A Model of Strategic High-Frequency Trading and For-Profit Exchanges with Intentional Delays

Jun Aoyagi*

July 30, 2021

Abstract

This paper studies competition between strategic high-frequency traders (HFTs) and multiple for-profit exchanges. In the model, HFTs play a dual role as liquidity snipers and market makers and strategically decide on their trading venue, the intensity of market monitoring, a bid-ask spread, and speed technologies. With the strategic liquidity provision and HFTs' dual role in the market, I show that the expected bid-ask spread can shrink when adverse selection becomes more severe. I also derive the HFTs' demand for speed services and demonstrate that it can be an increasing function of the length of intentional delays imposed on trade execution by exchange platforms (e.g., speed bumps). In the second part, for-profit exchanges try to maximize their revenues from supplying speed services to HFTs by controlling the speed of order execution. Since the demand for speed services can positively react to the intentional delays in order execution, exchanges are willing to introduce them to boost their profits. Thus the imposition of delays, which mitigates adverse selection and improves liquidity, is supported as an equilibrium outcome even without government intervention.

Keywords: high-frequency trading, for-profit exchanges, intentional delays, market liquidity

JEL Classification: D40, D47, G10, G18, G20

*The Department of Finance, The Hong Kong University of Science and Technology. E-mail junaoyagi@ust.hk.
I deeply acknowledge that Nicolae Gârleanu and Christine Parlour provided invaluable assistance. I appreciate the constructive comments from Kosuke Aoki, Kerry Back, Matteo Benetton, Michael Brolley, Joel Hasbrouck, Terry Hendershott, Ohad Kadan, Hayden Melton, Sophie Moinas, Emi Nakamura, David Sraer, Alberto Teguia, Yingge Yan, Mao Ye, Haoxiang Zhu, and Marius Zoican.

1 Introduction

The ever-increasing speed of high-frequency trading has attracted the attention of finance researchers and practitioners.[1] For traders, trading speed matters both when taking liquidity and when making markets. Upon the arrival of new information, liquidity takers can glean (almost) risk-free profit opportunities by placing liquidity-taking orders and sniping a standing limit order before a market maker updates her old quote. To avoid being picked off by snipers, in turn, market makers try to cancel and reprice their stale limit orders as quickly as possible by leveraging their speed technologies. Exchanges also deploy ultra-fast information processors[2] and provide *speed services*, such as colocation services and direct data feeds, to high-frequency traders (HFTs) for fees.

Overall, high-frequency trading attains massive presence in the market. They account for more than 50% of trading volume in the U.S. and about 40% in Europe. At the same time, however, a race to an arbitrage opportunity is highly concentrated. Aquilina, Budish and O'Neill (2020) find that 6 large high-frequency financial institutions are serving both as liquidity takers and providers in about 43% of races for FTSE 100 on London Stock Exchange, suggesting that each race is played by a limited number of players. The entire trading volume is also concentrated: Smith (2015) reports that the top 5 high-frequency financial institutions account for about 70% of the trading volume on BrokerTeck (which handles more than half of the U.S. Treasury). The literature has developed models where HFTs are competitive and serving both as takers and makers, reflecting their dual role in the real financial market (e.g., Menkveld and Zoican, 2017). This paper extends the existing environment to obtain further insights into the behavior of HFTs and exchange platforms with a new market structure.

First, I consider the strategic behavior of HFTs rather than focusing on a competitive environment. This is motivated by the above-mentioned concentration of players in the high-frequency races. Several papers (see below) have analyzed strategic liquidity takers or market makers separately, but my model is the first to analyze HFTs' strategic behavior when they play both as takers and makers.

Also, the existing studies introduce speed as a binary choice variable (i.e., a trader's choice is being an HFT with an exogenous speed level or staying as a regular slow trader), and the equilibrium is characterized by a unique bid-ask spread that makes all HFTs indifferent between taking and

[1] Quincy Data, one of the third-party institutions supplying the fastest data access, claims that they can send a piece of information form Aurora, Chicago, to Secaucus, New Jersey, in less than 4 microseconds (4×10^{-6} seconds).

[2] For example, throughput at NYSE has declined dramatically from 350 milliseconds in 2007 to 5 milliseconds in 2009.

providing liquidity. In my model, it is no longer an equilibrium. HFTs endogenously choose their speed technologies as a continuous variable,[3] meaning that they are potentially equipped with heterogeneous speed technologies and have heterogeneous indifference conditions. My model provides a tractable framework to analyze such a situation and derives an explicit formula for the HFTs' optimal demand for speed services in the continuous domain. It leads to the optimization of exchanges that try to maximize fee revenues from supplying speed services.

Strategic high-frequency traders play a one-shot trading game à la Glosten and Milgrom (1985) with N competing exchanges. Before the trading game starts, HFTs purchase speed services from exchanges. At the beginning of the game, each HFT serves as an HFM: she places a limit order on one of the exchanges.[4] Then, by allocating the limited information capacity, she starts processing data about the limit order book on each exchange (e.g., direct data feeds) to locate her rivals' limit orders and profitable trading opportunities. Responding to public news, she serves as a high-frequency sniper (an HFS) by placing a market order to the exchange with an arbitrage opportunity. She simultaneously sends a cancellation/repricing request to revoke her previous limit order and to avoid being picked off by her rivals.

My result shows that the equilibrium bid-ask spread can shrink when the adverse selection cost for market makers becomes more severe. As in Dennert (1993), an HFM adopts a mixed strategy and randomizes her quote when placing a limit order. In expectation, she sets her quote in order to make her rival HFM indifferent between all feasible strategies, i.e., the rival breaks even. When HFT i increases her speed, she faces a lower risk of being picked off and earns a higher expected profit as an HFM. To make HFT i break-even as a market maker, HFT j must reduce HFT i's expected market-making profit by quoting a narrower bid-ask spread and taking profitable noise trading away from HFT i. At the same time, however, HFT i also serves as a sniper, and her speed-up exacerbates the adverse selection cost for HFT j as a market maker. Therefore, even though HFT j faces more severe adverse selection, she proposes a narrower spread.

The above result goes counter to the conventional theory, in which the equilibrium bid-ask spread

[3] In reality, there are many factors that affect the speed and traders have ample choice sets. See, for example, IEX (2019) for a variety of speed services provided by the U.S. exchange families. Also, there are third party speed providers, such as Quincy Data and Quod Financial.

[4] Throughout the paper, I use the terms "market orders" and "limit orders" to represent liquidity-taking and liquidity-providing orders. Alternatively, we can think of other types of orders, such as Immediate-or-Cancel (IOC) to take liquidity and Post Only (PO) to provide liquidity.

positively reflects the adverse selection cost for market makers. The difference arises because the conventional argument considers a common source of adverse selection, while my model allows heterogeneous sources, and HFTs play their dual role. For example, if noise traders become more active, it works as a common factor that mitigates adverse selection for *all* market makers, and the bid-ask spread declines. In contrast, if a reduction in the adverse selection cost occurs due to idiosyncratic reasons, such as a speed-up by one HFT, the reaction of the bid-ask spread is not trivial—it depends on whose quote we are looking at and what is the source of a change in adverse selection.

Thirdly, my model incorporates the latest market structure. While providing speed services to HFTs and adopting ultra-fast information processors, exchanges start introducing a somewhat contradicting market structure, called *speed bumps*. They impose intentional delays on the execution of (a part of) trading orders. As tabulated in Table 1 in Appendix A, most delays are asymmetric and applied only to liquidity-taking orders. The exchanges (and SEC) argue that the delays aim at slowing down high-frequency snipers and protecting market makers against latency arbitrage so that exchanges attain more liquidity.

My model shows that intentional delays in execution of liquidity-taking orders can *increase* the HFTs' demand for speed services. Delays directly hamper sniping and mitigate adverse selection for market makers, leading to a narrower bid-ask spread. A narrower spread, in turn, implies a larger profit margin for snipers, and adding one unit of speed becomes more valuable. Moreover, the impact of a speed-up on the bid-ask spread and the sniping probability tends to diminish more slowly if intentional delays kick in. Thus HFTs can increase their speed more aggressively compared to the case with no delays. Of course, the imposition of delays makes it harder to snipe stale limit orders and reduce the expected prize for snipers, making them more reluctant to pay the fees to acquire speed services. Due to these competing effects, the impact of delays on HFTs' speed acquisition tends to be ambiguous.

Finally, I analyze competition between for-profit exchanges and endogenize the intensity (or the length) of delays in order execution. Since the demand for speed services exhibits a hump-shaped reaction to the intensity of delays, the model explicitly derives the optimal delays for a for-profit exchange. Therefore, I provide an answer to the question; "Do for-profit exchanges have an incentive to introduce intentional delays?" Budish, Lee and Shim (2020) show that, if the cost of adopting a market structure that prevents sniping is small, exchanges will not introduce it.[5] They argue that

[5] Budish, Lee and Shim (2020) focus on frequent batch auctions (FBAs) proposed by Budish, Cramton and Shim (2015)

the recent exchanges earn a huge portion of profits from supplying speed services, and introducing a new structure that invalidates HFTs' speed advantage will harm exchanges' profits. In my model, however, the optimal speed for HFTs can increase due to the imposition delays, allowing exchanges to boost their profit by intentionally delaying order execution.

My results have an important policy implication. Firstly, it suggests that intentional delays can mitigate the adverse selection problem for high-frequency market makers and improve market liquidity. Although delays can increase the trading speed of HFTs, they leverage it both as takers and makers, leaving the adverse selection cost in the symmetric equilibrium unaffected by HFTs' speed-up (Budish, Cramton and Shim, 2015; Brolley and Zoican, 2020). Secondly, the adoption of intentional delays is consistent with the profit maximization of competing exchanges. Thus, as in the real financial market, exchanges will self-impose them without government intervention. By appropriately adjusting the intensity of delays, we can simultaneously achieve improved market liquidity—the purported rationale of the imposition of delays—and higher fee revenue—the primary objective of exchange platforms.

This paper contributes to the literature on high-frequency trading and market microstructure. Endogenous speed acquisition of traders has been analyzed in the existing studies, such as Foucault, Roell and Sandas (2003), Liu (2009), Foucault, Kadan and Kandel (2013), Foucault, Kozhan and Tham (2016) but they abstract away from a strategic motive of HFTs, a slow market structure (i.e., order execution delays), or the dual role played by HFTs.[6] Several studies work on slow market structures, such as frequent batch auctions (Budish, Cramton and Shim, 2015; Haas and Zoican, 2016) and speed bumps (Aldrich and Friedman, 2018; Baldauf and Mollner, 2020; Brolley and Cimon, 2020),[7] by focusing on the HFTs' binary choice on speed and a competitive environment.[8] Starting from Kyle (1985), a large body of literature has analyzed strategic behavior of liquidity takers (e.g., Kyle and Wang, 1997; Back and Baruch, 2004) and market makers (e.g., Dennert, 1993; Baruch and Glosten,

but the similar discussion can be applied to analyze intentional delays.

[6] Jones (2013), O'Hara (2015), and Menkveld (2016) provide wholistic reviews on high-frequency trading.

[7] Du and Zhu (2017), Kyle and Lee (2017), Menkveld and Zoican (2017), and Pagnotta and Philippon (2018) analyze the exchange speed, i.e., the common speed for all traders.

[8] There are several empirical studies on the impact of intentional delays, such as those by Hu (2018), Chakrabarty, Huang and Jain (2019), Shkilko and Sokolov (2016), Chen, Foley, Goldstein and Ruf (2017), Anderson, Andrews, Devani, Mueller and Walton (2018), and Khapko and Zoican (2019), but they report somewhat mixed results. An experimental study by Khapko and Zoican (2019) finds that a marginally longer speed bump stimulates the traders' investment in speed when the execution price is endogenous.

2013; Ait-Sahalia and Saglam, 2017; Baruch and Glosten, 2019) separately. Roşu (2009) is an exception and considers strategic traders choosing between placing limit orders and market orders in an environment with symmetric information and speed. My paper incorporates all the above factors: strategic HFTs, speed acquisition as a continuous choice variable, intentional delays, and the dual role played by HFTs.

The positive reaction of bid-ask spreads (and the inverse market depth) to the adverse selection risk for market makers is formalized by Glosten and Milgrom (1985) and Kyle (1985). However, empirical studies have reported somewhat nuanced results, especially in the context of high-frequency trading. They take developments of speed technologies as a source of changes in adverse selection, and the result depends on whether the faster speed technologies are exploited more by snipers or market makers. For example, Hendershott, Jones and Menkveld (2011) and Boehmer, Fong and Wu (2015) find that the bid-ask spread shrinks due to increases in algorithmic trading, while Foucault, Kozhan and Tham (2016) and Brogaard, Hendershott and Riordan (2017) report the opposite result. My model shows that even if an HFT adopts the same speed to play both as a taker and a maker, a speed-up causes an ambiguous reaction of a bid-ask spread due to their strategic behavior.

