\gamma}{K}+\frac{1-\gamma}{M^*(K)})+(\epsilon-\epsilon')Q_F(K)]$$
$$+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}\phi(\delta_F)[1+\frac{1}{2K}+(\epsilon+\frac{1-\phi(\delta_F)}{\phi(\delta_F)}\delta_F)Q_F(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon')-\phi(\delta_F)][1+(\epsilon-\epsilon'+\frac{\epsilon'-\delta_F}{\phi(\epsilon'-\delta_F)})$$
$$\times Q_F(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon)-\phi(\epsilon')][1+\frac{\epsilon-\epsilon'}{\phi(\epsilon-\epsilon')}Q_F(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[1-\phi(\epsilon)]\cdot\frac{K+1}{2K} \quad (A.37)$$

Which implies that:

$$Q_F^*(K)=\lambda_I[\frac{\gamma}{K}+\frac{1-\gamma}{M^*(K)}]+\lambda_J+\frac{\lambda_U\lambda_J[\phi(\delta_F)+(1-K)(1-\phi(\epsilon))]}{2K\Sigma\lambda}+\frac{\lambda_U\lambda_I[1-\phi(\epsilon')](1-\gamma)M_S^*(K)}{2KM^*(K)\Sigma\lambda} \quad (A.38)$$

Similarly, for a slow exchange, we have:

$$(\frac{1}{\Sigma\lambda}+\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q_S(K)=\frac{\lambda_I}{\Sigma\lambda}[\frac{M_S^*(K)}{M-K}\frac{1-\gamma}{M^*(K)}+(\frac{1}{\lambda_I+\lambda_J}+\epsilon)Q_S(K)]+\frac{\lambda_J}{\Sigma\lambda}[\frac{M_S^*(K)}{M-K}+(\frac{1}{\lambda_I+\lambda_J}+\epsilon)\times$$
$$Q_S(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}\phi(\epsilon')[\frac{M_S^*(K)}{M-K}\frac{1-\gamma}{M^*(K)}+(\epsilon+\frac{1-\phi(\epsilon')}{\phi(\epsilon')}\epsilon')Q_S(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_I}{\lambda_I+\lambda_J}[1-\phi(\epsilon')][\frac{M_S^*(K)}{M-K}\frac{1-\gamma}{2M^*(K)}$$
$$+(\epsilon-\epsilon')Q_S(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}\phi(\delta_F)[\frac{M_S^*(K)}{M-K}+(\epsilon+\frac{1-\phi(\delta_F)}{\phi(\delta_F)}\delta_F)Q_S(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon')-\phi(\delta_F)][\frac{M_S^*(K)}{M-K}$$
$$+(\epsilon-\epsilon'+\frac{\epsilon'-\delta_F}{\phi(\epsilon'-\delta_F)})Q_S(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[\phi(\epsilon)-\phi(\epsilon')][\frac{M_S^*(K)}{2(M-K)}+\frac{\epsilon-\epsilon'}{\phi(\epsilon-\epsilon')}Q_S(K)]+\frac{\lambda_U}{\Sigma\lambda}\frac{\lambda_J}{\lambda_I+\lambda_J}[1-\phi(\epsilon)]$$
$$\times\frac{M_S^*(K)}{2(M-K)} \quad (A.39)$$

which implies that:

$$Q_S^*(K)=\frac{M_S^*(K)}{M-K}[\frac{\lambda_I(1-\gamma)}{M^*(K)}+\lambda_J]\{1-\frac{\lambda_U}{2\Sigma\lambda}[1-\phi(\epsilon')]\} \quad (A.40)$$

