Speed Choice by High-Frequency Traders with Speed Bumps*

Jun Aoyagi†

November 28, 2019

Abstract

This paper studies the backfire effect of a speed bump, which is a delay imposed on trade execution to slow down HFTs. In my model, the equilibrium bid-ask spread works as an endogenous cost of being faster for liquidity takers because their speed imposes adverse selection and widens the spread. Moreover, the *sensitivity* of the bid-ask spread to the speed negatively depends on the length of a speed bump. This is because the takers' speed matters *less* to liquidity providers' expected return when a speed bump exogenously reduces takers' speed. Hence, in contrast to the intended purpose of the regulation, a longer speed bump may promote HFTs' investment in high-speed technology by reducing the marginal cost of being faster and worsens the adverse selection problem. My model investigates when this backfire effect dominates the direct impact of the regulation, thus explaining ambiguous empirical results on the impact of a speed bump.
Keywords: high-frequency trading, speed bumps, adverse selection, strategic speed decision.
JEL Classification: D40, D47, G10, G18, G20

1 Introduction

The ever-increasing speed of electronic financial markets pushes traders to be lightning fast. They are obsessed with being faster than others to acquire information for trading purposes, spending significant amounts of money on high-speed technologies. With the sophisticated tools, high-frequency traders (HFTs) can extract information from massive layers of signals at the speed of light.

Regulators are concerned about how quickly HFTs can access and act on information because HFTs' speed advantage exposes other traders to the cost of adverse selection in the sense of Glosten and Milgrom (1985). That is, HFTs "snipe" stale quotes provided by market makers if they receive news that is not yet publicly available and find existing quotes outdated and mispriced (Budish et al., 2015).

*First draft: June 2018. This paper was previously circulated under the title "Speed Choice by High-Frequency Traders with Speed Bumps."

†Department of Economics, University of California at Berkeley. E-mail jun.aoyagi@berkeley.edu.

I appreciate the constructive comments from Kosuke Aoki, Michael Brolley, Nicolae Garleanu, Terry Hendershott, Ryo Horii, Ulrike Malmendier, Hayden Melton, Sophie Moinas (discussant), Emi Nakamura, Christine Parlour, Yoko Shibuya, David Sraer, Yingge Yan, Haoxiang Zhu, and Marius Zoican, and from seminar/conference participants at the 9th Stern Annual Microstructure Meeting, ISER at Osaka University, UC Berkeley and University of Tokyo. I am grateful to ClaireValgardson for copy-editing. This paper was awarded the 2018 Moriguchi Prize by the Institute of Social and Economic Research, Osaka University.

Table 1: Top 5 Firms by Volume on BrokerTec

Firm	Volume($ millions)	Market Share
Jump Trading	2,291,000	28%
Citadel LLC	1,004,000	12%
Teza Technologies	905,000	11%
KCG	798,000	10%
JP Morgan	649,000	8%

Note: It tabulates shares in May-June, 2015. Data regarding top-10 HFTs is also available and indicates a similar result. Source: Risk.com, October 2015, Issue 10.

The speed race by HFTs has prompted some exchange platforms to slow down HFT-involved transactions by introducing *speed bumps*.[1] A speed bump imposes an intentional delay on the arrival or execution of orders at a market, aiming to protect traders from exposure to the above-mentioned risks. For example, the Investors Exchange (IEX) adopts a 350-microsecond speed bump on incoming orders and outgoing information from the exchange. The Aequitas NEO and TMX Group, both Canadian exchanges, also apply a few milliseconds of random delay to non-cancellation orders (Appendix C provides more examples).

This paper analyzes the effect of speed bumps on the speed acquisition of HFTs and the adverse selection cost for market makers. The key result is that a speed bump can *increase* the equilibrium speed of HFTs and *worsen* adverse selection for market makers, in contrast to its intended purpose.

Specifically, my model takes into account the HFTs' strategic motive in their speed decision (i.e., they internalize the impact of their speed decision on a market). The strategic motive arises because major high-speed financial institutions have significant shares in the trading volume in markets.[2] For example, Table 1 shows the top five high-frequency financial institutions and their shares in the BrokerTec platform, through which more than half of the U.S. Treasury is traded. When an institution decides a speed technology, she becomes aware of the market impact of her choice because a sizable number of transactions (as in Table 1) involve the same speed technology and may affect the equilibrium price.

First, I consider a simple benchmark structure to separate the key mechanism. There exists a single HFT as an informed liquidity taker (sniper), facing a random speed bump. Market makers set the bid-ask spread that reflects the adverse selection cost, as in Glosten and Milgrom (1985). As the literature (e.g., Biais, Foucault and Moinas, 2015) points out, the faster the HFT, the more severe adverse selection the market makers face. Thus, the HFT's speed puts positive pressure on the bid-ask spread and, in turn, reduces the sniping profit of the HFT. Therefore, the bid-ask spread works as an *endogenous* upward-sloping cost of being faster for the HFT.

A marginally longer speed bump brings about two positive impacts on the HFT's incentive to

[1] Budish et al. (2015) propose frequent batch auctions (FBA) as an alternative trading structure to mitigate the problem. However, adopting FBA needs considerable structural changes in the current continuous markets. In contrast, a speed bump is easier to adopt and has become more popular.

[2] There is anecdotal evidence for HFTs being aware of the market impact of their speed choice. For example, clients who purchase a speed device from a trade technology company often try to hide it by asking to peel corporate logos from shipments due to confidentiality clauses. See, for example, https://www.wsj.com/ and Lewis (2014).

be faster in the equilibrium. Firstly, market makers are less likely to be picked off due to a longer delay. As a result, they do not care much about a marginal increase in the HFT's speed, and their pricing behavior (the bid-ask spread) becomes less responsive to the HFT's speed-up. This means that the HFT's sniping profit does not deteriorate even if she becomes faster, thereby generating further room for speed-up. Secondly, the HFT cares less about the adverse price movement, or the decline in the sniping profit, caused by her speed-up; Even if she becomes faster and the spread widens, her *expected* profit is affected slightly because she is less likely to snipe due to a speed bump. These two effects lower the endogenous marginal cost of speed-up for the HFT, thus providing her with a strong incentive to be faster.

The above idea is applied to analyze a speed competition among *multiple* HFTs—an "arms race"—where they serve not only as snipers but also as high-frequency market makers. My model shows that strategic complementarity in an arms race can arise because the speed-up by an HFT as a market maker has the same effect on her pricing behavior as a longer speed bump, i.e., she is less likely to be picked off, and the bid-ask spread grows insensitive to snipers' speed-up.

Due to the strategic complementarity, the introduction of a speed bump can backfire in terms of adverse selection. A longer delay makes each HFT willing to be faster by the mechanism in the benchmark model, thereby triggering a fiercer speed competition and positive externality. Although a speed bump protects market makers and mitigates adverse selection via its direct effect, the equilibrium speed increases substantially and dominates the direct protection, worsening adverse selection risk.

Moreover, the backfire effect of a speed bump becomes stronger if my model introduces price-elastic *discretionary* liquidity traders (LTs). A speed bump exogenously mitigates adverse selection, tightens the bid-ask spread, and facilitates LTs' participation. A large set of LTs, in turn, render market makers less likely to be picked off, which has the same impact on the bid-ask spread as extending a speed bump. Thus, a speed bump incentivizes HFTs to be faster not only via the benchmark channel but also through the LTs' participation behavior.