The scope of this paper extends to the behavior of exchange platforms. Foucault and Parlour (2004) analyze competition between two exchanges seeking to earn listing fees. Pagnotta and Philippon (2018) consider competition on trading speed to attract competitive latency-sensitive traders. The closest to my paper is the theoretical study by Budish, Lee and Shim (2020), in which they deal with competitive HFTs serving both as takers and makers. They consider competing for-profit exchanges deciding on the adoption of frequent batch auctions (FBAs) as a tool to hamper latency arbitrage. They argue that exchanges have no incentive to adopt FBA in the equilibrium, as it disincentivizes HFTs to purchase speed services and reduces exchanges' profit.[9] My model complements their results, as it takes into account strategic HFTs and shows that exchanges are willing to introduce intentional delays.

2 The model

[9]Their result is conditional on a small cost of adopting FBAs for copycats. The first exchange can introduce FBAs and earn more trading fees by attracting liquidity, perhaps making positive profits net of adoption costs. However, other exchanges may follow the first mover, as the cost of adoption is small for them, diluting the profit of the first mover. By taking steps backward, the first mover has no incentive to introduce FBAs.

2.1 Trading environment

Consider a one-shot exchange of an asset between two high-frequency traders (HFTs) and a liquidity trader.

Asset. A single risky asset is traded. When the market opens at time $t = 1$, the asset has value $v = 0$, which is common knowledge. After $t = 1$, a trade takes place either by a shock on the common value of the asset (v) or a shock on the liquidity trader's private value of holding the asset. A common-value shock hits with a Poisson rate z_c, and the value of the asset becomes $\tilde{v} = \pm\sigma$ with the same probability.[10] In contrast, a private-value shock happens with a Poisson rate z_p. It makes a liquidity trader need to buy or sell one unit of the asset with the same probability, causing noise trading.

I focus on a very short time interval so that a trigger shock occurs at most once, and two types of shocks are mutually exclusive, as in Menkveld and Zoican (2017) and Brolley and Zoican (2020). Since they arrive as a Poisson event, a common-value shock triggers a trade with probability $\eta = \frac{z_c}{z_c+z_p}$, while a private-value shock triggers it with the complementary probability, $1-\eta$.

Exchanges. There are $N \geq 2$ exchange platforms operating in parallel with each other. They are indexed by $k \in \mathcal{E} = \{1, 2, \cdots, N\}$ where $\mathcal{E}$ denotes the set of exchanges. I first take N as an exogenous parameter, but Section 4 endogenizes it as an equilibrium variable.

Following the trading rules in the U.S. (see SEC, 2005 or Budish et al., 2020 for more details), a trader can trade the asset on all exchanges at the best quoted price among them. Therefore N exchanges are perfect substitutes in terms of *trade execution services*. However, they differentiate themselves by monopolistically providing *speed services*, such as colocation of information servers. As described in Budish et al. (2020), a trader needs to obtain speed services on exchange k in order to conduct ultra-fast trading on that exchange—e.g., colocating an information server to exchange 1's information center does not provide fast access to exchange 2. To make the model as simple as possible, I assume that routing an order from one exchange to another takes a deterministic delay $\Delta > 0$ where Δ is relatively large compared to latency of HFTs so that I focus on the equilibrium where

[10] Innovations in v can be thought of as a new arbitrage opportunity triggered by a jump in the asset's price that is not yet reflected by prices of other highly correlated assets (Budish et al., 2015). They could also stem from some public news, such as Fed announcements and the release of government statistics, or execution of Intermarket Sweep Orders (Baldauf and Mollner, 2020).

HFTs place orders on exchange k if they want to execute orders on that exchange rather than routing from other exchanges.[11]

As Subsection 2.3 describes in detail, each exchange can randomly impose intentional delays on execution of liquidity-taking orders (e.g., asymmetric speed bumps). Exchanges are for-profit entities, and they endogenously choose the speed fees and the intensity of random delays to maximize their profits.

2.2 Traders and trading speed

A liquidity trader. There is a risk neutral liquidity trader who represents (slow) passive investors with no material information in the real world. Her behavior stems from some exogenous reasons, such as hedging motives and margin constraints.

Upon hit by a private-value shock, a liquidity trader exogenously submits a single-unit buy or sell market order with equal probability.[12] Since the private-value shock does not trigger a race between sniping and cancellation of limit orders by HFTs (see below), the liquidity trader can take liquidity for sure. Thus the liquidity trader is agnostic about her trading speed.

Moreover, following the U.S. trading rules, a market order posted on any exchange is matched with the limit order that proposes the best price across N exchanges. Therefore, the liquidity trader is indifferent between posting a market order on any exchanges upon hit by the shock.

High-frequency traders. There are two risk-neutral *high-frequency traders (HFTs)* indexed by i and j. Each HFT serves both as a market maker (a high-frequency market maker; an HFM) and a liquidity taker (a high-frequency sniper; an HFS). I assume that the trading game occurs in a very short interval and, as in Baldauf and Mollner (2020), HFTs are prohibited from sending multiple market orders in any two periods that are infinitely close.[13] Their behavior involves the following steps.

Step 1. At time $t = 0$ (prior to the trading game), each HFT decides on her trading speed, such as

[11] See Budish, Lee and Shim (2020) for a similar setting. In their model, an HFT cannot participate in a race on exchange k if she does not subscribe to exchange k's speed services, i.e., routing an order from other exchanges to exchange k is not fast enough to compete in a race on exchange k.

[12] I assume that the liquidity trader takes only one unit of the asset. The unit-trading assumption is widely used in the literature (see, for example, Glosten and Milgrom, 1985) and can be thought of as a capacity constraint of a liquidity trader or the size/magnitude of a private-value shock.

[13] Without this assumption, HFTs may send redundant orders simultaneously in the expectation that one of them hits a limit order faster than other orders. This behavior complicates the discussion without adding new implications.

high-bandwidth connectivity to exchanges and colocation services.[14] In reality, speed services are obtained via monthly or annual subscriptions and cannot be adjusted during a given trading game (see IEX, 2019).

In the model, HFT i purchases those speed services from exchanges at $t = 0$ and they are denoted as $\phi_i = (\phi_{i,k})_{k\in\mathcal{E}}$. Exchange k charges speed fee $p_k \geq 0$ per unit of speed services. ϕ_i determines HFT i's speed of access to the market. In particular, it takes stochastic time $\tau_{i,k} \sim \exp(\phi_{i,k})$ for HFT i to execute her market order/cancellation request on exchange k. Put differently, the expected time between order placement and execution is inversely proportional to $\phi_{i,k}$.[15]

Step 2. At time $t = 1$, each HFT posts a single-unit limit order as an HFM.[16] At that time, HFM i chooses (i) on which exchange she posts a limit order and (ii) at what prices she quotes for one unit of the asset. The bid and the ask prices are denoted as $(-s_i, s_i)$, and s_i is referred to as the (half) bid-ask spread.

Step 3. The current state of limit order books is not observable *per se*, and each HFT needs to monitor N exchanges to promptly react to the arrival of news. In reality, direct data feeds from an exchange provide raw information about the exchange's limit order book, and a trader needs to process them to observe the current state of the book and to locate profitable trading opportunities.

In the model, HFT i possesses one unit of perfectly divisible monitoring capacity (or information processors) and allocates $\tilde{q}_{i,k} \in [0, 1]$ fraction to exchange k.[17] By allocating $\tilde{q}_{i,k}$, HFT i can observe the limit order book of exchange k at the timing of a common-value shock with probability $\tilde{q}_{i,k}$. It allows her to immediately react to news by sending a market order to the exchange. With the complementary probability, $1 - \tilde{q}_{i,k}$, the HFT cannot promptly react to the arrival of news, and her sniping probability on exchange k (defined below) is discounted by $\alpha \in (0, 1]$.[18]

$\tilde{q}_{i,k}$ can be seen as the speed of information processing in a reduced form. For example, it could be the probability that an HFT can complete analyzing direct data feeds from exchange k to recon-

[14] The *ex-ante* choice of speed is in line with the literature, e.g., Foucault et al. (2013) and Brolley and Zoican (2020).

[15] See, for example, Foucault, Kozhan and Tham (2016) for a similar setting.

[16] An HFM does not provide quote for more than on unit of the asset, as the liquidity trader takes only one unit of the asset.

[17] In Appendix B, I consider a more general setting where HFT i chooses her monitoring capacity h_i before the trading game and allocates $\tilde{q}_{i,k}$ fraction of it to monitor exchange k. As long as the cost of acquiring h_i is relatively small, which is the case in the real financial market with multiple trading rounds, the implications of the main model stay the same.

[18] I take α as a given parameter, but Appendix B derives α from some additional delays in reaction time that follow a Poisson distribution.

struct its limit order book before the arrival of a shock. If she does not monitor the exchange based on its direct data feeds, she must trade based on consolidated data (or the SIP data), which is slow compared to trading with direct data feeds and a market maker's cancellation request.[19]

Step 4. Finally, if a common-value shock hits, both HFTs immediately observe the realized value of $v = \pm\sigma$, and a sniping race takes place. This setting allows us to focus on latency arbitrage, i.e., arbitrage opportunities that stem from speed asymmetry, rather than that from information asymmetry.

Upon observing v, HFT i places a cancellation request of her limit order to revoke her stale quote and to avoid being picked off. At the same time, she tries to send a single-unit market order to snipe her rival's limit order, which is immediate depending on her allocation of monitoring capacity, $\tilde{q}_{i,k}$, as defined in Step 3.

Remark. When sniping a limit order, HFTs take a pure strategy, i.e., they try to place a market order immediately upon the arrival of a common-value shock. This strategy differs from that of a strategic liquidity taker in Back and Baruch (2004) where an informed taker adopts a mixed strategy and randomizes the timing of her order placement. This is because the informed trader in Back and Baruch (2004) needs to hide her private information. In contrast, HFTs in my model face no asymmetric information at the timing of a race, as they know the realized value of v. Thus randomizing (or intentionally delaying) the timing of order placement is suboptimal. Instead, the following analyses show that HFTs may take a mixed strategy in their decisions prior to a race.

2.3 Intentional delays in order execution

A typical exchange imposes delays on execution of market orders.[20] They reflect the implementation of speed bumps in the real markets (see Table 1) where most delays are *asymmetrically* imposed only on liquidity-taking orders.

To capture the random delays, I allow exchange k to impose intentional delays with frequency $\delta_k \in [0, 1]$. If the delays are imposed, a market order fails to snipe liquidity with probability $\beta \in (0, 1)$. This implies that, on exchange k, the expected probability of successful sniping for an HFT (defined below) is discounted by

$$\lambda(\delta_k) = 1 - \delta_k + \delta_k(1 - \beta) \in [1 - \beta, 1]. \tag{1}$$

[19] By construction, a market maker knows on which exchange she posts her limit order and does not need to process data to send cancellation request.

[20] Liquidity-providing orders, as well as cancellation requests, are not delayed.

To further interpret the above setting, some background might be useful. In the real financial market, the U.S. trading rule requires all exchanges to be "immediately accessible (Regulation NMS; Rule 611)" and to function as one consolidated exchange.[21] The imposition of long intentional delays goes counter to this notion, and it must be sufficiently short and regarded as *de minimis* by SEC to be approved.

In light of this, we can think of λ as a result of the random imposition of relatively short delays, where δ_k is the probability that delays are imposed. Alternatively, δ_k can also be seen as the stochastic length of delays. In both cases, $1-\beta$ is the probability of successful sniping when a market order is delayed, where $\beta < 1$ captures the fact that everything can be stochastic on the microsecond/nanosecond timescale due to randomness in information processing by traders and order handling at exchanges' matching engine or gateways.[22] In what follows, δ_k is referred to as the intensity of intentional delays.

Moreover, I assume that the imposition of delays takes an *ex-ante* fixed cost, denoted by $C > 0$. It captures operational and administrative costs, e.g., an exchange must obtain the SEC approval, which is time consuming and requires human resources.[23]

Prior to the trading game, exchange k determines the intensity of delays, δ_k, and the fee for speed technologies, p_k, to maximize her fee revenues from supplying speed services.