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

A.8 Proof of Corollary 2

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

$$Q_F^*(K)-Q^*(\delta_F)=\lambda_I\frac{1}{K}+\lambda_J\frac{M_F^*(K)}{K}+\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\phi(\delta_F)\frac{1}{K}-\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M_F^*(K)-1}{K}-\lambda_I\frac{1}{M}-\lambda_J\frac{M^*(\delta_F)}{M}$$
$$-\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\phi(\delta_F)\frac{1}{M}+\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M^*(\delta_F)-1}{M}=\lambda_I(\frac{1}{K}-\frac{1}{M})+\lambda_J[\frac{M_F^*(K)}{K}-\frac{M^*(\delta_F)}{M}]+\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\phi(\delta_F)(\frac{1}{K}-$$
$$\frac{1}{M})+\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}(\frac{1}{K}-\frac{1}{M})+\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}[\frac{M^*(\delta_F)}{M}-\frac{M_F^*(K)}{K}]>\lambda_J[\frac{M_F^*(K)}{K}-\frac{M^*(\delta_F)}{M}]+\frac{1-\phi(\epsilon)}{2}\times$$
$$\frac{\lambda_U\lambda_J}{\Sigma\lambda}[\frac{M^*(\delta_F)}{M}-\frac{M_F^*(K)}{K}]=\lambda_J[\frac{M_F^*(K)}{K}-\frac{M^*(\delta_F)}{M}][1-\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U}{\Sigma\lambda}]\ge 0 \quad (A.41)$$

The first inequality holds because $K < M$. The second inequality holds because $\frac{M_F^*(K)}{K} \geq \frac{M^*(\delta_F)}{M}$ (recall that $M_F^*(K) = \min\{M^*(\delta_F), K\}$ 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 $Q_S^*(K)$ is increasing in $M^*(K)$ for a given K:

$$\frac{dQ^*(\delta_F)}{dM^*(\delta_F)} = [1 - \frac{1-\phi(\epsilon)}{2}\frac{\lambda_U}{\Sigma\lambda}]\frac{\lambda_J}{M} > 0 \tag{A.42}$$

From equation (21):

$$Q_S^*(K) = [\frac{\lambda_I(1-\gamma)}{M-K}(1 - \frac{K}{M^*(K)}) + \frac{M^*(K)-K}{M-K}\lambda_J][1 - \frac{\lambda_U}{2\Sigma\lambda}(1-\phi(\epsilon'))] \tag{A.43}$$

Thus, $Q_S^*(K)$ is increasing in $M^*(K)$ for a given K. From equation (19) and (11), we have:

$$\begin{aligned} Q_F^*(K) - Q^*(\delta_F) &= \lambda_I[\frac{\gamma}{K} + \frac{1-\gamma}{M^*(K)}] + \lambda_J + \frac{\lambda_U\lambda_J[\phi(\delta_F) + (1-K)(1-\phi(\epsilon))]}{2K\Sigma\lambda} + \frac{\lambda_U\lambda_I[1-\phi(\epsilon')](1-\gamma)M_S^*(K)}{2KM^*(K)\Sigma\lambda} \\ &- \frac{\lambda_I}{M} - \lambda_J\frac{M^*(\delta_F)}{M} - \frac{\lambda_U\lambda_J}{2\Sigma\lambda}\frac{\phi(\delta_F)}{M} + \frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M^*(\delta_F)-1}{M} > \lambda_I[\frac{\gamma}{K} + \frac{1-\gamma}{M^*(K)}] + \lambda_J + \frac{\lambda_U\lambda_J}{2K\Sigma\lambda}[\phi(\delta_F) + \\ &(1-K)(1-\phi(\epsilon))] - \frac{\lambda_I}{M} - \lambda_J\frac{M}{M} - \frac{\lambda_U\lambda_J}{2\Sigma\lambda}\frac{\phi(\delta_F)}{M} + \frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M-1}{M} > \frac{\lambda_U\lambda_J}{2\Sigma\lambda}[\frac{\phi(\delta_F) + (1-K)(1-\phi(\epsilon))}{K} - \frac{\phi(\delta_F)}{M} \\ &+ (1-\phi(\epsilon))\frac{M-1}{M}] = \frac{\lambda_U\lambda_J}{2\Sigma\lambda}(\frac{1}{K} - \frac{1}{M})[\phi(\delta_F) + 1 - \phi(\epsilon)] > 0 \end{aligned} \tag{A.44}$$