My theory characterizes when and why the speed competition between HFTs exhibits complementarity or substitution, suggesting that the effectiveness of a speed bump depends on some observable parameters, such as the price of speed technologies and the expected length of a speed bump. As well, my model provides some policy implications. For example, it suggests that the SEC's policy in 2017 that approved the IEX (with a speed bump) as a National Securities Exchange can strengthen HFTs' demand for speed technologies. This may allow other exchange platforms to charge expensive fees for the fast data access (e.g., colocation service), as we have observed in the real world.[3]

1.1 Literature review

This paper contributes to the literature on high-frequency trading and market structure (see Jones, 2013; O'Hara, 2015; Menkveld, 2016 for reviews). Biais et al. (2015) analyze the effect of an arms race and show that a higher speed triggers more severe adverse selection for slow traders. Delaney

[3]As for the SEC's approval of the IEX, see Hu (2018). For the recent proposals regarding the increasing price of direct data feed charged by exchange platforms, see https://www.wsj.com/articles/nyse-nasdaq-take-it-on-the-chin-in-washington-1539941404.

(2018) describes the speed decision as irreversible investments with an optimal stopping time, while Bongaerts and Van Achter (2016) view it from a perspective of high-frequency market making.[4] However, the speed decision in these models is discrete (i.e., being fast or not). Based on Foucault et al. (2003), studies by Liu (2009), Foucault et al. (2013), and Foucault et al. (2016) investigate a continuous choice of the speed level based on the monitoring intensity of traders. However, traders decide on the speed level simultaneously with other players (e.g., market makers), which requires them to focus on the exogenous cost of speed investment. My model differs from theirs, as the speed decision is continuous and bears an endogenous cost due to the strategic motive of HFTs. It empowers my model to analyze more in depth how speed choice is affected by a speed regulation.

As traders get faster, questions arise regarding the speed and frequency of executions by a trading platform. Du and Zhu (2017) show that a low-frequency platform works better to reallocate assets. Pagnotta and Philippon (2018) also consider platforms' decisions regarding execution frequency and fees to attract speed-sensitive traders. Menkveld and Zoican (2017) explore the effect of latency on HFTs' strategy and spread, citing risk aversion as a key to generating the non-monotonic spread against trading speed. In their analyses, the frequency of transactions is determined at a market level and applies to all investors, thus they pay little attention to the speed choice of HFTs.

Moreover, my model shares the same interests as the studies on the impact of slow market structures, such as frequent batch auctions (Budish et al., 2015; Haas and Zoican, 2016) and speed bumps (Baldauf and Mollner, 2017; Brolley and Cimon, 2017; Aldrich and Friedman, 2018), on HFTs' behavior and adverse selection for market makers.[5] However, they do not consider a continuous optimal speed decision by HFTs with a delay-sensitive endogenous cost. Thus, they conclude that these mechanisms mitigate adverse selection for market makers, an assertion that will be overturned in this paper.

The scope of the literature extends to other empirical findings regarding HFTs and the effect of speed bumps.[6] Hu (2018) analyzes the SEC approval of the IEX as a national securities exchange and finds a net improvement in market quality measured by the spreads. Shkilko and Sokolov (2016) exploit interruptions of messaging caused by precipitation to find a reduction in quoted spreads.[7] Chen et al. (2017) investigate a speed bump at the TMX Alpha, reporting an increase in quoted spreads. Consistent with my model, a recent experimental study by Khapko and Zoican (2019) finds that a marginally longer speed bump promotes the traders' investment in the speed technology when the execution price is endogenously determined by liquidity providers.[8] Overall, empirical and experimental results of the speed bumps look somewhat ambiguous. My model can reconcile these results because, depending on the relative significance of the exogenous cost and the level

[4] Aït-Sahalia and Saglam (2013), Hoffmann (2014), Foucault et al. (2016) construct models with HFTs to address the effect of high-frequency market making. See Conrad et al. (2015) for the empirical study of high-frequency quoting.

[5] Kyle and Lee (2017) argue that submitting trading rate rather than the standard limit orders can reduce the HFTs' speed advantage, as a market becomes fully continuous in terms of price, quantity, and time.

[6] Other empirical studies on HFTs and the transaction speed include, for example, Hendershott and Moulton (2011), Hasbrouck and Saar (2013), Riordan and Storkenmaier (2012),Ye et al. (2013), Frino et al. (2014), Boehmer et al. (2015), and Brogaard et al. (2015).

[7] Although the interruption by precipitation may have a similar effect to a speed bump, this phenomenon is not paid much attention by financial institutions. In contrast, traders anticipate a speed bump and take it into their decision making.

[8] If the price and trading profit are exogenously fixed, Khapko and Zoican (2019) find that a speed bump diminishes equilibrium investment into speed technologies.

of expected delay, a speed bump affects the optimal speed and the equilibrium spread negatively, positively, or not at all.

2 The benchmark model

This section proposes a simple benchmark model to separate the main mechanism. Consider a one-shot exchange of an asset. The asset's value jumps at $t = 0$ and becomes $v = \pm\sigma$ with equal probability, while it has zero-value before the jump. v is publicly announced at stochastic time $T \sim \exp(\gamma)$. Trade occurs during $t < T$ due to liquidity needs or the arrival of a privately informed trader. The structure of the market, as well as the occurrence of the jump, is common knowledge, while traders do not know if it has gone up ($v = \sigma$) or down ($v = -\sigma$) until the information arrival. Following the convention of market microstructure, I assume that each trader can take only a unit position.

2.1 Traders

Market makers There is a continuum of competitive slow, uninformed *market makers* with a unit mass. At $t = 0$, all market makers submit a single-unit limit order with (half) spread s to commit to trade at this price. The order will disappear from the limit order book (LOB) if there is a taker or if the market maker cancels it based on public news/liquidation events. To focus on short-horizon behavior, market makers do not return to the market once they exit.

High-frequency trader There is one ($N = 1$) risk-neutral *high-frequency trader (HFT)*. Before $t = 0$, the HFT invests in a technology that provides speed ϕ, e.g., colocation service for a reduced latency. Given ϕ, she can privately observe v and immediately submits marketable limit orders (i.e., market orders) to "snipe" stale limit orders provided by market makers.[9] This event happens with an exponential probability with intensity ϕ, thus the arrival time of the HFT (with no speed bums) is defined as $T_H \sim \exp(\phi)$.

Liquidity traders In addition, there is a continuum of *liquidity traders* (LTs) who are exposed to a liquidity shock. The shock exogenously makes them submit a buy or sell market order with equal probability, so they convey no information to market makers. Let $T_L \sim \exp(\beta)$ be the stochastic timing of the liquidity shock with intensity $\beta(\geq \gamma)$.

As in Haas and Zoican (2016) and Brolley and Cimon (2017), assume that trading information, including traders' identity, becomes public immediately after an order is executed, i.e., the market is perfectly transparent.

[9] HFT does not intentionally delay the timing of the order submission: if she gets information at t, she immediately sends the order at t. Putting a time lag between obtaining the information and submitting the order can reduce a spread and increase sniping profit. However, without a commitment device, this cannot be an equilibrium since it is always optimal for the HFT at the information arrival time T_H to snipe immediately given the lag she announces at $t = 0$, i.e., there is a time inconsistency.