2.4 Equilibrium

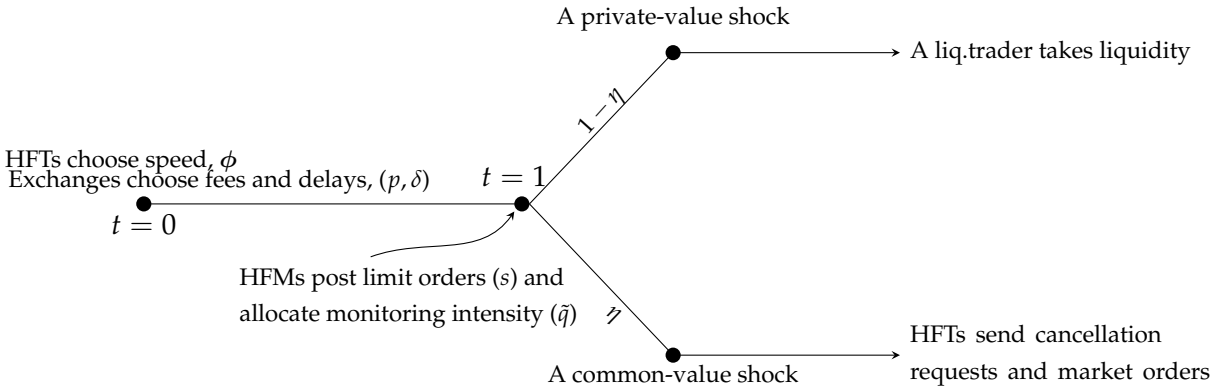
The model is conceptualized as a sequential game with three stages, and the equilibrium concept is the subgame perfect equilibrium. Figure 1 visualizes the timeline of the game. In the first stage ($t = 0$), all exchanges simultaneously determine the price of speed services, p_k, and the intensity of intentional delays, δ_k, to maximize their fee revenues, and HFTs decide on their trading speed, ϕ, by purchasing speed services. In the second stage ($t = 1$), HFTs post limit orders by choosing their trading venue from N exchanges, and allocate monitoring capacity, $\tilde{q}$. Finally, traders move as specified in Subsection 2.2 following a trigger event: either a common-value or a private-value shock.

[21] Rule 611: https://www.sec.gov/spotlight/emsac/memo-rule-611-regulation-nms.pdf

[22] It is possible that a message sent by the winner of a race actually arrives to an exchange slightly later than the first loser's message but nevertheless can get processed first. Aquilina, Budish and O'Neill (2020) report that the probability of the above event is about 4%.

[23] I assume that operational costs after technology installations are ignorable. This is in line with the real financial market, as suggested by IEX (2019).

Figure 1: Timeline of the game



In what follows, I assume that all random variables are independent. Also, without loss of generality, I analyze how the ask prices (s_i, s_j) are determined when the asset's value experiences a positive innovation, $v = +\sigma$.[24]

2.5 Profit of a high-frequency sniper

Consider a profit of HFT i as a sniper (HFS i). If HFT j posts her limit order on exchange k, and there is no discount in the sniping probability, HFS i with speed $\phi_{i,k}$ can snipe her rival's limit order with the following probability:

$$\psi_{i,k} = \Pr(\tau_{i,k} < \tau_{j,k}) = \frac{\phi_{i,k}}{\phi_{i,k} + \phi_{j,k}}.$$

$\psi_{i,k}$ represents the intrinsic sniping probability of HFT i as a sniper to exchange k. It also captures the degree of adverse selection that HFT j faces as a market maker.

When HFS i allocates $\tilde{q}_{i,k}$ of her attention on exchange k, and the exchange imposes intentional delays with intensity δ_k, the overall sniping probability of HFS i is discounted and given by

$$q_{i,k}\lambda(\delta_k)\psi_{i,k} \tag{2}$$

where

$$q_{i,k} = \tilde{q}_{i,k} + \alpha(1 - \tilde{q}_{i,k})$$

denotes the expected discounted probability that HFS i can immediately react to the news, and $\lambda(\delta_k)$ is the impact of delays, given by (1). Thus, given that HFM j quotes s_j on exchange k and HFT i

[24]The model is symmetric around zero, and the symmetric argument gives the results for the bid side of the market.

allocates $\tilde{q}_{i,k}$ to monitor it, HFT i's expected profit as an HFS is given by

$$V_{i,k}^{HFS}(\tilde{q}_{i,k}, \phi_{i,k}) = q_{i,k}\lambda(\delta_k)\psi_{i,k}(\sigma - s_j). \tag{3}$$

2.6 Profit of a high-frequency market maker

Consider HFT i's profit as a market maker (HFM i) when her rival places a limit order on exchange l with ask and bid prices $(s_j, -s_j)$.

As in the canonical model by Glosten and Milgrom (1985), HFM i obtains positive profits in expectation when her limit order is taken by the liquidity trader. This is because liquidity trading implies no innovations in $\tilde{v}$, $\mathbb{E}[s_i - \tilde{v}] = s_i > 0$. Since the liquidity trader takes only one unit, however, only one of HFMs who proposes a better price can glean this profit opportunity, causing pricing competition. In this game, I impose the following tie-breaking rule.

Assumption 1. *If both HFMs post the same price, $s_i = s_j$, the liquidity trader's market order is matched with HFM i's limit order with probability $a_i \in (0, 1)$.*

Assumption 1 is consistent with the fact that the liquidity trader is indifferent between trading on all N exchanges, as her order is executed at the best price. Also, the following analyses show that Assumption 1 (i.e., the value of a_i) is irrelevant to the equilibrium result.

Suppose that HFM i posts her limit order on exchange k with prices $(s_i, -s_i)$, and HFT j allocates $\tilde{q}_{j,k}$ of her monitoring capacity to analyze exchange k. Then HFM i's expected profit from making market is given by

$$V_{i,k}^{HFM}(s_i, \phi_{i,k}) = (1 - \eta)\theta_i(s_i, s_j)s_i + \eta q_{j,k}\lambda(\delta_k)\psi_{j,k}(\sigma - s_i), \tag{4}$$

where θ_i represents the indicator function of trading with the liquidity trader conditional on the shock:

$$\theta_i(s_i, s_j) = \mathbb{I}_{\{s_i < s_j\}} + a_i\mathbb{I}_{\{s_i = s_j\}}.$$

The first term of (4) shows the case of liquidity trading. If HFM i proposes a better price than her rival ($s_i < s_j$), the liquidity trader's order is matched with HFM i's limit order. By the symmetric logic, HFM i cannot trade with the liquidity trader if $s_i > s_j$. The second term of θ_i captures the case with $s_i = s_j$ that Assumption 1 stipulates.

The second term of (4) captures the case that a trade is triggered by a common-value shock. With probability $q_{j,k}\lambda(\delta_k)\psi_{j,k}$, HFS j snipes HFM i's limit order, causing the adverse selection cost to HFM i, as $s_i - \sigma < 0$. Otherwise, HFM i cancels (or revises) her stale quote and avoids the cost.

Note that intentional delays lower the sniping probability of an HFS, i.e., λ declines with δ_k. In other words, the intentional delays protect a market maker against latency arbitrage and mitigate the adverse selection cost.

3 Equilibrium

In this section, I solve for the optimal behavior of HFTs given the market structure imposed by exchanges (i.e., prices of speed technologies and the intensity of delays). I take steps backward to solve for the equilibrium.

3.1 Liquidity provision and market monitoring

When HFT i decides on her behavior as a market maker and her allocation of monitoring capacity as a sniper, she has not yet observed her rival's strategy. Thus decision regarding these variables constitutes one subgame.

As analyzed by Dennert (1993) and Baruch and Glosten (2013, 2019), strategic behavior of HFTs leads to the following result.

Lemma 1. *In the liquidity-provision and monitoring stage, there exists no equilibrium in pure strategies.*

Appendix C provides formal proofs for all analytical results, but some brief explanation is provided here. Regarding the choice on the trading venue and market monitoring, intuition is quite clear. For example, if HFM i takes a pure strategy and posts her limit order on exchange 1 for sure, her rival allocates full monitoring intensity to exchange 1 as well. However, given that HFT j allocates full capacity to exchange 1, posting a limit order on exchange 1 is no longer optimal—HFM i finds it optimal to post on one of other exchanges. This argument implies that HFMs do not take pure strategies in their venue choice and market monitoring.

On the pricing decision, if both HFMs take pure strategies, the profit function as a market maker, $V_{i,k}^{HFM}(s_i, \phi_{i,k})$ in equation (4), involves a discontinuity at $s_i = s_j$. Namely, if HFM i slightly undercuts her rival, she can obtain strictly positive profits by attracting the liquidity trader with probability $1 - \eta$, whereas she loses this opportunity if her price is slightly higher than her rival's quote. Since

both HFMs are strategic and comprehend the above structure, they try to exploit the discontinuity to earn more, which in turn eliminates pure strategy equilibrium.

In light of Lemma 1, I search for an equilibrium where

(i) HFT i posts her limit order on exchange k with probability $m_{i,k} \in (0,1)$,

(ii) she randomizes her quote s_i on exchange k with cdf $F_{i,k}(s) \equiv \Pr(s_i < s)$ over $s \in [\underline{s}_{i,k}, \sigma]$, and

(iii) allocates $\tilde{q}_{i,k} \in (0,1)$ fraction of her monitoring capacity to exchange k,

where $\underline{s}_{i,k}$ denotes an endogenous lower bound of the feasible quote. As shown by Proposition 1, the above equilibrium is indeed unique.

When HFM j adopts the above strategies, HFM i's quote s_i becomes the best price and can attract liquidity trading with the following expected probability.

$$\begin{aligned}\mathbb{E}[\theta_i(s_i, s_j)] &= \sum_{k\in\mathcal{E}} m_{j,k} \Pr(s_i < s_j | \text{HFT } j \text{ posts on exchange } k) \\ &= 1 - \sum_{k\in\mathcal{E}} m_{j,k} F_{j,k}(s_i).\end{aligned}$$

Accordingly, the expected profit of HFM i from posting her limit order on exchange k is rewritten as

$$V_{i,k}^{HFM}(s_i, \phi_{i,k}) = (1-\eta)\left(1 - \sum_{k\in\mathcal{E}} m_{j,k} F_{j,k}(s_i)\right) s_i + \eta q_{j,k} \lambda(\delta_k) \psi_{j,k} (\sigma - s_i). \tag{5}$$

Moreover, the expected sniping profit of HFT i is

$$V_i^{HFS}(\tilde{q}_i, \boldsymbol{\phi_i}) = \eta \sum_{k\in\mathcal{E}} q_{i,k} m_{j,k} \lambda(\delta_k) \psi_{i,k} \left(\sigma - \mathbb{E}_{j,k}[s_j]\right) \tag{6}$$

where $\mathbb{E}_{j,k}[s_j] = \int_{\underline{s}_{j,k}}^{\sigma} s_j dF_{j,k}(s_j)$ denotes the expected spread posted on exchange k by HFM j.

Equilibrium monitoring. Firstly, the mixed strategy regarding venue choice requires that HFM i is indifferent between posting her limit order on all exchanges. Since HFT i's profit from sniping, V_i^{HFS} in (6), is independent of on which exchange she posts her limit order, the indifference condition is given by $V_{i,k}^{HFM}(s_i, \phi_{i,k}) = V_{i,l}^{HFM}(s_i, \phi_{i,l})$ for all $k, l \in \mathcal{E}$. This is equivalent to $q_{j,k}\lambda(\delta_k)\psi_{j,k} = q_{j,l}\lambda(\delta_l)\psi_{j,l}$

for all $k, l \in \mathcal{E}$, and $\sum_{u\in\mathcal{E}} q_{j,u} = 1$ leads to

$$\gamma_j \equiv q_{j,1}\lambda(\delta_1)\psi_{j,1} = \frac{1}{\sum_{k\in\mathcal{E}} \frac{1}{\lambda(\delta_k)\psi_{j,k}}}.$$

Adopting the above strategy, HFM i's profit in (5) is rewritten as

$$V_i^{HFM}(s_i, \boldsymbol{\phi}_i) = (1-\eta)\left(1 - \sum_{k\in\mathcal{E}} m_{j,k}F_{j,k}(s_i)\right)s_i + \eta\gamma_j(\sigma - s_i). \tag{7}$$

γ_j can be seen as the sniping probability of HFS j, as well as the adverse selection risk for HFM i, after incorporating HFS j's monitoring strategy, $\tilde{q}_{j,k}$. In what follows, γ_j is referred to as the compound sniping probability (resp. adverse selection cost) of HFT j (resp. for HFM i).

Moreover, equation (7) implies that HFMs are indifferent between all exchanges when they post a limit order, i.e., V_i^{HFM} is independent of index k. Therefore, for HFM i, randomizing her limit order by using a homogeneous distribution function is an equilibrium, leading to $F_{i,k} = F_{i,l} = F_i$ and $\underline{s}_{i,k} = \underline{s}_{i,l} = \underline{s}_i$ for all $k, l \in \mathcal{E}$ and both i and j.

Equilibrium bid-ask spread. The next step is to derive the equilibrium pricing strategy. As in Dennert (1993), sustaining the mixed strategy regarding the ask price requires HFM i to be indifferent between all feasible prices, $s_i \in [\underline{s}_i, \sigma]$. Since posting $s_i = \sigma$ is always feasible and provides zero profits, it must hold that $0 = V_i^{HFM}(\sigma) = V_i^{HFM}(s_i)$. By using (7), it implies that

$$F_j(s) = 1 - \frac{\eta}{1-\eta}\gamma_j\frac{\sigma - s}{s}.$$

Also, the lower bound of the feasible strategy for HFT j ($\underline{s}_i$) must make the RHS of the above equation zero, meaning that

$$\underline{s}_j = \frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}\sigma.$$

Moreover, the indifference condition for HFM i implies that her profits from market making shrink to zero in expectation.

Equilibrium venue choice. Finally, consider the allocation of monitoring capacity by HFT i, which in turn pins down the venue choice of her rival.