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 $0 < s/2 < \sigma$ and $\gamma \geq \bar{\gamma}$:

$$\pi_F(\frac{s}{2}, X_F, X - X_F) \geq \pi_S(\frac{s}{2}, X) \tag{A.45}$$

For any $1 \leq X_F \leq X$ if $\pi_S(\frac{s}{2}, X) \geq 0$, where $X = X_F + X_S$ and $\pi_S(\frac{s}{2}, X) = \pi_S(\frac{s}{2}, X_F, X_S)$ because the later only depends on the sum of $X_F + X_S$. From equation (18), if $s^*(\delta_F)/2 < \sigma$ we have:

$$\pi_S(\frac{s^*(\delta_F)}{2}, K + M_S^*(K)) = \pi_S(\frac{s^*(\delta_F)}{2}, K, M_S^*(K)) \geq 0 \tag{A.46}$$

Thus, from (A.45) we have:

$$\pi(\frac{s^*(\delta_F)}{2}, K + M_S^*(K) | \delta = \delta_F) = \pi_F(\frac{s^*(\delta_F)}{2}, K + M_S^*(K), 0) \geq \pi_S(\frac{s^*(\delta_F)}{2}, K + M_S^*(K)) \geq 0 \tag{A.47}$$

Therefore, from equation (7) we conclude that $M^*(\delta_F) \geq M^*(K) = K + M_S^*(K)$. From equation (11) and (21) we have:

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

$$+\lambda_J]\{1-\frac{\lambda_U}{2\Sigma\lambda}[1-\phi(\epsilon')]\} \geq \lambda_I\frac{1}{M}+\lambda_J\frac{M^*(\delta_F)}{M}+\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\phi(\delta_F)\frac{1}{M}-\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M^*(\delta_F)-1}{M}-\frac{M^*(\delta_F)-K}{M-K}\times$$
$$[\frac{\lambda_I(1-\gamma)}{M^*(\delta_F)}+\lambda_J]\{1-\frac{\lambda_U}{2\Sigma\lambda}[1-\phi(\epsilon')]\} > \lambda_J(\frac{M^*(\delta_F)}{M}-\frac{M^*(\delta_F)-K}{M-K})-\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M^*(\delta_F)-1}{M}+\frac{M^*(\delta_F)-K}{M-K}$$
$$\times[\frac{\lambda_I(1-\gamma)}{M^*(\delta_F)}+\lambda_J]\frac{\lambda_U}{2\Sigma\lambda}[1-\phi(\epsilon')] > \lambda_J(\frac{M^*(\delta_F)}{M}-\frac{M^*(\delta_F)-K}{M-K})[1-\frac{\lambda_U}{2\Sigma\lambda}[1-\phi(\epsilon)]\geq 0 \quad \text{(A.48)}$$

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 $\frac{1}{M} \geq \frac{M^*(\delta_F)-K}{M-K}\frac{1}{M^*(\delta_F)}$. The third inequality holds because $\phi(\epsilon') < \phi(\epsilon)$.

(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 (19) and (11), we have:

$$Q_F^*(1)-Q^*(\delta_S) = \lambda_I+\lambda_J+\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\phi(\delta_F)-\frac{\lambda_I}{M}-\lambda_J\frac{M^*(\delta_S)}{M}-\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\frac{\phi(\delta_S)}{M}+\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M^*(\delta_S)-1}{M} \geq \lambda_I+\lambda_J$$
$$+\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\phi(\delta_F)-\frac{\lambda_I}{M}-\lambda_J\frac{M-1}{M}-\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\frac{\phi(\delta_S)}{M}+\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M-2}{M} > \frac{\lambda_J}{M}[1-\frac{\lambda_U}{2\Sigma\lambda}\phi(\delta_S)] > 0 \quad \text{(A.49)}$$

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:

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

If:

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

Similarly, when $M^*(\delta_F) > K$, $\pi_S(s^*(\delta_F)/2, K, 1) \geq 0$ and $M^*(\delta_S) < M$ from equation (19) and (11) we have:

$$Q_F^*(1)-Q^*(\delta_S) = \lambda_I[\gamma+\frac{1-\gamma}{M^*(1)}]+\lambda_J+\frac{\lambda_U\lambda_J\phi(\delta_F)}{2\Sigma\lambda}+\frac{\lambda_U\lambda_I[1-\phi(\epsilon')](1-\gamma)M_S^*(1)}{2M^*(1)\Sigma\lambda}-\frac{\lambda_I}{M}-\lambda_J\frac{M^*(\delta_S)}{M}$$
$$-\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\frac{\phi(\delta_S)}{M}+\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M^*(\delta_S)-1}{M} > \lambda_J-\lambda_J\frac{M-1}{M}-\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\frac{\phi(\delta_S)}{M}=\frac{\lambda_J}{M}[1-\frac{\lambda_U}{2\Sigma\lambda}\phi(\delta_S)] > 0 \quad \text{(A.52)}$$

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:

$$Q_F^*(1)-Q^*(\delta_S) = \lambda_I[\gamma+\frac{1-\gamma}{M^*(1)}]+\lambda_J+\frac{\lambda_U\lambda_J\phi(\delta_F)}{2\Sigma\lambda}+\frac{\lambda_U\lambda_I[1-\phi(\epsilon')](1-\gamma)M_S^*(1)}{2M^*(1)\Sigma\lambda}-\frac{\lambda_I}{M}-\lambda_J\frac{M}{M}$$
$$-\frac{\lambda_U\lambda_J}{2\Sigma\lambda}\frac{\phi(\delta_S)}{M}+\frac{1-\phi(\epsilon)}{2}\frac{\lambda_U\lambda_J}{\Sigma\lambda}\frac{M-1}{M} > \frac{\lambda_U\lambda_J}{2\Sigma\lambda}[\phi(\delta_F)-\frac{\phi(\delta_S)}{M}+(1-\phi(\epsilon))\frac{M-1}{M}] \geq 0 \quad \text{(A.53)}$$

if (A.51) holds.

A.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:

$$\bar{f}Q^*(\delta_S) < \bar{f}Q_F^*(1) - C_{speed} \Longleftrightarrow \frac{C_{speed}}{\bar{f}} < Q_F^*(1) - Q^*(\delta_S) \tag{A.54}$$

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:

$$\bar{f}Q_S^*(K) < \bar{f}Q^*(\delta_F) - C_{speed} \Longleftrightarrow \frac{C_{speed}}{\bar{f}} < Q^*(\delta_F) - Q_S^*(K) \tag{A.55}$$

This is because $\bar{f}Q^*(\delta_F) - C_{speed} \leq \bar{f}Q_F^*(K+1) - C_{speed}$ from Corollary 2 (i).[27] 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 (21). Note that the maximum $X_F + X_S$ such that equation (17) is non-negative at $s^*(\delta_F)$ is independent of K and denote this maximum depth as $\bar{M}$. Thus, $M_S^*(K) = \bar{M} - K$. Equation (21) can be rewritten as:

$$Q_S^*(K) = \frac{\bar{M}-K}{M-K}[\frac{\lambda_I(1-\gamma)}{\bar{M}} + \lambda_J]\{1 - \frac{\lambda_U}{2\Sigma\lambda}[1-\phi(\epsilon')]\} \tag{A.56}$$

which is clearly decreasing in K. Therefore, as long as $C_{speed}/\bar{f} < min\{Q_F^*(1) - Q^*(\delta_S), Q^*(\delta_F) - Q_S^*(1)\}$ (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:

$$\bar{f}Q^*(\delta_F) - C_{speed} < \bar{f}Q^*(\delta_S) \Longleftrightarrow \frac{C_{speed}}{\bar{f}} > Q^*(\delta_F) - Q^*(\delta_S) \tag{A.57}$$

(iii) is the same result as in Proposition 2 (iii) by simply minus the taker fee from investor's welfare defined in equation (10). Thus, it is the same proof of Proposition 2 (iii).