2.2 Market structure: speed bump

A market imposes a random *asymmetric speed bump* on incoming liquidity-taking orders. Specifically, an order submitted to the market at date t arrives at $t + \delta$, where δ is a random delay. Cancellation requests from market makers are executed promptly (i.e., the speed bump is asymmetric). Thus, during $\tau \in (T_H, T_H + \delta)$, outstanding limit orders can be illusory for the HFT if liquidity traders took it or if market makers canceled due to public news. For notational simplicity, let $\delta \sim \exp(b)$, thus measuring the expected length of a delay by $\lambda \equiv \mathbb{E}[\delta] = b^{-1}$.[10]

The assumption that market makers are not subject to a speed bump captures the structure of asymmetric speed bumps at ANEO, TSX, Chicago Stock Exchange, and NASDAQ OMX PHLX, just to name a few, where they impose a speed bump only on liquidity-taking orders. Table 2 in Appendix C reproduces the list of exchanges in Baldauf and Mollner (2017) and Khapko and Zoican (2019) to show that the asymmetric speed bumps have been popular in actual implementations.

Also, for simplicity, I assume that liquidity traders are not subject to a seed bump. This assumption is to capture the primary purpose of a speed bump, i.e., hampering active strategies by informed HFTs but not active-neutral behavior, such as those by liquidity-driven traders. The assumption has several justifications. First, it considers a speed bump at ANEO, where they categorize traders into latency-sensitive HFTs and other traders (e.g., liquidity traders) and impose delays only on the first type of traders. Second, it also captures the design of a speed bump at IEX: since outgoing information is also delayed for δ, information-driven HFTs incur a delay of at least 2δ after the jump—first δ to obtain information and another δ to execute orders—while liquidity-driven traders incurs only δ delay, as they are not motivated by new information. Thus, the gap between them is δ, and I can normalize it by shifting the model's timeframe by δ-period. Finally, Appendix A.1 shows that my results do not change even if this assumption is relaxed and liquidity traders incur a speed bump.

Extensions: alternative market structures

As an extension, I analyze a case with multiple HFTs in Section 3, in which each HFT serves not only as a sniper but also as a high-frequency market maker. Also, Section 5 considers the *discretionary liquidity traders* (DLTs) and shows that the price-elastic liquidity traders strengthen my results. Appendix A.2 considers a case where market makers can continuously update (cancel and resubmit) their limit orders, and Appendix A.3 analyzes the coexistence of slow and fast markets.

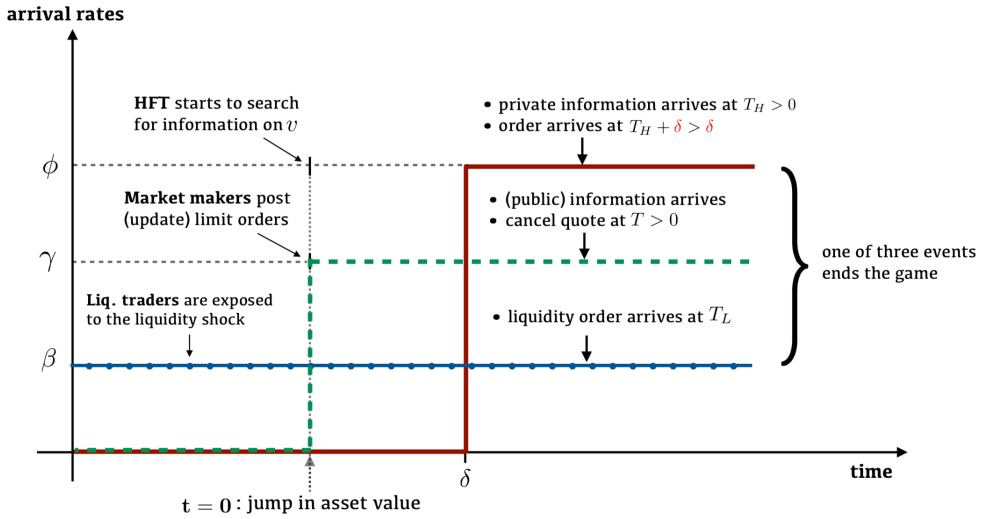
2.3 Equilibrium

The model is conceptualized as the sequential game with two stages. In the first stage, the HFT decides the level of speed ϕ.[11] In the next stage, each market maker submits a competitive limit order, anticipating a confrontation with the informed HFT and liquidity traders. In the trading stage, the HFT looks for an opportunity to snipe. Figure 1 shows the flow of the events in the second stage (i.e., the trading game).

[10] The randomness of δ does not significantly affect my results, while it makes the solution simpler. The case with a deterministic δ is available on request.

[11] The *ex-ante* choice of speed is in line with Foucault et al. (2013) and Brolley and Zoican (2019).

Figure 1: Timeline and Arrival Rates



Note: This figure illustrates the arrival rates of orders. The values of ϕ, β, and γ shown in the figure are set for illustrative reasons and do not exhibit actual equilibrium values.

The equilibrium concept is the subgame perfect equilibrium, and the HFT chooses the optimal level of speed ϕ in light of the optimal reaction of market makers. That is, the HFT knows the price impact of her *speed choice*, as the monopolist in Kyle (1985) knows the price impact of her trading behavior. This behavior is supported by the fact provided in the introduction and Table 1.

2.4 Optimal behavior of market makers

Consider a continuum of market makers quoting a competitive bid-ask price. As in Glosten and Milgrom (1985), serving as a market maker yields zero expected profit. Without loss of generality, I consider how ask price s is determined when $v = \sigma$.[12]

The key effect of a speed bump is to make a market maker less likely to be picked off by the HFT or, put differently, to increase the probability that she observes public news to cancel her limit order. The value function of a market maker with a (half) spread s is given by the following:

$$V = \mathbb{E}_\delta \left[\int_0^\delta s\beta e^{-(\beta+\gamma)t} dt + \int_\delta^\infty (\beta s + \phi(s-\sigma)) e^{-\psi(t-\delta)} e^{-(\beta+\gamma)\delta} dt \right], \tag{1}$$

where $\mathbb{E}_\delta$ is the expectation on δ, and $\psi \equiv \phi + \beta + \gamma$.

Firstly, suppose that trade takes place at $t < \delta$. Then, there is no fear of facing an information-driven HFT due to the speed bump because the fastest possible arrival of the HFT occurs δ period after the jump in the asset value at $t = 0$. On the other hand, liquidity traders can arrive during this "safe interval," because their behavior is independent of the asset value, as captured by the assumption that they are not subject to the delay. Hence, the liquidity traders arrive at t (before the public news) with density $\beta e^{-\beta t} e^{-\gamma t}$. In this case, a market maker obtains $s - \mathbb{E}[v] = s$. The first

[12] Results for the opposite case are given by the symmetric argument.

term in (1) captures this event. Appendix A.1 shows that the following results do not change even if a speed bump delays the liquidity traders, in which case the density of their arrival is proportional to $\beta e^{-\beta(t-\delta)}$.