Since the expected profit from market making converge to zero in expectation, HFT i obtains the

following expected profit by allocating full monitoring capacity to exchange k conditional on the arrival of a common-value shock.

$$V_i^{HFS}(q_{i,k}=1,\phi_i) = m_{j,k}\lambda(\delta_k)\psi_{i,k}\left(\sigma - \mathbb{E}_j[s_j]\right).$$

For HFT i to take a mixed strategy regarding the monitoring intensity, she must be indifferent between analyzing all N exchanges, leading to

$$m_{j,k}\lambda(\delta_k)\psi_{i,k} = m_{j,l}\lambda(\delta_l)\psi_{i,l}, \ \ \forall k,l \in \mathcal{E}.$$

Solving the above equations for $k=1$ by using $\sum_{k\in\mathcal{E}} m_{j,k} = 1$ yields the equation below.

$$m_{j,1}\lambda(\delta_1)\psi_{i,1} = \frac{1}{\sum_{k\in\mathcal{E}} \frac{1}{\lambda(\delta_k)\psi_{i,k}}} = \gamma_i. \tag{8}$$

The following Proposition summarizes the strategy of HFTs in the trading game, and intuition is discussed in Subsection 3.2.

Proposition 1. *There is a unique mixed strategy equilibrium in the trading subgame where*
(i) *HFT i posts a limit order on exchange k with probability*

$$m_{i,k} = \Pr(\textit{HFM i posts on ex. } k) = \frac{1}{\sum_{l\in\mathcal{E}} \frac{\psi_{j,k}\lambda(\delta_k)}{\psi_{j,l}\lambda(\delta_l)}}. \tag{9}$$

(ii) *When posting a limit order, HFM i randomizes her quote s_i by using cdf F_i over $s_i \in [\underline{s}_i, \sigma]$ with*

$$F_i(s) = 1 - \frac{\eta}{1-\eta}\gamma_i\frac{\sigma - s}{s}. \tag{10}$$

and

$$\underline{s}_i = \frac{\eta\gamma_i}{1-\eta+\eta\gamma_i}\sigma$$

where

$$\gamma_i = \frac{1}{\sum_{k\in\mathcal{E}} \left(\lambda(\delta_k)\psi_{i,k}\right)^{-1}}.$$

(iii) *HFTs' expected profits from market making shrink to zero in expectation, $\mathbb{E}[V_i^{HFM}(s_i,\phi_i)] = 0$.*

(iv) *HFT i allocates $\tilde{q}_{i,k}$ fraction of monitoring capacity to analyze exchange k where*

$$\tilde{q}_{i,k} = \frac{1}{1-\alpha}\left(\frac{\gamma_i}{\lambda(\delta_k)\psi_{i,k}} - \alpha\right).$$

By using all the strategies derived above, the ex-ante expected profit for HFT i is given by the following.

Proposition 2. *The ex-ante expected profit of HFT i, net of the speed fees, is given by*

$$V_i(\phi) = \eta\sigma\gamma_i\left(1 + \frac{\eta}{1-\eta}\gamma_j \log\frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}\right) - \sum_{k\in\mathcal{E}} p_k\phi_{i,k}.$$

Proof. By using the indifference conditions and (8), the ex-ante expected profit of HFT i is rewritten as

$$V_i(\phi_i) = \eta\gamma_i(\sigma - \mathbb{E}_j[s_j]) - \sum_{k\in\mathcal{E}} p_k\phi_{i,k}. \tag{11}$$

On the other hand, equilibrium condition (10) implies that the following equation holds regarding the expected bid-ask spread.

$$\mathbb{E}_j[s] = -\sigma\frac{\eta}{1-\eta}\gamma_j \log\frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}. \tag{12}$$

□

3.2 Comparative statics: the expected bid-ask spread

From the *ex-ante* perspective, HFT i anticipates to pay the following expected spread when sniping her rival's limit order.

$$\bar{s}_j \equiv \sum_{k\in\mathcal{E}} m_{j,k}\mathbb{E}_{j,k}[s_j] = -\sigma\frac{\eta}{1-\eta}\gamma_j \log\frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}.$$

It has the following properties.

Proposition 3. *(i) The expected trading cost for HFT i is decreasing in the HFT i's speed level and increasing*

in her rival's speed level, i.e., for all $k \in \mathcal{E}$,

$$\frac{\partial \bar{s}_j}{\partial \phi_{i,k}} < 0 \text{ and } \frac{\partial \bar{s}_j}{\partial \phi_{j,k}} > 0.$$

(ii) The expected trading cost declines as exchanges impose delays more intensively, i.e., for all $k \in \mathcal{E}$,

$$\frac{\partial \bar{s}_j}{\partial \delta_k} < 0.$$

The above proposition concludes that the expected bid-ask spread that HFT j posts shrinks and market liquidity improves when HFT i becomes faster by purchasing more speed technologies, $\phi_{i,k}$. Since HFT i's speed is the source of the adverse selection cost for HFM j, it differs from the conventional results that positively relate adverse selection to the bid-ask spread either in a competitive environment (e.g., Glosten and Milgrom, 1985) or a strategic environment (Dennert, 1993 and Baruch and Glosten, 2013).

The difference from the existing models emanates from strategic liquidity provision and the dual role played by the HFTs. Firstly, strategic liquidity provision induces a market maker to quote the bid-ask spread that makes her rival break even. Importantly, this means that HFM i's quote positively reflects the adverse selection cost that HFM j faces, rather than the cost that she herself incurs. For example, if HFM i increases $\phi_{i,k}$, she faces less severe adverse selection, and her expected profit increases. To make HFM i break even, in turn, HFM j must quote a narrower spread to take the profitable liquidity trading away from HFM i. However, a speed-up by HFM i is the source of the cost for HFM j, as HFT i is also serving as a sniper. Thus HFM j quotes a narrower bid-ask spread in expectation, while adverse selection deteriorates.

Most of the existing models of market making deal with homogeneous market makers facing a common source of adverse selection i.e., informed trading relative to noise trading. In that case more intensive informed trading exacerbates adverse selection for all market makers, inducing all of them to quote wider bid-ask spreads. In contrast, market makers in my model have (potentially) heterogeneous speed levels and allow us to analyze how each trader's quote depends on different sources of adverse selection.

Comparison of points (i) and (ii) of Proposition 3 attests the above discussion: if a decline in the adverse selection cost happens due to a common source (intentional delays; a higher δ_k), it tightens the expected bid-ask spreads, as it mitigates the cost for all market makers. In contrast, if an idiosyn-

cratic factor, such as an HFT's speed, drives a change in adverse selection, the reaction of the bid-ask spread can be different across HFMs.

As discussed in Introduction, whether increases in the throughput of high-frequency trading widen the spread is inconclusive because the result depends on which one of takers and makers adopt speed technologies more intensively. My model proposes a new explanation for the ambiguity based on the heterogeneous sources of adverse selection and the HFTs' dual role in the market.

3.3 Equilibrium speed acquisition and the impact of intentional delays

As shown by Proposition 1, HFT's profit as a market maker shrinks to zero in expectation, and the *ex-ante* expected profit all comes from sniping her rival's quote. Thus at the speed-acquisition stage, HFT i solves the following problem.

$$\begin{aligned}\phi_i^* &= \arg\max_{\phi_i} V_i(\phi_i) \\ &= \arg\max_{\phi_i} \eta\sigma\gamma_i \left(1 + \frac{\eta}{1-\eta}\gamma_j \log\frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}\right) - \sum_{k\in\mathcal{E}} p_k\phi_{i,k}\end{aligned}$$

with

$$\gamma_i = \left(\sum_{k\in\mathcal{E}} \frac{\phi_{i,k}+\phi_{j,k}}{\lambda(\delta_k)\phi_{i,k}}\right)^{-1}.$$

The FOC with respect to $\phi_{i,k}$ is given by the following equation, and the SOC is satisfied.

$$\begin{aligned}(\phi_{i,k}): \frac{p_k}{\eta\sigma} = &\overbrace{\left(1 + \frac{\eta}{1-\eta}\gamma_j \log\frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}\right)}^{\text{the profit margin}} \overbrace{\frac{d\gamma_i}{d\phi_{i,k}}}^{\text{change in the sniping prob.}} \\ &+ \underbrace{\gamma_i\left(1 + \frac{1-\eta+\eta\gamma_j}{1-\eta}\log\frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}\right)\frac{\eta}{1-\eta+\eta\gamma_j}\frac{d\gamma_j}{d\phi_{i,k}}}_{\text{change in the trading cost}}.\end{aligned} \tag{13}$$

The LHS shows the normalized speed fee that exchange k charges, i.e., the exogenous cost of speed technologies for HFT i, normalized by the size of an arbitrage opportunity (σ) and its arrival frequency (η). In the RHS, the first term captures an increase in the sniping probability that is achieved by adopting faster speed technologies, i.e., $\frac{d\gamma_i}{d\phi_{i,k}} > 0$. It increases the overall expected profit, as a faster technology makes sniping easier given the profit margin. The second term in the RHS represents the impact of a speed-up on the trading cost, i.e., the expected bid-ask spread that her rival posts. As Proposition 3 attests, an increase in $\phi_{i,k}$ mitigates adverse selection for HFM i, induces HFM j to post

a narrower spread, and lowers the HFT i's trading cost. Note that this term arises because HFTs are strategic and incorporate the price impact of their decision on speed.

To understand the impact of delays on the optimal speed acquisition, I investigate comparative statics of the HFT i's sniping probability, as well as those of the adverse selection cost, i.e., γ_i and γ_j.

Lemma 2. *(i) For all $k \in \mathcal{E}$, the HFT i's sniping probability, γ_i, is an increasing function of $\phi_{i,k}$. Also, the marginal impact of a speed-up, defined by $|\frac{d\gamma_i}{d\phi_{i,k}}|$, is decreasing in δ_k if and only if $\left(1+\frac{\phi_{j,k}}{\phi_{i,k}}\right)\sum_{l\neq k}\lambda(\delta_l)^{-1}\psi_{i,l}^{-1} > \lambda(\delta_k)$.*

(ii) For all $k \in \mathcal{E}$, the adverse selection cost that HFT i incurs, γ_j, is a decreasing function of $\phi_{i,k}$. Also, the marginal impact of a speed-up, defined by $|\frac{d\gamma_j}{d\phi_{i,k}}|$, is decreasing in δ_k if and only if $\left(1+\frac{\phi_{i,k}}{\phi_{j,k}}\right)\sum_{l\neq k}\lambda(\delta_l)^{-1}\psi_{j,l}^{-1} > \lambda(\delta_k)$.

The above lemma shows that a more intensive imposition of delays has an ambiguous impact on the marginal effect of a speed-up. Mathematically, the chain rule leads to the following derivative of HFT i's sniping probability.

$$\frac{d\gamma_i}{d\phi_{i,k}} = \lambda(\delta_k)\underbrace{\frac{\phi_{j,k}}{(\phi_{i,k}+\phi_{j,k})^2}}_{=\frac{d\psi_{i,k}}{d\phi_{i,k}}}\underbrace{\left(\frac{1}{1+\lambda(\delta_k)\psi_{i,k}\sum_{l\neq k}\lambda(\delta_l)^{-1}\psi_{i,l}^{-1}}\right)^2}_{=\frac{d\gamma_i}{d\psi_{i,k}}} > 0. \tag{14}$$

Thus more intensive delays (i.e., a higher δ_k) have two competing effects on the above derivative, as visualized by Figures 2 and 3.

Firstly, a higher δ_k weakens the positive impact of $\phi_{i,k}$ on γ_i: even if HFT i acquires faster speed, intentional delays slow down her market order, weakening the impact of her speed-up on the overall sniping probability. It is represented by the fact that the first two terms in (14), $\lambda(\delta_k)\frac{d\psi_{i,k}}{d\phi_{i,k}}$, is decreasing in δ_k.

Secondly, however, an increase in δ_k makes it more valuable for HFT i to be faster. Intuitively, the positive impact of a speed-up on the sniping probability tends to saturate and diminish as the level of $\phi_{i,k}$ becomes high (Figure 2). However, with intentional delays, the impact of $\phi_{i,k}$ on the sniping probability decays slowly (Figure 3). They exogenously lower $\lambda(\delta_k)\psi_{i,k}$ and provide HFT i with an additional value to increase her speed. This is captured by the second term of (14), $\frac{d\gamma_i}{d\psi_{i,k}}$, which is decreasing in $\lambda(\delta_k)$.