Appendix B Equilibrium Analysis with Exchange Speed Heterogeneity When $\gamma < \bar{\gamma}$

When $\gamma < \bar{\gamma}$ it is possible that $\pi_F(\frac{s}{2}, X_F, X_S) < \pi_S(\frac{s}{2}, X_F, X_S)$ 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

[27] 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 δ_F. We have $Q_F^*(M) = Q^*(\delta_F)$).

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 $M_S^*(1)$ (determined in equation (18)) slow exchanges. But if $\pi_F(\frac{s^*(\delta_F)}{2}, 1, M_S^*(1)) < \pi_S(\frac{s^*(\delta_F)}{2}, 1, M_S^*(1))$, it is possible that $\pi_F(\frac{s^*(\delta_F)}{2}, 1, M_S^*(1)) < 0$. 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 $\gamma < \bar{\gamma}$, 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:

$$\max_{1 \leq X_F \leq K; J \leq X_S \leq M-K} X_F \pi_F(s/2, X_F, X_S) + (X_S - J)\pi_S(s/2, X_F, X_S) \tag{B.1}$$

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

$$\pi_F(s/2, X_F + 1, X_S) < 0 \quad \text{if} \quad X_F + 1 \leq K$$

what problem (B.1) 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 (B.1) always exists because $X_F = K$ and $X_S = M - K$ satisfy all conditions in problem (B.1). Further, we denote:

$$J^*(s/2, K) = \max\{1 \leq J \leq M - K \;\; s.t. \;\; \pi(s/2, J|\delta = \delta_S) \geq 0\} \tag{B.2}$$

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 (5)). Define:

$$TP_1(s/2, K, J) = X_F^* \pi_F(s/2, X_F^*, X_S^*) + [X_S^* - J]\pi_S(s/2, X_F^*, X_S^*) \tag{B.3}$$

Where $X_F^* = X_F^*(s/2, K, J)$ and $X_S^* = X_S^*(s/2, K, J)$ are the solutions for problem (B.1). And:

$$TP_2(s/2, K) = J^*(s/2, K)\pi(s/2, J^*(s/2, K)|\delta = \delta_S) \tag{B.4}$$

We will discuss three potential equilibriums. Denoting E_1 as the case when a single HFT provides liquidity on $X_F^*(s^*(\delta_F)/2, K, 0)$ fast exchanges and $X_S^*(s^*(\delta_F)/2, K, 0)$ slow exchanges with bid-ask spread $s^*(\delta_F)$; E_2 as the case when a single HFT provides liquidity on $X_F^*(s^*(\delta_S)/2, K, 0)$ fast exchanges and $X_S^*(s^*(\delta_S)/2, K, 0)$ 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 7. (Equilibrium with Exchange Speed Heterogeneity When $\gamma < \bar{\gamma}$) *If there are K fast exchanges with order processing speed δ_F and $M - K$ slow exchanges with order processing speed δ_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 $TP_1(s^*(\delta_F)/2, K, 0) \geq 0$, E_1 is an equilibrium;*
(ii) *If $s^*(\delta_F) = s^*(\delta_S)$ or $TP_1(s^*(\delta_F)/2, K, 0) < 0$, either E_2 or E_3 could be an equilibrium: 1) if $TP_1(s^*(\delta_S)/2, K, 0) > TP_2(s^*(\delta_S)/2, K)$, E_2 is an equilibrium; 2) if $TP_1(s^*(\delta_S)/2, K, 0) < 0$, E_3 is an equilibrium; 3) if $0 \leq TP_1(s^*(\delta_S)/2, K, 0) \leq TP_2(s^*(\delta_S)/2, K)$, then E_2 is an equilibrium if:*

$$TP_1(s^*(\delta_S)/2, K, J^*(s^*(\delta_S)/2, K)) \geq 0 \tag{B.5}$$

Otherwise, E_3 is an equilibrium.