If trade occurs at $t \geq \delta$, on the other hand, it bears an adverse selection cost: the HFT buys an asset only if a limit order is mispriced given the true information. The HFT can snipe the stale quote at t if she becomes informed at $t-\delta$, and there are no liquidity shocks or public news events during $(t-\delta, t)$. In this case, a market maker obtains $s - \sigma \leq 0$. A market maker can also trade with liquidity traders at t if there is a liquidity shock at t, and the HFT becomes informed *after* $t-\delta$. In this case, the trading profit is $s - \mathbb{E}[v] = s \geq 0$. The expected profit in this second case is represented by the second term in (1), where I use the following probabilities to formulate it: given δ,

$$\begin{aligned}
\Pr(\text{HFT arrives at } t) &= \phi e^{-\phi(t-\delta)} e^{-(\beta+\gamma)t}, \\
\Pr(\text{Liq. traders arrive at } t) &= \beta e^{-\phi(t-\delta)} e^{-(\beta+\gamma)t}, \\
\Pr(\text{cancellation at } t) &= \gamma e^{-\phi(t-\delta)} e^{-(\beta+\gamma)t}.
\end{aligned}$$

It is then possible to get the equilibrium spread from the break-even condition:

Proposition 1. *The equilibrium (half) spread is given by*

$$s = \frac{\phi \mathbb{E}_\delta[e^{-(\beta+\gamma)\delta}]}{(\phi+\beta)\mathbb{E}_\delta[e^{-(\beta+\gamma)\delta}] + \frac{\beta\psi}{\beta+\gamma}(1 - \mathbb{E}_\delta[e^{-(\beta+\gamma)\delta}])} = \frac{\frac{\phi}{1+\lambda\psi}}{\frac{\phi}{1+\lambda\psi} + \beta}\sigma. \tag{2}$$

In (2), a direct effect of the speed bump appears in the form of the discount on the arrival rate of the HFT, which is given by $(1+\lambda\psi)^{-1}$. This term stems from the safe interval and mitigates adverse selection risk. Thus, if ϕ is fixed, a higher λ induces a smaller spread. This direct effect is consistent with the existing works, such as Budish et al. (2015) and Baldauf and Mollner (2017).

However, this argument is ignorant of an endogenous reaction of ϕ. Specifically, a speed bump affects the expected profit of the HFT and her optimal speed, ϕ^*. The existing models argue that the incentive to be faster diminishes as a speed bump becomes longer, i.e., ϕ^* is a *decreasing* function of λ, whereas I propose the opposite effect in this paper.

Sensitivity of the bid-ask spread

Note that the price impact of the speed is positive $\frac{ds}{d\phi} > 0$. This is because a faster HFT worsens the adverse selection problem, and market makers charge a wider spread to break even. Moreover, the reaction of the spread to the speed exhibits the following property:

Lemma 1. *The sensitivity of the price to the speed is decreasing in λ, i.e.,*

$$\frac{\partial}{\partial\lambda}\left(\frac{ds}{d\phi}\right) < 0.$$

Therefore, the longer the expected delay, the less sensitive the spread becomes. A market with high λ is protected by a long (expected) safe interval, and market makers behave as if the share (i.e., the arrival rate) of the HFT is small. Hence, market makers care *less* about the speed investment by

the HFT compared to a market with small λ. Thus, market makers' pricing behavior becomes less sensitive to ϕ if λ is large.

2.5 Expected profit of the HFT

If the HFT becomes informed, it is optimal for her to immediately submit a unit order since she exits the market once her orders are executed, i.e., she is a short-term investor.[13]

The HFT can snipe at $t+\delta$ with the following probability:

$$\pi_t(\phi,\delta) \equiv \Pr(T_H = t, \min\{T, T_L\} > t+\delta) = \underbrace{\phi e^{-\psi t}}_{\text{prob of (i)}} \times \underbrace{e^{-(\beta+\gamma)\delta}}_{\text{prob of (ii)}}. \tag{3}$$

Equation (3) is the probability that (i) the HFT observes news at t before other two events occur, $\phi e^{-\psi t}$, multiplied by the probability that (ii) there are no liquidity shocks or cancellation during $(t, t+\delta)$, which is $e^{-(\beta+\gamma)\delta}$. The second part is generated by a speed bump, i.e., risk of front-running by liquidity traders or cancellation by market makers.

Then, the objective function of the HFT in the first stage takes a simple form:

$$W(\phi) \equiv \mathbb{E}_\delta\left[\int_0^\infty \pi_t(\phi,\delta)(\sigma - s)dt\right] = \pi(\phi,\lambda)(\sigma - s), \tag{4}$$

where $\pi(\phi,\lambda) \equiv \mathbb{E}_\delta\left[\int_0^\infty \pi_t(\phi,\delta)dt\right]$ denotes the expected probability that the HFT can snipe. Moreover, due the independence of $\{T, T_H, T_L\}$ and δ, $\pi(\phi,\lambda)$ is decomposed into two factors:

$$\pi(\phi,\lambda) = \frac{\tilde{\pi}(\phi)}{h(\lambda)} \tag{5}$$

with $\tilde{\pi}(\phi) = \frac{\phi}{\psi}$ and $h(\lambda) = 1 + \lambda(\beta+\gamma)$. $\tilde{\pi}(\phi)$ is the *fundamental sniping probability* when the HFT has speed ϕ, i.e., the probability of sniping when there are no speed bumps ($\lambda = 0$). $h(\lambda)$ is the *discount* on $\tilde{\pi}$ caused by a speed bump.

For later use, we rewrite the equilibrium spread in (2) by using $\pi(\phi,\lambda)$ as follows:

$$s = \frac{y(\pi(\phi,\lambda))}{\beta + y(\pi(\phi,\lambda))}\sigma \tag{6}$$

with

$$y(\pi) \equiv \frac{\pi}{1-\pi}(\beta+\gamma).$$

Note that y represents how likely a market maker is picked off by the HFT. From (6), a speed bump, λ, affects the equilibrium spread only through this probability of being picked off, $y(\pi(\phi,\lambda))$.

[13] See Appendix A.2 for a more general setting with continuous updating by market makers and time-dependent s_t.

2.6 The optimal speed

Consider the speed choice by the HFT in the first stage. To obtain an interior solution, I make the expected length of delay relatively short:

$$\lambda < \frac{1}{\sqrt{\beta(\beta+\gamma)}}. \tag{7}$$

The intuition behind this condition will be provided after offering the main propositions.

The optimization problem of the HFT is

$$\begin{aligned} \max_{\phi} W(\phi) &\equiv \pi(\phi,\lambda)(\sigma - s(\pi(\phi,\lambda))) - C(\phi), \\ s.t.\ s &= \frac{y(\pi(\phi,\lambda))}{\beta + y(\pi(\phi,\lambda))}\sigma. \end{aligned} \tag{8}$$

$C(\phi)$ denotes the exogenous cost of speed. In the following, I set $C(\phi) = 0$ to separate the key mechanism, while Subsection 4 analyzes non-zero costs.

The problem in (8) indicates that the HFT decides ϕ knowing the price impact of her speed decision, i.e., she is strategic. In this case, being faster pushes up the price charged by a market maker (s) and reduces her sniping profit. For this reason, the equilibrium spread can be seen as an *endogenous* cost of being faster.

To understand how $\phi^* \equiv \arg\max_\phi W$ is determined, consider the marginal impact of being faster:

$$\begin{aligned} W'(\phi) &= (\sigma - s)\frac{d\pi}{d\phi} + \pi\frac{d(\sigma - s)}{d\phi}. \\ &= \frac{d\pi}{d\phi}\left[(\sigma - s) + \pi\frac{d(\sigma - s)}{d\pi}\right] \end{aligned} \tag{9}$$

When the HFT becomes marginally faster, she faces a price-liquidity tradeoff. On the one hand, W increases, as the HFT is more likely to snipe (the first term in [9]). On the other hand, it reduces her profit as the trading cost s goes up (the second term in [9]). These factors constitute the marginal gain and cost of being faster.