As a result, the reaction of the marginal impact of a speed-up ($\phi_{i,k}$) on the sniping probability (γ_i)

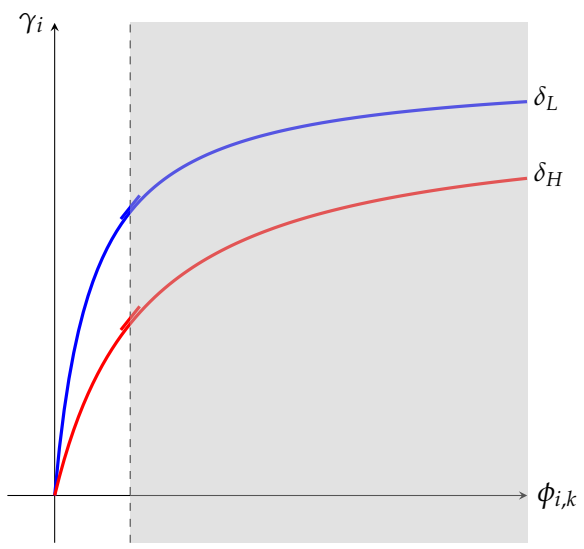


Figure 2: Reaction of γ_i to $\phi_{i,k}$ and δ_k

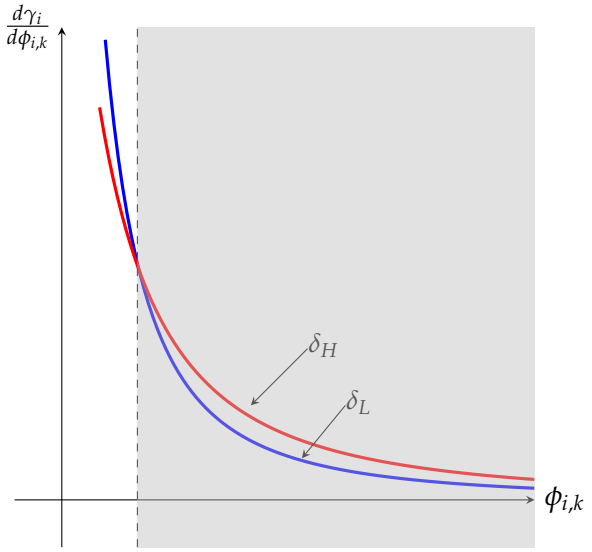


Figure 3: Reaction of $\frac{\partial \gamma_i}{\partial \phi_{i,k}}$ to $\phi_{i,k}$ and δ_k

Note: The left panel plots γ_i against $\phi_{i,k}$. The right panel plots the slope of γ against $\phi_{i,k}$. In both panels, the shaded area represents the region in which more intensive delays magnify the marginal impact of a speed-up on the sniping probability.

becomes hump-shaped against the intensity of delays (δ_k). The symmetric discussion can be applied to analyze the marginal impact of speed on the adverse selection cost, γ_j.

Since the adverse selection cost that HFT i faces determines her trading cost, the effect of intentional delays on the price impact of speed exhibits the similar ambiguity.

Lemma 3. *More intensive delays strengthen the (negative) price impact of $\phi_{i,k}$ if and only if*

$$\left(\frac{1-\eta}{1-\eta+\eta\gamma_j}\right)^2 < -\left(1-\lambda(\delta_k)\psi_{j,k}\sum_{l\neq k}\lambda(\delta_l)^{-1}\psi_{j,l}^{-1}\right)\left(\log\frac{\eta\gamma_j}{1-\eta+\eta\gamma_j}+\frac{1-\eta}{1-\eta+\eta\gamma_j}\right). \tag{15}$$

Lemma 3 means that, if the above inequality holds, HFT i can reduce her expected trading cost more by increasing her speed level when intentional delays are imposed more intensively. Together with Lemma 2, the above result leads to the possibility that the imposition of intentional delays promote HFTs' incentive to acquire faster speed technologies, as they increase the marginal benefit of adding one unit of speed.

3.4 Demand for speed in the symmetric equilibrium

Now, I focus on the symmetric equilibrium where both HFTs choose the same level of trading speed on each exchange, i.e., $\phi_{i,k} = \phi_{j,k} = \phi_k$ for all $k \in \mathcal{E}$. The equilibrium speed is given by solving the FOC:

Proposition 4. *In the symmetric equilibrium with intentional delays $\boldsymbol{\delta} = (\delta_k)_{k\in\mathcal{E}}$, HFTs obtain the following speed technologies from exchange k:*

$$\phi_k = \frac{1}{4p_k}\frac{\eta\sigma}{\lambda(\delta_k)}\frac{(1-\eta)g^2(\boldsymbol{\delta})}{1-\eta+\frac{\eta}{2}g(\boldsymbol{\delta})} \tag{16}$$

with

$$g(\boldsymbol{\delta}) = \frac{1}{\sum_{l\in\mathcal{E}}\lambda(\delta_l)^{-1}}. \tag{17}$$

The above equation represents the HFTs' demand for speed on exchange k. It is negatively proportional to the level of the speed fee that exchange k charges, linearly increasing in the volatility of the asset (σ), and exhibits an ambiguous reaction to other parameters.

Corollary 1. *The demand for speed technologies ϕ_k takes a U-shaped curve against the frequency of a common-value shock η with a unique tipping point given by*

$$\hat{\eta} = \frac{1}{1+\sqrt{\frac{1}{2}g(\boldsymbol{\delta})}}.$$

On the one hand, a higher probability of a common-value shock implies that HFTs are more likely to face a profitable sniping opportunity. It dwarfs the speed fees charged by exchanges and promotes HFTs' speed acquisition. On the other hand, a trade is more likely to involve adverse selection for market makers. Since η is a common source of adverse selection, it leads to a wider bid-ask spread and a smaller profit margin of sniping. The latter channel makes HFTs more reluctant to pay the fees to increase their speed. As a result of the above two competing effects, the impact of η on ϕ_k becomes non-monotonic.

The next section analyzes the impact of changing the delay intensity on the demand for speed and asks whether and how many exchanges optimally introduce them in the equilibrium.

4 Do for-profit exchanges introduce delays?

As discussed in Introduction, exchanges in the real market start adopting intentional delays (see Table 1 and Appendix A). On the one hand, the rationale for the imposition of delays is to protect market makers against latency arbitrage and improve market liquidity. It is based on the notion that delays can reduce the speed of HFTs by directly hampering sniping and by indirectly curtailing their

profit and an incentive to be faster. On the other hand, the theoretical study by Budish et al. (2020) suggests that, when the cost of adopting delays are sufficiently small, for-profit exchanges have no incentive to adopt delays. The following discussion reconciles the above discussion by Budish et al. (2020) to the speed bumps in the real financial market by showing that exchanges introduce delays (i.e., $\delta > 0$) in the equilibrium.

4.1 The optimal intensity of intentional delays

Remember that each exchange tries to maximize her fee revenues by controlling the level of the speed fee, p_k, and the intensity of intentional delays, δ_k. Proposition 4 implies that the speed fee is irrelevant to the exchange's profit, as the demand for speed is inversely proportional to p_k. Thus, any p_k can be an equilibrium and it is indeterminate. In reality, the price for speed services is skyrocketing (see, for example, IEX, 2019 and Glosten, 2020). Since the main focus of my model is on the intentional delays, I will leave the equilibrium determination of p_k and its analyses as a topic of future research.

By contrast, a for-profit exchange can affect her profit by controlling the intensity of delays, as it has a disproportional impact on the demand for speed. Exchange k solves the following problem.

$$\begin{aligned}\delta_k^* &= \arg\max_{\delta_k\in[0,1]} 2\phi_k p_k - C\mathbb{I}_{\{\phi_{i,k}>0\}} \\ &= \arg\max_{\delta_k\in[0,1]} 2\frac{\eta\sigma}{\lambda(\delta_k)}\frac{(1-\eta)g^2(\boldsymbol{\delta})}{1-\eta+\frac{\eta}{2}g(\boldsymbol{\delta})} - C\mathbb{I}_{\{\phi_{i,k}>0\}}\end{aligned}$$

where g is given by (17), and C is the fixed cost of adopting delays with $\mathbb{I}_X$ being the indicator function for X. I consider the exchange k's optimal strategy given that her rival exchanges take the symmetric strategy, i.e., $\delta_l = \delta_h = \delta_{-k}$ for all $l, h \neq k$.

Proposition 5. *(i) The demand for speed services on exchange k takes a single-peaked reaction to the intensity of intentional delays on that exchange with a unique maximizer being*

$$\delta_k^* \equiv \frac{1}{\beta}\left(1 - \frac{\lambda(\delta_{-k})}{N-1}\sqrt{\frac{1-\eta}{1-\eta+\frac{\eta}{2}\frac{\lambda(\delta_{-k})}{N-1}}}\right) \in (0,1].$$

Therefore, the optimal intensity of delays for exchange k is given by $\delta_k = \delta_k^*$.

(ii) The demand for speed services on exchange k is monotonically decreasing in the intensity of delays on other exchanges, i.e., $\frac{d\phi_k}{d\delta_{-k}} < 0$.

(iii) The exchange k's optimal intensity of delays is increasing in the intensity of delays on other exchanges,

i.e., $\frac{d\delta_k^*}{d\delta_{-k}} > 0$.

The imposition of intentional delays on exchange k has an ambiguous impact on the demand for speed services on her own platform. This is the reflection of Lemmas 2 and 3, i.e., the hump-shaped reaction of the marginal impact of a speed-up on the sniping probability and on the adverse selection to more intensive delays. When delays become longer, they can generate additional room for HFTs to increase the speed level to boost the sniping probability. It may also make it marginally more valuable to add speed to reduce the bid-ask spread.

Moreover, we can think of the result in point (iii) as the strategic complementarity between for-profit exchanges. When rival exchanges impose longer intentional delays, it reduces the level of demand for speed on exchange k. However, the demand function tilts toward right and exchange k can increase the demand by imposing marginally more intensive delays.

Due to the strategic complementarity, exchanges try to impose delays with their full capacity in the symmetric equilibrium.

Proposition 6. *In the symmetric equilibrium, the intensity of delays converges to a corner solution, i.e., the equilibrium intensity is given by* $\delta_k = \delta^* = 1$ *for all* $k \in \mathcal{E}$.

The above result shows that, conditional on paying the fixed cost for delays C, each for-profit exchange fully leverages her capacity to delay order execution. Note that it leads the sniping probability to be $\lambda(1) = 1 - \beta$.

Remark. The possibility of more intensive delays increasing the demand for speed is absent in the models with competitive HFTs and those with a binary choice on speed. In the case of competitive HFTs, delays may narrow the bid-ask spread and makes it more valuable to increase the sniping probability. This effect, however, is not enough to overturn the negative impact of delays on speed acquisition via a reduced sniping probability and a more salient cost of speed (i.e., fees). The HFTs' strategic behavior generates the positive impact of delays on the demand for speed via the slope of the bid-ask spread in Lemma 3, which helps overturn the existing result.

4.2 Entry decision of for-profit exchanges

This subsection endogenizes the number of exchanges, N, and characterizes how many exchanges may enter the financial market with intentional delays.

At the symmetric equilibrium, each exchange expects to earn the following profits from supplying speed services, net of the fixed cost of delays.

$$\Pi_k = \Pi = \frac{\eta\sigma}{2}\frac{\beta}{N^2}\frac{1-\eta}{1-\eta+\frac{\eta}{2}\frac{\beta}{N}} - C$$

Obviously, the above profit function is decreasing in the number of entrants, N, and converges to $-C$ as $N \to \infty$. Whenever N satisfies $\Pi > 0$, some exchanges may introduce intentional delays and enter the market, increasing N. This process continues until it holds that $\Pi \leq 0$. In contrast, if $\Pi < 0$, some exchanges stop operating. It reduces N and increases Π until $\Pi \geq 0$ holds. Therefore, the equilibrium $N = N^*$ is determined by the break-even condition and is stable. As long as condition () below is satisfied, a unique interior solution for $\Pi = 0$ exists:

$$\frac{\sigma\beta}{8}\frac{\eta(1-\eta)}{1-\eta+\eta\frac{\beta}{4}} > C. \tag{18}$$

If the cost is large and violates the above condition, no exchanges are willing to enter the market. Thus the following discussion assumes that (18) holds.

Proposition 7. *(i) The equilibrium number of for-profit exchanges in operation with intentional delays is given by*

$$N^* = \lfloor N_{BE} \rfloor,$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, and N_{BE} is given by

$$N_{BE} = \frac{\eta\beta}{4}\frac{\sqrt{1+\frac{8\sigma}{C\eta\beta}(1-\eta)^2}-1}{1-\eta}.$$

(ii) N^* is (weakly) increasing in β and σ and (weakly) decreasing in C.

In the equilibrium, a larger number of exchanges adopt intentional delays when the cost of adopting delays is small, the market is more volatile, and the delays are more likely to prohibit sniping. The reaction of N^* to β and σ is predictable, as both of they make HFTs' demand for speed more responsive to the intensity of delays by magnifying the profitability of a speed-up.

The impact of the cost (C) to the adoption of delays is opposite to that suggested by Budish et al. (2020). In my model, even if the cost is small and all exchanges adopt delays, delaying order execution can generate additional profits by boosting the demand for speed technologies. Thus the

smaller the cost, the more exchanges attempt to introduce the delays.[25]

4.3 Market quality

How does the behavior of strategic HFTs and for-profit exchanges affect the measure of market quality, such as liquidity and price discovery? Do intentional delays can achieve its primary purpose? This subsection provides an answer to the above questions.