Proof. In (i) a single HFT can make non-negative profits by providing liquidity on $X_F^*(s^*(\delta_F)/2, K, 0)$ fast exchanges and $X_S^*(s^*(\delta_F)/2, K, 0)$ slow exchanges with bid-ask spread $s^*(\delta_F)$. Since $X_F^*(s^*(\delta_F)/2, K, 0)$ and $X_S^*(s^*(\delta_F)/2, K, 0)$ are solutions for problem (B.1), 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 a HFT can quote. 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 $X_F^*(s^*(\delta_F)/2, K, 0)$ fast exchanges and $X_S^*(s^*(\delta_F)/2, K, 0)$ 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 (B.1) hold. The original single HFT's profit would be smaller than the profits by providing liquidity on $X_F^*(s^*(\delta_F)/2, K, 0)$ fast exchanges and $X_S^*(s^*(\delta_F)/2, K, 0)$ slow exchanges with bid-ask spread $s^*(\delta_F)$ because the later maximizes the total liquidity provision profits (solving problem (B.1) with $J = 0$). 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 $TP_1(s^*(\delta_S)/2, K, 0) > TP_2(s^*(\delta_S)/2, K)$, the single HFT provides liquidity on $X_F^*(s^*(\delta_S)/2, K, 0)$ fast exchanges and $X_S^*(s^*(\delta_S)/2, K, 0)$ 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 $TP_1(s^*(\delta_S)/2, K, 0) < 0$, 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 (B.1), 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 $0 \leq TP_1(s^*(\delta_S)/2, K, 0) \leq TP_2(s^*(\delta_S)/2, K)$, a single HFT has lager 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 $X_F^*(s^*(\delta_S)/2, K, 0)$ fast exchanges and $X_S^*(s^*(\delta_S)/2, K, 0)$ slow exchanges with bid-ask spread $s^*(\delta_S)$ has no incentive to change her quotes. Since $TP_1(s^*(\delta_S)/2, K, 0) \leq TP_2(s^*(\delta_S)/2, K)$, 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 $TP_1(s^*(\delta_S)/2, K, 0)$. 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 7 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 λ_I, λ_J and λ_U. But for most cases, the equilibrium features the same spread and depth as stated in Proposition 7.

References

Angel, J. J., L. E. Harris, and C. S. Spatt (2015): "Equity trading in the 21st century: An update," *The Quarterly Journal of Finance*, 5(01), 1550002.

Baldauf, M., and J. Mollner (2017): "High-frequency trading and market performance," *working paper*.

Baruch, S., and L. R. Glosten (2016): "Strategic foundation for the tail expectation in limit order book markets," *working paper*.

Bennett, P., and L. Wei (2006): "Market structure, fragmentation, and market quality," *Journal of Financial Markets*, 9(1), 49–78.

Biais, B., F. Declerck, and S. Moinas (2016): "Who supplies liquidity, how and when?," *working paper*.

Biais, B., T. Foucault, and S. Moinas (2015): "Equilibrium fast trading," *Journal of Financial Economics*, 116(2), 292–313.

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan (2015): "Trading fast and slow: Colocation and liquidity," *Review of Financial Studies*, 28(12), 3407–3443.

Brogaard, J., T. Hendershott, and R. Riordan (2014): "High-frequency trading and price discovery," *The Review of Financial Studies*, 27(8), 2267–2306.

Budish, E., P. Cramton, and J. Shim (2014): "Implementation details for frequent batch auctions: Slowing down markets to the blink of an eye," *The American Economic Review*, 104(5), 418–424.

——— (2015, BCS): "The high-frequency trading arms race: Frequent batch auctions as a market design response," *The Quarterly Journal of Economics*, 130(4), 1547–1621.

Caglio, C., and S. Mayhew (2012): "Equity trading and the allocation of market data revenue," *working paper*.

Chao, Y., C. Yao, and M. Ye (2017): "Why Discrete Price Fragments US Stock Exchanges and Disperses Their Fee Structures?," Discussion paper, Conditionally Accepted, Review of Financial Studies.

Colliard, J.-E., and T. Foucault (2012): "Trading fees and efficiency in limit order markets," *Review of Financial Studies*, 25(11), 3389–3421.

Du, S., and H. Zhu (2017): "What is the Optimal Trading Frequency in Financial Markets?," *The Review of Economic Studies*, p. rdx006.