Also, (6) induces the second line in (9): ϕ affects the sniping profit, $\sigma - s$, only through a change in π. Namely, a faster HFT increases her arrival rate ($\frac{d\pi}{d\phi} > 0$), making market makers charge a wider spread ($\frac{ds}{d\pi} > 0$).

Furthermore, W' is expressed by using the *elasticity* of the sniping profit, denote by $\varepsilon(\phi)$:[14]

$$W' = (\sigma - s)\frac{d\pi}{d\phi}\left[1 - \varepsilon(\phi,\lambda)\right] \tag{10}$$

with

$$\varepsilon \equiv -\frac{d\log(\sigma - s(\phi,\lambda))/d\phi}{d\log\pi(\phi,\lambda)/d\phi} > 0.$$

From (10), when the equilibrium spread is more sensitive to ϕ than the sniping probability (i.e., $\varepsilon > 1$), being faster harms the HFT's profit, and her incentive to increase ϕ dwindles. On the other hand,

[14] Appendix B.1 provides an explicit formula for ε.

if the price impact of speed is sufficiently small, the positive effect of a higher sniping probability dominates a decline in the sniping profit ($\varepsilon < 1$), luring the HFT to be faster.

Appendix B.1 shows that W satisfies the SOC so that the optimal speed ϕ^* is derived by solving for the FOC. Therefore, the HFT selects $\phi = \phi^*$ that makes the sensitivity of the spread (or the trading profit) and that of the sniping probability identical, i.e., $\varepsilon(\phi^*) = 1$. We obtain the following explicit formula:

Proposition 2. *(i) The optimal speed is given by*

$$\phi^* = \frac{\sqrt{\beta+\gamma}(1+\lambda(\beta+\gamma))}{1-\lambda\sqrt{\beta(\beta+\gamma)}}. \tag{11}$$

(ii) ϕ^* is increasing in λ.

Proof. See Appendix B.1. □

In contrast to the existing models, Proposition 2 demonstrates that a speed bump increases the equilibrium speed of the HFT. Once again, equilibrium spread s serves as an *endogenous* cost of being faster for the HFT: this not only guarantees a bounded solution even without an exogenous cost of speed (i.e., $C(\phi) = 0$) but also overturns the traditional result regarding speed bumps (point [ii] in Proposition 2).

Intuition

To understand why a speed bump promotes the HFT's speed, consider the partial derivative of $\frac{dW}{d\phi}$ in (9) with respect to λ:

$$\begin{aligned}\frac{\partial}{\partial\lambda}\left(\frac{dW}{d\phi}\right) = &\overbrace{\frac{\partial(\sigma-s)}{\partial\lambda}\frac{d\pi(\phi,\lambda)}{d\phi}}^{\text{effect (i)}>0} + \overbrace{(\sigma-s)\frac{\partial}{\partial\lambda}\frac{d\pi(\phi,\lambda)}{d\phi}}^{\text{effect (ii)}<0} \\ &+ \underbrace{\frac{\partial\pi(\phi,\lambda)}{\partial\lambda}\frac{d(\sigma-s)}{d\phi}}_{\text{effect (iii)}>0} + \underbrace{\pi(\phi,\lambda)\frac{\partial}{\partial\lambda}\frac{d(\sigma-s)}{d\phi}}_{\text{effect (iv)}>0}.\end{aligned} \tag{12}$$

Increasing λ has three positive effects and one negative effect. Effects (i) and (ii) in the first line come from the non-strategic behavior, whereas effects (iii) and (iv) in the second line are from the strategic behavior.

1. Firstly, a speed bump increases the sniping profit by $\frac{\partial(\sigma-s)}{\partial\lambda} = \tilde{\pi}\frac{\partial h^{-1}}{\partial\lambda}\frac{d(\sigma-s)}{d\pi}$. A longer delay enlarges the discount h on the fundamental sniping probability $\tilde{\pi}$ which, in turn, allows a market maker to narrow the spread down due to less severe adverse selection risk. As a result, increasing the sniping probability earns a larger profit by $\frac{\partial(\sigma-s)}{\partial\lambda}\frac{d\pi}{d\phi} = \tilde{\pi}\frac{\partial h^{-1}}{\partial\lambda}\frac{d(\sigma-s)}{d\pi}\frac{1}{h}\frac{d\tilde{\pi}}{d\phi}$. This positive impact is shown by effect (i) in (12).

2. Secondly, even if the HFT increases the fundamental sniping probability, $\frac{d\tilde{\pi}}{d\phi}$, a large discount, h, makes it less worthwhile. This effect reduces the marginal benefit of being faster by $(\sigma - s)\frac{\partial}{\partial\lambda}\frac{d\pi(\phi,\lambda)}{d\phi} = (\sigma-s)\frac{\partial h^{-1}}{\partial\lambda}\frac{d\tilde{\pi}}{d\phi}$ and constitutes the negative second term in (12), i.e., effect (ii).

3. Thirdly, the HFT knows that she is less likely to snipe a stale quote if there is a speed bump, i.e., $\frac{\partial \pi}{\partial \lambda} < 0$. Then, she becomes careless of the adverse price movement caused by her speed up, $\frac{d(\sigma - s)}{d\phi} < 0$, i.e., even if she becomes faster and the spread widens, this is less likely to affect her expected profit. This reduces the marginal cost of being faster and has a positive impact on W', as shown by effect (iii) in (12).

4. Lastly, a speed bump makes the spread less responsive to the speed-up by the HFT, as Lemma 1 shows. This reduces the marginal cost of being faster, as effect (iv) in (12) shows.

The first two effects (i.e., non-strategic channels) have been analyzed in the literature, such as Biais et al. (2015) and Foucault et al. (2016), in the context of a speed choice. Importantly, they cannot be positive at the equilibrium (Foucault et al., 2016). Intuitively, effect (ii) represents how λ *directly* deteriorates the marginal profit W' by reducing the sniping probability, while effect (i) captures the positive *indirect* effect of λ, i.e., a reduction in the sniping probability improves W' by changing the spread s. At the equilibrium, the positive indirect effect cannot dominate the direct effect. Thus, by letting $\Delta_{h\tilde{\pi}} \equiv \frac{\partial h^{-1}}{\partial \lambda} \frac{d\tilde{\pi}}{d\phi}$, the first line in (12) is rewritten as follows:

$$\text{effects (i) + (ii)} = \Delta_{h\tilde{\pi}} \left[(\sigma - s) + \pi \frac{d(\sigma - s)}{d\pi} \right] \leq 0. \tag{13}$$

From (9) and the envelope condition, effects (i) and (ii) cancel out each other, and (13) holds with equality at the equilibrium if there are no exogenous costs of increasing ϕ. If I add an exogenous cost of ϕ, as in the existing models, the FOC implies $(\sigma - s) + \pi \frac{d(\sigma - s)}{d\pi} < 0$, thus making (13) hold with strict inequality (see Subsection 4).

As a result, in the benchmark model, two positive effects in the second line of (12) become dominant. A longer expected delay (λ) makes the HFT more willing to invest into the speed technology (ϕ) because it makes the HFT care less about the adverse price movement caused by her speed up (effect [iii]), as well as it makes market makers (and the equilibrium spread) less responsive to the speed-up of the HFT (effect [iv]).