Liquidity. As in the literature, the expected bid-ask spread can be used as the metric to gauge market liquidity. After incorporating the speed and delays in the symmetric equilibrium, the expected bid-ask spread is given by

$$\bar{s} = -\sigma \frac{\eta}{1-\eta} \frac{1-\beta}{N^*} \log \frac{\eta \frac{1-\beta}{N^*}}{1-\eta+\eta \frac{1-\beta}{N^*}},$$

where the equilibrium number of exchanges, N^*, is given by Proposition 7.

Firstly, in the symmetric equilibrium, the impact of delays on HFTs' speed acquisition does not matter because all HFTs end up having the same level of speed and the intrinsic sniping probability becomes constant, $\frac{\phi_{i,k}}{\phi_{i,k}+\phi_{j,k}} = \frac{1}{2}$, as in Menkveld and Zoican (2017) and Budish, Cramton and Shim (2015). Therefore, only the direct impact of intentional delays remains effective, and the adverse selection problem is mitigated for market makers.

Moreover, the number of exchanges, N^*, negatively affects the bid-ask spread. When a large number of exchanges are operating in parallel with each other, it becomes harder for HFTs to concentrate their monitoring capacity to analyze a certain exchange. It means that each market receives dispersed monitoring attention and dissipates the market makers' risk of being picked off by snipers.

Price discovery. The price discovery process is another important metric to evaluate market efficiency. It represents how quickly the equilibrium price impounds material information about the value of an asset. Conditional on a jump in $\tilde{v}$, the standing limit order with stale bid and ask prices reflects material information if (i) an HFT snipes her rival's limit order or (ii) reprices her stale limit order.[26]

Firstly, denote the latency of price discovery on exchange k when HFT i is serving as a sniper by

[25] The result is robust even if the market has M other exchanges who stick to the market structure with no delays.

[26] In the model, sending a cancellation request is equivalent to sending a repricing request, as there is no asymmetric information, and the model focuses on a one-shot trading game.

$T_{i,k}$. It has the following distribution.

$$T_{i,k} \sim \begin{cases} \min\{\tau_{i,k}, \tau_{j,k}\} & \text{with prob. } q_{i,k}\lambda(\delta_k), \\ \tau_{j,k} & \text{with prob. } 1 - q_{i,k}\lambda(\delta_k), \end{cases}$$

with the expected time being

$$\mathbb{E}[T_{i,k}] = \begin{cases} \frac{1}{\phi_{i,k}+\phi_{j,k}} & \text{with prob. } q_{i,k}\lambda_k, \\ \frac{1}{\phi_{j,k}} & \text{with prob. } 1 - q_{i,k}\lambda_k. \end{cases}$$

With probability $q_{i,k}\lambda_k$, HFT i can immediately react to the arrival of news and intentional delays do not prohibit her sniping, leading to competition with no discounts. In this case, price discovery is triggered by HFS i with latency $\tau_{i,k}$ or HFM j with latency $\tau_{i,j}$. With the complementary probability, HFM j can reprice prior to being picked off, which happens with latency $\tau_{j,k}$.

By aggregating across traders and exchanges, the latency T for price discovery has the following expected value.

$$\bar{T} = \mathbb{E}[T] = \sum_{k\in\mathcal{E}} \sum_{l=i,j} m_{l,k}\mathbb{E}[T_{l,k}].$$

The value of $\bar{T}$ is easy to compute in the symmetric equilibrium.

Corollary 2. *In the symmetric equilibrium, information on $\tilde{v}$ is reflected by the price with the following expected latency.*

$$\bar{T} = 2\left(1 - \frac{\lambda(1)}{N^*}\right)\frac{1}{N^*}\sum_{k\in\mathcal{E}^*}\frac{1}{\phi_k^*} \tag{19}$$

Due to the indeterminacy of the speed fees, the price discovery process is not fully characterized. However, equation (19) provides several observations. Firstly, intentional delays may have two competing effects on the price discovery process. On the one hand, they directly slow down the process by hampering HFTs' sniping behavior. This channel is captured by the fact that more intensive delays ($\delta \to 1$) increase $\bar{T}$, i.e., it takes longer for the price to incorporate information. On the other hand, it can facilitate the trading speed of HFTs, as suggested by Proposition 5. It shortens $\bar{T}$ by allowing HFTs to inject information into price both as snipers and market makers.

Moreover, the number of exchange, N^*, matters. Firstly, it increases $\bar{T}$ via the first term. This is because a larger number of exchanges dissipates monitoring attention of snipers, making a race less

likely to happen. In contrast, a larger N^* has ambiguous impact on the second component, i.e., the average latency of HFTs, given by $\frac{1}{N^*}\sum_{k\in\mathcal{E}}\frac{1}{\phi_k^*}$. If an increase in the number of exchanges is triggered by an exchange with low p_k and high ϕ_k, the addition of an exchange facilitates price discovery, and vice versa.

5 Discussion

5.1 Policy implication

In reality, exchanges have introduced intentional delays in the expectation that "[delays] will facilitate for passive liquidity providers an increased likelihood for of interacting with active orders of natural investors, while protecting against opportunistic, latency sensitive active strategies" (TSX, 2014). That is, delays are expected to mitigate adverse selection for market makers and improve liquidity.

In contrast, theoretical study by Budish, Lee and Shim (2020) argues that exchanges will not introduce new innovations, such as FBAs and speed bumps, as they earn large portion of profits from supplying speed services to HFTs and innovations may hinder speed acquisition.

My model reconciles the above view by Budish, Lee and Shim (2020) and the adoption of delays in the real markets. As Proposition 5 attests, in the certain parameter region, intentional delays facilitate speed acquisition by HFTs and boost their demand for speed services. This is due to the strategic nature of HFTs and their speed acquisition in the continuous domain. It leads for-profit exchanges to introduce delays to earn more from providing speed services.

Furthermore, intentional delays always mitigate adverse selection in the symmetric equilibrium and improve liquidity,[27] achieving their purported target. At the same time, they are consistent with the profit maximization of exchange platforms. This could be a theoretical background for the fact that exchanges in the real market volunteer to introduce speed bumps even without SEC or other government entities imposing them as a government policy.

[27] Aoyagi (2018) shows that whether the adoption of delays mitigates adverse selection for market makers depends on parameters, such as the length of delays, costs of speed acquisition, and the volatility of the asset's value.

5.2 Robustness and limitations of the model

Although my paper provides a stylized model of limit order markets, my main result does not hinge on modeling choices. The key economic forces that drive the paper is the strategic nature of HFTs who choose speed technologies as a continuous choice variable.

5.2.1 A model with slow competitive market makers

What if liquidity is provided by competitive slow market makers? They quote the competitive bid-ask spread, as in the canonical Glosten and Milgrom (1985) model, which positively reflects the speed of an HFS and negatively reacts to the intensity of delays. It is easy to show that longer delays increase the sniper's marginal benefit of being faster by (i) increasing the profit margin and (ii) making the spread less responsive to a speed-up. Channel (i) is trivial, as the spread declines as the result of mitigated adverse selection. Channel (ii) is a version of Lemma 2: knowing that the sniping probability is discounted by the delays, market makers downplay the impact of sniper's speed-up on the adverse selection cost. The latter effect makes it easier for a sniper to increase her speed without adversely affecting the price of the asset. Thus, the reaction of the demand for speed to more intensive delays stays the same even if market makers are competitive and equipped with exogenous speed technologies.

5.2.2 Long-lived information

One of the limitations of my model is that it focuses on a one-shot trading game and abstracts away from long-lived private information a la Kyle (1985) and Back and Baruch (2004). However, the key implication of my model stays the same even if a strategic trader possesses long-lived private information. With long-lived information, a strategic liquidity taker incorporates the impact of her trading activity on the information revelation by the equilibrium price, and she holds back from trading by fully leveraging her private information. When delays are imposed, the price impact of informed trading weakens due to the same logic as Lemma 2. This implies that an informed trader can exploit her private information more aggressively, making faster information acquisition more valuable. To my knowledge, strategic market making with (potentially) heterogeneous speed is hard to incorporate into the Kyle-type environment.[28] Thus considering HFTs playing a dual role

[28]For the case of strategic market makers with no speed/information acquisition, see Bondarenko (2001) and Nishide (2006).

with endogenous speed acquisition in the Kyle-type environment is one of the topics for the future research.

5.2.3 Trader welfare

My model is not suitable for analyzing welfare implication due to the existence of the liquidity trader with exogenous trading motives. Of course, the model can incorporate discretionary liquidity traders, as in Admati and Pfleiderer (1988), and endogenize their participation into the market. In this situation, their trading surplus depends on the expected bid-ask spread, meaning that the adoption of intentional delays leads to higher welfare. They also generate a positive feedback effect that shows "liquidity begets liquidity." Namely, a larger set of active liquidity traders mitigate adverse selection for market makers, and the bid-ask spread shrinks. In turn, a narrower spread facilitates liquidity traders' participation even more. Thus the imposition of delays may improve liquidity via the endogenous reaction of liquidity traders as well.

In the long run, the liquidity traders may face a tradeoff between trading as quickly as possible at the positive bid-ask spread and waiting to trade at a narrower spread after HFTs learn information. In other words, their welfare depends on the adverse selection, which is a transfer of market makers' cost, and the price discovery. As discussed in Subsection 4.3, intentional delays not only reduce spread, but they can also promote price discovery by making HFTs faster to trade. Thus even in a long-run model, intentional delays are expected to improve liquidity trader welfare.

5.2.4 Intensive margin and extensive margin

My model focuses on the speed choice in the continuous domain, meaning that HFTs are choosing their intensive margin of speed. In reality, being an HFT may take a fixed investment cost, such as those to set up specific computers and to colocate information servers (see IEX, 2019). Since the intentional delays reduce the indirect profit of HFTs, they can confound exchanges' introduction of delays. However, I believe that the effect via fixed costs is limited. Although my model considers a one-shot trading game, HFTs in the real financial market can exploit their speed advantage for multiple trading rounds. As Aquilina, Budish and O'Neill (2020) estimate, the aggregate prize of latency arbitrage amounts to about GBP 60 million per year in the UK (and $5 billion across global equity markets). Since the fixed cost should be amortized over the entire trading opportunities, the impact of the effective fixed costs should be minimal.

6 Conclusion

This paper studies the model of strategic high-frequency traders with for-profit exchanges. The model characterizes the equilibrium behavior of HFTs who serve both as liquidity takers (snipers) and liquidity providers. In the equilibrium, HFTs take a mixed strategy by randomizing the venue to post a limit order, the degree of market monitoring, and the bid-ask spread. In contrast, they deterministically snipe liquidity, as the arbitrage opportunity arises due to asymmetric trading speed rather than asymmetric information. In this equilibrium, the expected bid-ask spread negatively reacts to a decline in adverse selection that stems from some common sources, such as more active noise trading and the imposition delays in order execution. However, if idiosyncratic factors, such as a speed-up by one HFT, worsen the adverse selection cost for a market maker, bid-ask spreads exhibit an ambiguous reaction, which depends on whose quote we are looking at and what is the source of adverse selection. This is because of the strategic liquidity provision, combined with the dual role played by HFTs.

The second part of the model considers intentional delays imposed by for-profit exchanges. Exchanges earn fee revenues from supplying speed services to HFTs. My model shows that the demand for speed can be an increasing function of the length of intentional delays, as HFTs try to compensate for the imposition of delays. In this situation, exchanges have an incentive to introduce delays to boost the demand for speed services. Therefore, my paper suggests that intentional delays not only improve liquidity by mitigating adverse selection but also increase exchanges' profits. Thus it provides one explanation for the real market where exchanges self-impose intentional delays even without government intervention.

References

Admati, Anat R and Paul Pfleiderer, "A theory of intraday patterns: Volume and price variability," *The Review of Financial Studies*, 1988, *1* (1), 3–40.

Ait-Sahalia, Yacine and Mehmet Saglam, "High frequency market making: Optimal quoting," *Available at SSRN 2331613*, 2017.

Aldrich, Eric M and Daniel Friedman, "Order protection through delayed messaging," Technical Report, WZB Discussion Paper 2018.

American, NYSE, "NYSE American to Eliminate Speed Bump, Restore Floor-Based Trading," January 2019.

Anderson, Lisa, Emad Andrews, Baiju Devani, Michael Mueller, and Adrian Walton, "Speed segmentation on exchanges: Competition for slow flow," Technical Report, Bank of Canada Staff Working Paper 2018.

Aoyagi, Jun, "Strategic Speed Choice of High-Frequency Traders under Speed Bumps," *SSRN Electronic Journal*, 2018.

Aquilina, Matteo, Eric B Budish, and Peter O'Neill, "Quantifying the high-frequency trading "arms race": A simple new methodology and estimates," *Chicago Booth Research Paper*, 2020, (20-16).

Back, Kerry and Shmuel Baruch, "Information in securities markets: Kyle meets Glosten and Milgrom," *Econometrica*, 2004, *72* (2), 433–465.