Foucault, T., R. Kozhan, and W. W. Tham (2017): "Toxic arbitrage," *The Review of Financial Studies*, 30(4), 1053–1094.

Foucault, T., and A. J. Menkveld (2008): "Competition for order flow and smart order routing systems," *The Journal of Finance*, 63(1), 119–158.

Glosten, L. R. (1998): "Competition, design of exchanges and welfare," *working paper*.

Glosten, L. R., and P. R. Milgrom (1985): "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of financial economics*, 14(1), 71–100.

Goldstein, M. A., A. V. Shkilko, B. F. Van Ness, and R. A. Van Ness (2008): "Competition in the market for NASDAQ securities," *Journal of Financial Markets*, 11(2), 113–143.

Griffith, T., and B. S. Roseman (2016): "Making Cents of Tick Sizes," *working paper*.

Hasbrouck, J. (2015): "High frequency quoting: Short-term volatility in bids and offers," *working paper*.

Hendershott, T., C. M. Jones, and A. J. Menkveld (2011): "Does algorithmic trading improve liquidity?," *The Journal of Finance*, 66(1), 1–33.

Hoffmann, P. (2014): "A dynamic limit order market with fast and slow traders," *Journal of Financial Economics*, 113(1), 156–169.

Jovanovic, B., and A. J. Menkveld (2016): "Middlemen in limit order markets," *working paper*.

Kirilenko, A., A. S. Kyle, M. Samadi, and T. Tuzun (2017): "The Flash Crash: High-Frequency Trading in an Electronic Market," *The Journal of Finance*.

Kyle, A. S. (1985): "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society*, pp. 1315–1335.

Lewis, M. M. (2014): *Flash boys: a Wall Street revolt*. WW Norton New York, NY.

Malinova, K., and A. Park (2016): "ModernMarket Makers," *working paper*.

Menkveld, A. J. (2013): "High frequency trading and the new market makers," *Journal of Financial Markets*, 16(4), 712–740.

——— (2016): "The economics of high-frequency trading: Taking stock," *Annual Review of Financial Economics*, 8, 1–24.

Menkveld, A. J., and M. A. Zoican (2016): "Need for speed? Exchange latency and liquidity," *working paper*.

Miller, R. S., and G. Shorter (2016): "High Frequency Trading: Overview of Recent Developments," *Washington: Congressional Research Service*.

O'Hara, M., G. Saar, and Z. Zhong (2015): "Relative tick size and the trading environment," *working paper*.

Pagnotta, E., and T. Philippon (2016): "Competing on speed," Discussion paper, National Bureau of Economic Research.

Petersen, M. A. (2004): "Information: Hard and soft," *working paper*.

Rindi, B., and I. M. Werner (2017): "US Tick Size Pilot," *working paper*.

Song, S., and C. Yao (2016): "The Real Costs of Natural Experiments," *working paper*.

Tabb, L. (2016): "Stock Exchanges Are Eating Your Returns," https://www.bloomberg.com/view/articles/2016-01-22/stock-exchanges-data-fees-harm-investors, Online; posted 22-January-2016.

Tuttle, L. A. (2013): "Alternative trading systems: Description of ATS trading in national market system stocks," *working paper*.

Wah, E., and M. P. Wellman (2013): "Latency arbitrage, market fragmentation, and efficiency: a two-market model," in *Proceedings of the fourteenth ACM conference on Electronic commerce*, pp. 855–872. ACM.

Wang, X. (2017b): "Can Stock Exchanges with 'Speed Bump' Survive?," *working paper*.

Wang, X., and M. Ye (2017): "Who Provides Liquidity and When: An Analysis of Price vs. Speed Competition on Liquidity and Welfare," *working paper*.

Yao, C., and M. Ye (2017): "Why trading speed matters: A tale of queue rationing under price controls," *Forthcoming Review of Financial Studies*.

Ye, M., C. Yao, and J. Gai (2013): "The externalities of high frequency trading," *working paper*.

Zhu, H. (2014): "Do dark pools harm price discovery?," *Review of Financial Studies*, 27(3), 747–789.