The upshot is that a speed bump does not prevent a speed race but promotes it, as Proposition 2 attests. This surprising finding highlights the main difference of my results from the literature on speed regulations, such as Budish et al. (2015), Haas and Zoican (2016), and Baldauf and Mollner (2017). In their models, the HFT does not care about the effect of her speed choice on the spread or does not optimize ϕ with a continuous domain (i.e., a discrete choice between being fast or not). In these existing cases, the strategic effects ([iii] and [iv]) in (12) disappear, making the optimal speed ϕ^* decrease in the length of a speed bump.

2.7 Adverse selection

A speed bump has two competing effects on adverse selection risk. First, as the literature suggests, a speed bump directly mitigates adverse selection because it protects market makers from the HFT. However, my strategic model adds an opposing channel: a speed bump promotes speed investment by the HFT (Proposition 2). In the following, the equilibrium half spread s is used as a measure of adverse selection.

Proposition 3. *The equilibrium spread is independent of the expected delay, i.e.,* $ds/d\lambda = 0$.

Proof. See Appendix B.2. □

This result shows that a speed bump cannot mitigate (or worsen) adverse selection for market makers in the benchmark model.

With ϕ fixed, a speed bump reduces the profit of the HFT, as the discount h on the sniping probability increases. To compensate for this disadvantage, the strategic HFT tries to be faster. Since she is a monopolistic HFT and anticipates the price impact of her speed investment, she can choose the level of ϕ that just offsets the cost from a speed bump. In other words, knowing that a speed bump slows the HFT down, the optimal reaction of the HFT is to move faster and offset the impact of a speed bump. As a result, the two competing consequences of a speed bump cancel each other out.

In the following sections, however, I show that this independence ($\frac{ds}{d\lambda} = 0$) is specific to the benchmark model and holds only in this special case: more general market structures change this result.

3 Multiple HFTs and high-frequency market making

In the real world, HFTs serve not only as takers (snipers) but also as liquidity providers. I modify the benchmark model to capture this fact.

To keep deviation from the benchmark model as minimal as possible, I adopt the following structure. There are three HFTs ($i = 1, 2, 3$),[15] and all of them determine the speed level ϕ_i before the game starts. Given the speed, nature at $t = 0$ selects two HFTs as high-frequency market makers (HFMs), while the remaining one serves as a high-frequency sniper (HFS), where the selection is random (each HFT becomes an HFS with probability $1/3$).

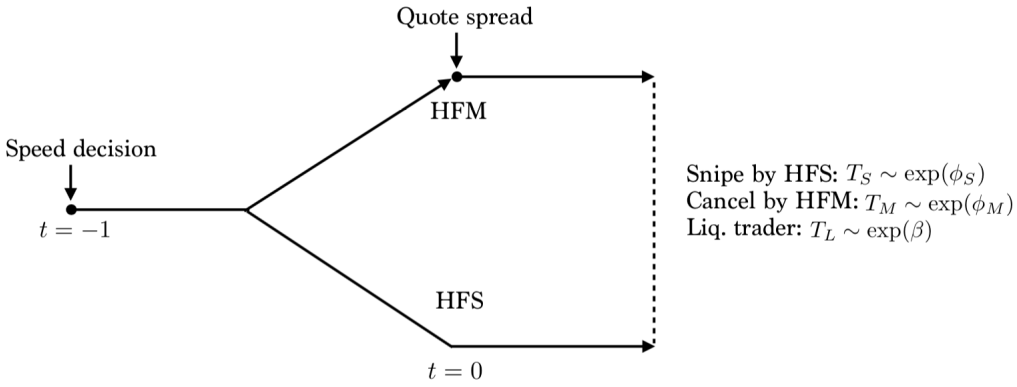
Once the role of HFTs is determined, HFMs at $t = 0$ simultaneously provide liquidity, i.e., submit a single-unit limit order by specifying the (half) spread. Only the best price prevails and is posted on the LOB: an HFM with the best price is entitled to trade subsequently, while one with an inferior price exits.[16] The existence of (at least) two HFMs renders the quoting behavior at $t = 0$ competitive: the Bertrand competition drives HFMs' expected profit at the quoting stage zero so that I can reuse the results in the benchmark model.

Given the limit order(s), the trading game ends with one of three possible trigger events. Firstly, each HFT i obtains the private news with a Poisson rate ϕ_i, and the true v is realized. If this event happens to the HFS, she submits a marketable limit order to snipe the stale quote. On the other hand, if the private news arrives at the HFM, she cancels her stale quote to avoid being picked off. Finally, a liquidity trader arrives with a Poisson rate β and takes the same behavior as in the benchmark model. I ignore the public news at $T \sim \exp(\gamma)$ because cancellation is made by an HFM in this extension. Other structures of the game, including a speed bump, stay the same. For technical reasons, assume that $\beta \geq 1$ and focus on the symmetric equilibrium. Figure 2 summarizes the timeline of the game.

[15] The case with $N \geq 4$ HFTs is easily analyzed with a minor modification.

[16] This is because of the unit trade assumption. If both HFMs provide the same price, which happens at the symmetric equilibrium, one of them is randomly selected and posted on the LOB. I assume that the unselected HFM stays inactive, while allowing her to serve as an HFS does not change my results.

Figure 2: Timeline of the game



Note: This figure illustrates the timeline of the game with multiple HFTs. After the speed decision at $t = -1$, two HFTs are selected as market makers (HFMs) and submit competitive quotes at $t = 0$. The HFM who provides the best price is entitled to trade, and the other HFM exits from the market. The remaining HFT serves as a sniper. The game ends if the HFS snipes (ϕ_S), the HFM cancels (ϕ_M), or a liquidity trader arrives (β).

3.1 Optimal behavior of HFTs

The problem is solved by backward induction. I specify the HFM and HFS by indices $i = M$ and S, respectively.

The high-frequency market maker

Firstly, consider an HFM deciding on the ask price, s. Her behavior is the same as that of ordinary competitive market makers in the benchmark model: her expected return from market making is given by V in (1), and the Bertrand competition with the other HFM drives her profit zero. The only difference from the benchmark model is that she can observe private news and cancel her limit orders at $T_M \sim \exp(\phi_M)$ instead of the public news arrival at T. Thus, ϕ and γ are replaced by ϕ_S and ϕ_M, respectively.

The break-even condition yields the following equilibrium spread:

$$s(\phi_S, \phi_M) = \frac{\frac{\phi_S}{1+\lambda(\phi_S+\phi_M+\beta)}}{\frac{\phi_S}{1+\lambda(\phi_S+\phi_M+\beta)} + \beta}\sigma. \tag{14}$$

This spread has the same structure as the benchmark s in (2): it reflects the expected value of v conditional on trade at the ask price, s. Note that the symmetric equilibrium with $\phi = \phi_i = \phi_{-i}$ makes two HFMs submit the same spread.