Baldauf, Markus and Joshua Mollner, "High-frequency trading and market performance," *The Journal of Finance*, 2020, *75* (3), 1495–1526.

Baruch, Shmuel and Lawrence R Glosten, "Flickering quotes," *Columbia University*, 2013.

_ **and** _ , "Tail expectation and imperfect competition in limit order book markets," *Journal of Economic Theory*, 2019, *183*, 661–697.

Boehmer, Ekkehart, Kingsley Fong, and Julie Wu, "International evidence on algorithmic trading," *SSRN Electronic Journal*, 2015.

Bondarenko, Oleg, "Competing market makers, liquidity provision, and bid–ask spreads," *Journal of Financial Markets*, 2001, *4* (3), 269–308.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, "High frequency trading and the 2008 short-sale ban," *Journal of Financial Economics*, 2017, *124* (1), 22–42.

Brolley, Michael and David A Cimon, "Order-flow segmentation, liquidity, and price discovery: the role of latency delays," *Journal of Financial and Quantitative Analysis*, 2020, *55* (8), 2555–2587.

_ **and Marius Zoican**, "Liquid speed: On-demand fast trading at distributed exchanges," *arXiv preprint arXiv:1907.10720*, 2020.

Budish, Eric, Peter Cramton, and John Shim, "The high-frequency trading arms race: Frequent batch auctions as a market design response," *The Quarterly Journal of Economics*, 2015, *130* (4), 1547–1621.

_ **, Robin Lee, and John Shim**, "A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?," *NBER Working Paper Series*, 2020, (No. 25855).

Chakrabarty, Bidisha, Jianning Huang, and Pankaj K Jain, "Effects of a Speed Bump on Market Quality and Exchange Competition," *Available at SSRN 3280645*, 2019.

Chen, Haoming, Sean Foley, Michael Goldstein, and Thomas Ruf, "The Value of a Millisecond: Harnessing Information in Fast, Fragmented Markets," *SSRN Electronic Journal*, 2017.

Dennert, Jürgen, "Price competition between market makers," *The Review of Economic Studies*, 1993, *60* (3), 735–751.

Du, Songzi and Haoxiang Zhu, "What is the optimal trading frequency in financial markets?," *The Review of Economic Studies*, 2017, *84* (4), 1606–1651.

Foucault, Thierry, Ailsa Roell, and Patrik Sandas, "Market making with costly monitoring: An analysis of the SOES controversy," *The Review of Financial Studies*, 2003, *16* (2), 345–384.

_ **and Christine A Parlour**, "Competition for listings," *Rand Journal of Economics*, 2004, pp. 329–355.

_ **, Ohad Kadan, and Eugene Kandel**, "Liquidity cycles and make/take fees in electronic markets," *The Journal of Finance*, 2013, *68* (1), 299–341.

_ **, Roman Kozhan, and Wing Wah Tham**, "Toxic arbitrage," *The Review of Financial Studies*, 2016, *30* (4), 1053–1094.

Glosten, Lawrence R, "Economics of the Stock Exchange Business: Proprietary Market Data," *Available at SSRN 3533525*, 2020.

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

Haas, Marlene and Marius Zoican, "Beyond the frequency wall: Speed and liquidity on batch auction markets," *SSRN Electronic Journal*, 2016.

Hendershott, Terrence, Charles M Jones, and Albert J Menkveld, "Does algorithmic trading improve liquidity?," *The Journal of finance*, 2011, *66* (1), 1–33.

Hu, Edwin, "Intentional access delays, market quality, and price discovery: Evidence from IEX becoming an exchange," *SEC Working Paper*, 2018.

IEX, "The cost of exchange services," Technical Report, IEX January 2019.

Jones, Charles M, "What do we know about high-frequency trading?," *Columbia University Working Paper*, 2013.

Khapko, Mariana and Marius Zoican, "Do Speed Bumps Curb Speed Investment? Evidence from a Pilot Experiment," *Evidence from a Pilot Experiment (March 12, 2019)*, 2019.

Kyle, Albert S and F Albert Wang, "Speculation Duopoly with Agreement to Disagree : Can Overconfidence Survive the Market Test?," *Journal of finance*, 1997, *52* (5), 2073–2090.

_ **and Jeongmin Lee**, "Toward a fully continuous exchange," *Oxford Review of Economic Policy*, 2017, *33* (4), 650–675.

Kyle, S Albert, "Continuous Auctions and Insider Trading," *Econometrica*, 1985, *53* (6), 1315–1335.

Liu, Wai-Man, "Monitoring and limit order submission risks," *Journal of Financial Markets*, 2009, *12* (1), 107–141.

Menkveld, Albert J, "The economics of high-frequency trading: Taking stock," *Annual Review of Financial Economics*, 2016, *8*, 1–24.

_ **and Marius A Zoican**, "Need for speed? Exchange latency and liquidity," *The Review of Financial Studies*, 2017, *30* (4), 1188–1228.

Nishide, Katsumasa, "Insider trading with imperfectly competitive market makers," *Working Paper, Kyoto University*, 2006, (85).

O'Hara, Maureen, "High frequency market microstructure," *Journal of Financial Economics*, 2015, *116* (2), 257–270.

Pagnotta, Emiliano S and Thomas Philippon, "Competing on speed," *Econometrica*, 2018, *86* (3), 1067–1115.

Table 1: Design of Speed Bumps

	Exchange	Date	Targets of delay	Length of delay
In operation	IEX	Octber 2013	All but pegged orders	350 microseconds
	Thomson Reuters*	June 2016	Non-cancellation	0-3 milliseconds
	Aequitas NEO*	March 2015	Liquidity takers	3-9 milliseconds
	TSX Alpha*	September 2015	Liquidity takers	1-3 milliseconds
	Eurex Exchange*	June 2019	Liquidity takers	1 or 3 milliseconds
	EBS Market*	July 2013	Liquidity takers	3-5 milliseconds
	ParFx*	March 2013	Liquidity takers	10-30 milliseconds
	Moscow Exchange*	April 2019	Liquidity takers	2-5 milliseconds
Proposed	CHX	Proposed	Liquidity takers	350 microseconds
	EDGA (Cboe)	Proposed	Liquidity takers	n/a
	NASDAQ OMX PHLX	Proposed	Liquidity takers	5 microseconds
	ICE Futures	Proposed	Liquidity takers	3 microseconds
	Interactive Brokers*	Proposed	Liquidity takers	10-200 milliseconds
	NYSE American**	July 2017	All but pegged orders	350 microseconds

Note: * indicates random speed bumps. **In December 2019, ICE announced removal of speed bumps from NYSE American based on their finding that speed bumps worsen liquidity and the trading share of the exchange. As of May 2020, exchanges with random speed bumps do not announce the distribution funcsion of random delays.

Roşu, Ioanid, "A dynamic model of the limit order book," *The Review of Financial Studies*, 2009, *22* (11), 4601–4641.

SEC, "REGULATION NMS," https://www.sec.gov/rules/final/34-51808.pdf August 2005.

Shkilko, Andriy and Konstantin Sokolov, "Every cloud has a silver lining: Fast trading, microwave connectivity and trading costs," *SSRN Electronic Journal*, 2016.

Smith, Robert Mackenzie, "Client list reveals HFT dominance on BrokerTec," *Risk*, 2015, *28* (10).

TSX, "Notice of Proposed Rule Amendments and Requests for Comments," 12 2014.

A The actual implementation of intentional delays

This section briefly describes the institutional details of speed bumps. Although the model in the main text is built to analyze asymmetric speed bumps rather than symmetric delays on IEX, starting with IEX would be helpful, as it is a precursor of all other speed bumps.

Symmetric speed bumps with deterministic delays

Who is protected? A symmetric and deterministic speed bump is first adopted by IEX in 2013 and followed by NYSE American (while NYSE American has decided to remove it based on NYSE, 2019).

It delays all incoming and outgoing orders by 350 microseconds. The type of orders protected by the speed bump is "Pegged Order." Pegged order is the type of non-displayed limit orders whose price is dynamically adjusted by reference to the national best bid and offer (NBBO).[29] Although IEX imposes a delay on incoming orders and outgoing information, the messaging of SIP-related information is not delayed. Thus, the price of pegged orders is dynamically adjusted by IEX with no delays. If the NBBO changes, a speed bump allows IEX to adjust the pegged orders, and HFTs cannot snipe them at the stale price.

Speed bump infrastructure All traders sending messages to IEX must enter the IEX's system from the Point of Presence (POP) in Secaucus, NJ. After entering via the POP, a message sent to IEX travels through a "coiled" fiber optic cable, which has a distance of 38 miles. After exiting the coil, the message travels an additional physical distance to the IEX trading system, located in Weehawken, NJ. Due to this travel distance, a message sent to IEX must incur 350 microseconds of additional travel time.

Asymmetric speed bumps

An asymmetric speed bump has been adopted by a growing number of exchanges. It delays all orders except liquidity-providing orders. Details in the implementation varies depending on exchanges. For example, a speed bump by Chicago Stock Exchange delays all orders except for visible limit orders from approved liquidity providers, while an exception in TSX Alpha is provided to visible Post-Only orders. Post-Only is the type of limit orders that is automatically rejected if it has a potential to cross a market and remove liquidity from the limit order book.

Empirically identifying the impact of speed bumps is not straightforward, as a speed bump typically comes with changes in other market structures and trading rules. For example, along with a speed bump, TSX Alpha sets a minimum size requirement for liquidity providing orders and adopts an inverted maker-taker fee structure. Due to these structural changes, market makers must pay

[29] There are three types of pegged-order: Primary Peg (P-Peg), Discretionary Peg (D-Peg), and Midpoint Peg (M-Peg). P-Peg and D-Peg orders are resting at one tick below or above the NBBO. P-Peg orders have discretion to trade at the NBBO, while D-Peg orders have discretion to trade up to the midpoint. M-peg orders stay and are traded at the midpoint of the NBBO and has a higher priority than D-Peg orders at the mid-point. Whether the discretion of each order is exercised is determined by the "IEX signal" that determines if the NBBO is volatile (i.e., "scrambling") by using a specific measure. The discretion is not exercised if the signal is "on," meaning that the bid-ask in the market is volatile.

some additional cost (in terms of monetary or risk exposure) in return for protection by a speed bump.

Randomness Randomizing the length of delay is expected to generate an additional benefit: it mitigates asymmetric information more effectively compared to a deterministic speed bump. The advantage of random speed bumps stems from a situation where an informed trader splits a large order and sends them to multiple exchanges, i.e., "sweep" or "sprayed" orders.

If a speed bump is deterministic, a trader can send split orders to multiple exchanges by adding or subtracting some time lags to synchronize the execution timing of her orders on all exchanges. The simultaneous execution of sweep orders is made possible by the smart order router (SOR) that calculates and predicts the execution timing of each order by incorporating the deterministic delay imposed by a speed bump.

For example, consider an informed trader who wants to fill a large order (say $1,000$ shares). Exchanges A has 300 shares available, and Exchange B has 500 shares. Thus, the informed trader may spray orders to both exchange to fill 1,000 shares. Suppose that it takes t_A and t_B to send and execute orders on Exchanges A and B, respectively. Now, a speed bump is applied in Exchange B. If the length of delay δ is deterministic, the informed trader can stagger the timing or order entry to make $t_A = t_B + \delta$.[30] By the synchronized execution, there is no information leakage, and the trader can fulfill all orders.

Randomness in a speed bump makes it harder for the SOR to predict the execution timing, so that synchronizing executions of split orders does not really work. Due to the failure in the synchronized execution, market makers in Exchange B observe the execution of informed order on Exchange A before a part of split orders arrive at Exchange B. The time lag generated by a random speed bump allows the market makers to cancel or reprice their limit orders to avoid being picked off. Hence, liquidity providers bear less severe adverse selection than in the case with a deterministic speed bump.

B Extension

[30] In reality, there must involve some unexpected delays due to random factors, such as precipitation and temperature, and even a SOR cannot perfectly synchronize the order arrival timing.

B.1 Endogenous monitoring capacity

Suppose that HFT i can choose her monitoring capacity h_i by paying cost $K(h_i) = ch_i$ before the trading game starts. h_i with a linear cost can be seen as the number of computer nodes purchased for information processing or compensation for labor force. In the monitoring stage, she allocates $\tilde{q}_{i,k}$ fraction of h_i to monitor exchange k. By leveraging the monitoring capacity, HFT i can analyze the limit order book on exchange k with latency $\tau_{i,monitor} \sim \exp(h_i)$ with probability $\tilde{q}_{i,k}$, while it causes some additional delays $\delta_{monitor} \sim \exp(b)$ with probability $1 - \tilde{q}_{i,k}$. This implies that HFT i can observe the limit order book before the arrival of a common-value shock with the following probability.

$$\begin{cases} \Pr(\tau_{i,monitor} < \tau_c) = \frac{h_i}{z_c + h_i} & \text{with probability } \tilde{q}_{i,k}, \\ \Pr(\tau_{i,monitor} + \delta_{monitor} < \tau_c) = \frac{h_i}{z_c + h_i} \frac{b}{b + z_c} & \text{with probabiliry } 1 - \tilde{q}_{i,k}. \end{cases}$$

Note that adding some other delays with the exponential distribution does not change the result. By denoting $\alpha = \frac{b}{b+z_c}$ and setting $h_i \to \infty$, which is the case if $c = 0$, the above environment becomes the same as that in the main model.