Speed decision

HFT i decides ϕ_i before her role in the game is determined. Since the Bertrand competition drives her expected profit from market making zero, her *ex-ante* expected gains come only from the sniping

profit. Thus, the speed decision is analogous to (4):[17]

$$\max_{\phi_i} W_i = \pi(\phi_i, \phi_M; \lambda)(\sigma - s(\phi_i, \phi_M)), \tag{15}$$

$$\text{s.t., } \pi(\phi_i,\phi_M;\lambda) = \frac{1}{1+\lambda(\beta+\phi_M)} \frac{\phi_i}{\phi_i+\phi_M+\beta}, \tag{16}$$

$$s(\phi_i, \phi_M) = \frac{\frac{\phi_S}{1+\lambda(\phi_S+\phi_M+\beta)}}{\frac{\phi_S}{1+\lambda(\phi_S+\phi_M+\beta)}+\beta}\sigma.$$

This problem has the same structure as the benchmark case with γ replaced by ϕ_M. Note that HFT i in the speed-choice stage decides ϕ_i anticipating that she becomes an HFS and one of the other HFTs becomes an HFM. Since a faster market maker posts a better price, we can define $\phi_M \equiv \max\{\phi_{-i}\}$ and, for notational convenience, suppose that $\phi_j = \max\{\phi_{-i}\}$.[18]

The best response function of HFT i is given by the following (see Proposition 2):

$$BR_i(\phi_j) = \frac{\sqrt{\beta+\phi_j}[1+\lambda(\beta+\phi_j)]}{1-\lambda\sqrt{\beta(\beta+\phi_j)}}. \tag{17}$$

This value is bounded as long as $1 > \lambda\sqrt{\beta(\beta+\phi_j)}$. Otherwise, $\phi_i = \infty$ is the best response. This section focuses on bounded responses, while Subsection 3.2 provides full characterization of BR.

Behavior of the best response function

The best response function has the following property:

Proposition 4. *(i) The arms race exhibits strategic complementarity, i.e.,* $\frac{dBR_i(\phi_j)}{d\phi_j} > 0$.
(ii) A speed bump, λ, has a positive impact on the best response function, i.e., $\frac{\partial BR_i}{\partial \lambda} > 0$.

Intuition for the impact of λ in point (ii) of Proposition 4 is the same as the benchmark model. The mechanism for the strategic complementarity (point [i]) should be clear by investigating the marginal gain of being faster for HFT i:

$$\frac{\partial W_i}{\partial \phi_i} = (\sigma - s(\phi_i, \phi_j))\frac{\partial \pi_i}{\partial \phi_i} + \pi_i \frac{\partial(\sigma - s(\phi_i, \phi_j))}{\partial \phi_i}, \tag{18}$$

where HFT i's sniping probability is $\pi_i = \pi(\phi_i, \phi_j; \lambda)$ in (16).

The first term is the marginal improvement in the sniping probability, and the second term is a decline in the profit. These terms can be seen as the marginal benefit and cost of being faster, as in the benchmark model, while they differ from the benchmark because (18) depends on the speed of the competitor, ϕ_j.

[17] The ex-ante expected profit should be multiplied by $1/3$ but I ignore this because it does not affect the result.

[18] More precisely, HFT i knows that there are two possibilities: (a) the case that other HFTs, say j and k, submit the different spreads and (b) the case with j and k submit the same spread. In case (a), I can define $\phi_M \equiv \max\{\phi_j, \phi_k\}$ in (15), as the fastest HFM provides the best price, and HFT i anticipates to trade with this HFM. In case (b), the objective function is $W_i = \frac{1}{2}\sum_{l=j,k} \pi(\phi_i, \phi_l; \lambda)(\sigma - s(\phi_i, \phi_l))$ due to the tie-breaking rule in the market-making sector. Obviously, this leads to (15) with $\phi_M = \phi_j = \phi_k$ because case (b) happens only if two HFMs quote the same speed.

To understand the reaction of BR_i to ϕ_j, the cross-derivative with respect to ϕ_j must be analyzed:

$$\frac{\partial^2 W_i}{\partial \phi_j \partial \phi_i} = \overbrace{\frac{\partial \pi_i}{\partial \phi_i}\frac{\partial(\sigma - s_j)}{\partial \phi_j}}^{\text{effect (i)}>0} + \overbrace{(\sigma - s_j)\frac{\partial^2 \pi_i}{\partial \phi_j \partial \phi_i}}^{\text{effect (ii)}<0} + \underbrace{\frac{\partial \pi_i}{\partial \phi_j}\frac{\partial(\sigma - s_j)}{\partial \phi_i}}_{\text{effect (iii)}>0} + \underbrace{\pi_i \frac{\partial^2(\sigma - s_j)}{\partial \phi_j \partial \phi_i}}_{\text{effect (iv)}>0}. \tag{19}$$

It is worth noting that an increase in the competitor's speed ϕ_j and a marginally longer speed bump λ have the same type of impact on the marginal profit of HFT i, $\frac{\partial W_i}{\partial \phi_i}$, because both effects make an HFM less likely to be picked off.

Therefore, a faster opponent charges a smaller spread. As effect (i) shows, the opponent's speed raises the sniping profit for HFT i, making it more worthwhile to have a higher π_i. However, the opponent's speed decreases the marginal improvement in the sniping probability due to the large discount, h, as effect (ii) shows. At the same time, a faster opponent reduces the (endogenous) marginal cost of being faster for HFT i because she is less likely to snipe and does not need to care much about the adverse price movement of the price due to her speed-up, which constitutes effect (iii). Moreover, effect (iv) stems from the fact that a faster opponent becomes more insensitive to HFT i's speed-up due to the same logic as Lemma 1. Following the same argument as in the benchmark, the overall effect of ϕ_j in (19) becomes positive, thereby making BR_i an increasing function of ϕ_j.

Moreover, "tit for tat" due to complementarity can be strong, and BR becomes convex when the opponent is sufficiently fast.

Corollary 1. *There is a unique $\phi_j = \phi_0$ such that*

$$\frac{d^2 BR_i(\phi_j)}{d\phi_j^2} > 0 \Leftrightarrow \phi_j > \phi_0. \tag{20}$$

Proof. See Appendix B.3. □

The convexity of BR helps us deriving the symmetric equilibrium below.

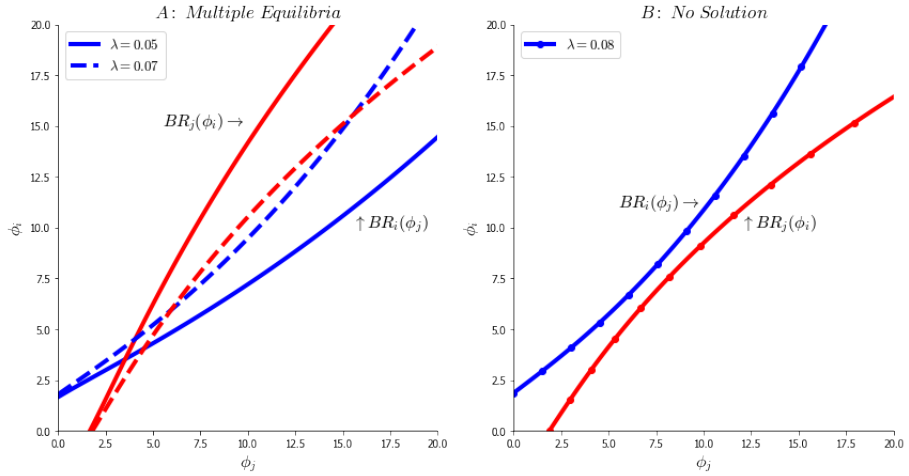
3.2 Equilibrium speed

From (17), we obtain $BR_i(0) > 0$, i.e., facing a zero-speed opponent, HFT i still maintains a positive speed. This is because $\phi_i > 0$ yields a positive profit, while $\phi_i = 0$ keeps it at zero. Together with (20), this implies that multiple symmetric equilibria can arise.