Now, consider the general case with $h_i < \infty$ due to $c > 0$, and denote $x_i = \frac{h_i}{z_c + h_i}$. For simplicity, I assume that $b = 0$ and thus $\alpha = 0$. In this case, the equilibrium condition for q is still given by $q_{i,k}\lambda_k\psi_{i,k} = q_{i,l}\lambda_l\psi_{i,l}$ for all $k, l \in \mathcal{E}$ with

$$q_{i,k} = \tilde{q}_{i,k} x_i.$$

By using $\sum_{k \in \mathcal{E}} \tilde{q}_{i,k} = 1$, γ_i in the main text is replaced by

$$\gamma_i = \frac{x_i}{\sum_{k=1}^{N} \left(\lambda(\delta_k)\psi_{i,k}\right)^{-1}}. \tag{20}$$

Note that the rest of the model remains the same as the main text. The FOC regarding $\phi_{i,k}$ is the same as the main model with γ_i given by (20), and that for h_i is given by

$$(h_i) : \frac{c}{\eta\sigma} = \left(1 + \frac{\eta}{1-\eta}\gamma_j \log \frac{\eta\gamma_j}{1 - \eta + \eta\gamma_j}\right) \frac{d\gamma_i}{dh_i}.$$

Note that the SOC is satisfied for the above FOC.

As in the main model, consider the symmetric equilibrium where both HFTs choose the same level of monitoring capacity, $h = h_i = h_j$, and speed level on each exchange $\phi_k = \phi_{i,k} = \phi_{j,k}$ that lead

to $\gamma = \gamma_i = \gamma_j$. Also, denote $\Lambda = \sum_{k=1}^{N} \lambda(\delta_k)^{-1}$. The FOCs are reduced to

$$(\phi_k) : \frac{p_k}{\eta\sigma} = \frac{1}{\phi_k}\frac{h+z_c}{\lambda_k h}\gamma^2\frac{1-\eta}{1-\eta+\eta\gamma},$$

$$(h) : (z_c+h)^2 = \frac{z_c\eta\sigma}{2c}\frac{1}{\Lambda}\left(1+\frac{\eta\gamma}{1-\eta}\log\frac{\eta\gamma}{1-\eta+\eta\gamma}\right),$$

with

$$\gamma = \frac{1}{2\Lambda}\frac{h}{h+z_c}.$$

Thus I obtain two equations that determine h^* as a function of γ.

$$h = H_{foc}(\gamma,\Lambda) = \sqrt{\frac{z_c\eta\sigma}{2c}\frac{1}{\Lambda}\left(1+\frac{\eta\gamma}{1-\eta}\log\frac{\eta\gamma}{1-\eta+\eta\gamma}\right)} - z_c,$$
$$h = H_{def}(\gamma,\Lambda) = z_c\frac{2\Lambda\gamma}{1-2\Lambda\gamma}.$$

H_{foc} and H_{def} are derived from the FOC for h_i and the definition of γ, respectively. As long as c is sufficiently small and satisfies

$$c < \frac{\eta\sigma}{2z_c}\frac{1-\beta}{N}, \tag{21}$$

there is a unique set of interior solutions (h^*,γ^*) that solve the above two equations. Note that γ^* is monotonically decreasing in Λ. The demand for speed is given by

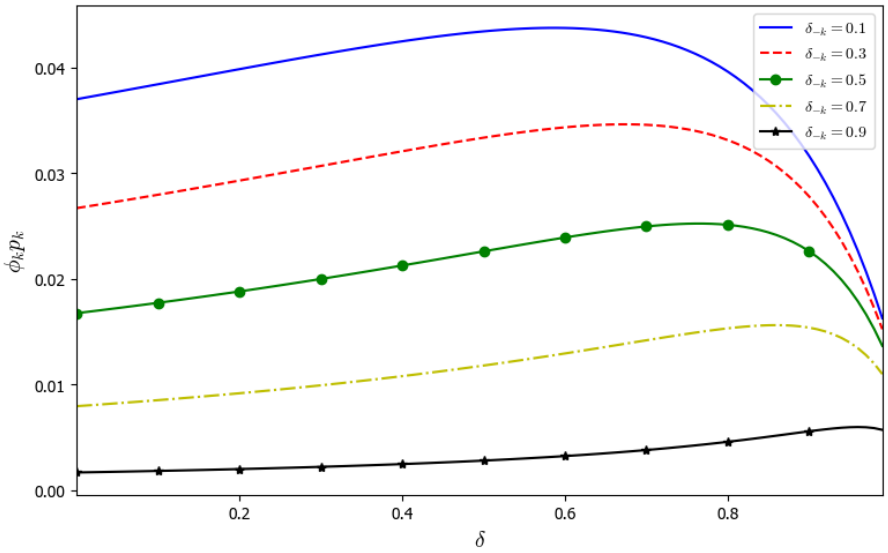
$$\phi_k = \frac{(1-\eta)\sigma}{2p_k\lambda_k}\frac{\eta\gamma^*}{1-\eta+\eta\gamma^*}\frac{1}{\Lambda}.$$

Suppose that all exchanges other than k adopt the same delays, meaning that $\lambda_l = \lambda_u = \lambda$ for $l,u \neq k$. Although the analytical solution is hard to obtain, numerical result in Figure 4 shows that the demand for speed takes a single-peaked curve against δ_k, and the optimal δ_k^* is an increasing function of the intensity of delays of rival exchanges. The result in Figure 4 is robust to a change in parameter values, as long as condition (21) is satisfied.

C Proof of Lemma 1 and Proposition 1

Since the derivation of the mixed strategies is provided in the text, the following proof focuses on the non-existence of equilibrium in pure strategy and the uniqueness of the mixed strategy equilibrium.

Figure 4: Demand for speed services and intentional delays



Note: This figure is illustrated by using the following parameter values: $c = 0.01, z_c = 1.0, \eta = 0.5, N = 3, \beta = 0.05$.

Incorporating the strategy regarding the venue choice, the expected profit for HFM i is given by

$$V_i^{HFM}(s_i, \phi_i) = (1-\eta)\underbrace{\big(1 - F_j(s_i)\big)}_{=\theta_i} s_i + \eta\gamma_j(\sigma - s_i).$$

Mixed strategy equilibrium

It is trivial that there is no pure strategy equilibrium in terms of the venue choice (m) and the decision on the monitoring intensity ($\tilde{q}$). Regarding the pricing decision, suppose that HFMs take a pure strategy. Also, define the break-even spread in the competitive environment with $F_j = 0$ as follows.

$$s_i^{BE} \equiv \frac{\eta\gamma_j}{1 - \eta + \eta\gamma_j}\sigma. \tag{22}$$

Note that $\underline{s}_i = s_j^{BE}$. Figure 5 draws HFM i's profit from market-making, V_i^{HFM}, as a function of her strategy, s_i. Figure 6 shows the best-response function of HFM l ($l = i, j$), denoted as s_l^*, to her rival's quote.

Figure 5 shows that it is optimal for HFM i to slightly undercut s_j as long as $s_j > s_i^{BE}$ because a better price attracts a profitable order flow from the liquidity trader. Thus, in Figure 6, the best-response price of HFM i for $s_j > s_i^{BE}$ is $s_i^* = s_j - \epsilon$ with $\epsilon \to +0$. Once s_j hits s_i^{BE}, however, quoting

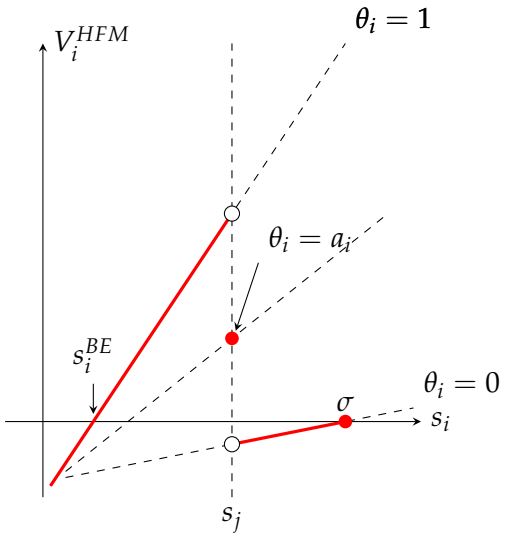


Figure 5: HFM i's profit

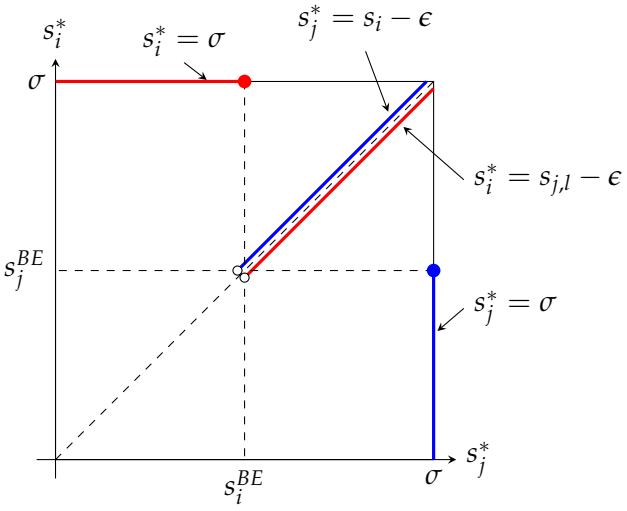


Figure 6: The best response

$s_i \in [0, \sigma)$ generates negative profits. Since placing $s_i = \sigma$ always grantees zero profit, HFM i's best response price jumps to $s_i^* = \sigma$. Symmetric arguments provide the best response of HFM j, denoted as s_j^* in Figure 6.

Figure 6 shows that price competition between strategic HFMs does not result in equilibrium in pure strategies. This is because HFMs comprehend how prices (s_i, s_j) affect their profit and try to exploit discontinuity at $s_i = s_j$.

Uniqueness

Firstly, suppose that HFM i puts a positive weight on $s_i = \sigma$. For $s_i = \sigma$ to obtain a positive weight, HFM j must charge prices above σ, which is not an equilibrium. Therefore, $s_i = \sigma$ cannot be an atom.

Secondly, suppose that HFM i puts positive wight w on $p \in (s_j^{BE}, \sigma)$. For this to be an equilibrium, there must exist positive ϵ such that HFM j does not charge prices in $[s_j^{BE}, p + \epsilon]$. If not, HFM j can exploit the profit discontinuity at p, and she undercuts HFM i to obtain positive profits. Also, if HFM j charges prices below p, it is not optimal for HFM i to put a positive weight on p. Thus, HFM j must charge prices above $p + \varepsilon$. In this case, however, it is optimal for HFM i to raise p.

Finally, suppose that $p = s_j^{BE}$ has a positive weight. For $s_j \geq p = s_j^{BE}$, the profits for HFM j satisfy

$$\begin{aligned}
V_j^{HFM}(s_j) &= \Pr(s_j < s_i)\left[(1-\eta)s_j + \eta\gamma_i(s_j - \sigma)\right] \\
&\quad + \Pr(s_j = s_i)\left[(1-a_i)(1-\eta)s_j + \eta\gamma_i(s_j - \sigma)\right] + \Pr(s_j > s_i)\eta\gamma_i(s_j - \sigma) \\
&= (1 - \eta + \eta\gamma_i)(s_j - s_j^{BE}) - [\Pr(s_j > s_i) + \Pr(s_j = s_i)a_i](1-\eta)s_j \\
&< (1 - \eta + \eta\gamma_i)(s_j - s_j^{BE}) - \Pr(s_j > s_i)(1-\eta)s_j
\end{aligned}$$

where the second line comes from subtracting $V_j^{BE}(s_j^{BE}) = 0$. Then, define

$$\epsilon_j \equiv \frac{\Pr(s_j > s_i)s_{i,k}^{BE}}{(1-\eta)\Pr(s_j \leq s_i) + \eta\gamma_i} > 0$$

so that posting $s_j \in [s_j^{BE}, s_j^{BE} + \epsilon_j]$ makes $V_j^{HFM}(s_j) < 0$. Thus, HFM j does not post prices in $[s_j^{BE}, s_j^{BE} + \epsilon_j]$. However, this implies that HFM i has an incentive to raise p from $p = s_j^{BE}$ to $p = s_j^{BE} + \epsilon$, leading to the same discussion as the case with $p \in (s_j^{BE}, 1)$. From the above argument, if HFM i randomizes quote over $[s_j^{BE}, \sigma]$ with an atom p, then p converges to σ. However, this contradict to the first case that shows $p = \sigma$ cannot be an atom.