Proposition 5. *(i) There exists a unique λ_0. If $\lambda > \lambda_0$, no bounded solutions exist. If $\lambda \leq \lambda_0$, there are two bounded solutions for $BR(\phi) = \phi$ denoted as $\phi_L^* < \phi_H^*$.*
(ii) ϕ_L^* is stable and ϕ_H^* is unstable.
(iii) At the stable equilibrium, $\phi_L^* \equiv \phi_i = \phi_{-i}$ is increasing in λ.

Proof. See Appendix B.3. □

Figure 3: Best Response Functions



Note that the effects of higher λ and ϕ_j act in the same direction to increase BR_i. Thus, due to the same logic as in Proposition 4 and Lemma 1, a sufficiently high λ makes the complementarity strong enough to eliminate a bounded solution, i.e., $\phi = \infty$ is optimal. On the other hand, when λ is small, we obtain bounded symmetric equilibria.

Following the convention (Hendershott and Mendelson, 2000; Zhu, 2014), I use stability as an equilibrium selection criterion. The unstable equilibrium ϕ_H^* is not robust to a small perturbation in a parameter value, whereas the stable equilibrium ϕ_L^* does not diverge even if a parameter changes slightly. Thus, my focus is on $\phi^* = \phi_L^*$.

Figure 3 plots the best response functions for different values of λ. The discussions so far have established that a speed bump λ increases the marginal benefit of being faster. Thus, as suggested by Proposition 4 and illustrated by Figure 3, an increase in λ shifts BRs upward, leading to a higher speed at the stable equilibrium.

3.3 Impact on market quality

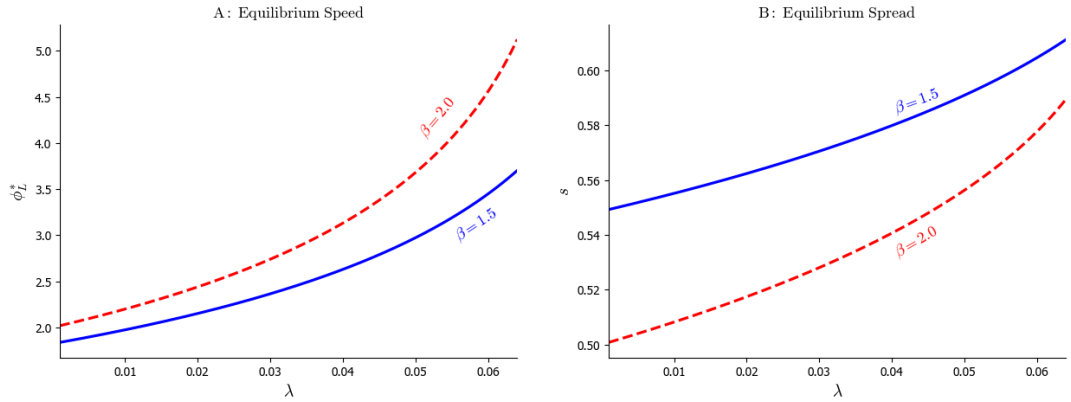
To measure the impact of a speed bump on market quality, this subsection analyzes the behavior of the adverse selection cost and price discovery. As in the literature (e.g., Glosten and Putnins, 2016; Rosu, 2018), faster HFTs lead to the tradeoff between more severe adverse selection and more efficient price discovery.

Adverse selection and the equilibrium spread

The effect of a speed bump on the spread and adverse selection can be derived analytically. Due to "fast market making," adverse selection risk is mitigated, and λ helps protect market makers. However, at the symmetric equilibrium, high-frequency snipers become faster by the same degree as market makers. As the strategic complementarity is sufficiently strong, this arms race outweighs the direct protection by the speed bump, expanding the spread.

Proposition 6. *A longer speed bump widens the spread;* $\frac{ds}{d\lambda} > 0$.

Figure 4: Impact of λ on speed and spread



Note: Panel A illustrates the speed level at the stable symmetric equilibrium, and Panel B plots the equilibrium spread at the same equilibrium. The solid and dashed lines represent different value of the liquidity traders' arrival rate, β

Proof. See Appendix B.4. □

Figure 4 plots this result. The introduction of a speed bump or a longer expected delay can backfire, increasing the equilibrium speed and worsening adverse selection.

In the benchmark model, the speed-up by the single HFT is an indirect consequence of the speed bump and cannot offset the direct protection of market makers, leading to $\frac{ds}{d\lambda} = 0$. By contrast, multiple HFTs generate a positive externality through strategic complementarity (Proposition 4). In this situation, an increase in λ indirectly affects the best response functions of both HFTs, thus shifting them upward, as shown in Figure 3. This triggers an arms race with positive externality and amplifies the indirect effect. As a result, the speed-up in the symmetric equilibrium dominates the direct protection of market makers, leading to more severe adverse selection.

Price discovery

How fast does the equilibrium quote incorporate new information? For information to be reflected by the price, a stale quote must be removed from the market either by an informed market maker canceling her quote or an informed high-frequency sniper taking it. These events are triggered by a Poisson arrival with intensity ϕ_S or ϕ_M but incur a delay due to a speed bump.

Let $\tau \equiv \min\{T_M, T_S + \delta\}$ be the arrival time of this trigger event. By exploiting the property of the exponential distribution, the expected length of time until the price discovery is then given by

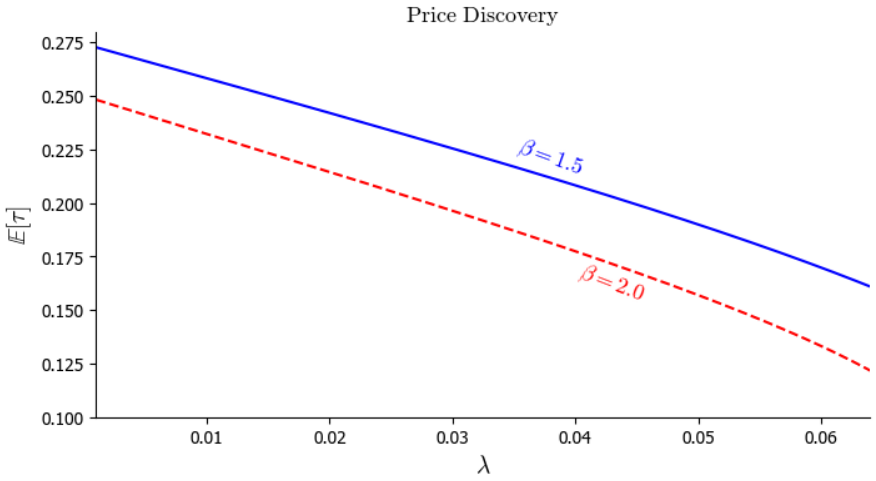
$$\mathbb{E}[\tau] = q\frac{1}{\phi_M + \phi_S} + (1-q)\frac{\lambda}{1+\lambda\phi_M}$$

with $q = \frac{1}{1-\lambda\phi_S}$. At the symmetric equilibrium, the speed of HFM and HFS must be the same. Thus, it reduces to

$$\mathbb{E}[\tau] = \frac{1}{2}\left(\frac{1}{\phi_L^*} + \frac{\lambda}{1+\lambda\phi_L^*}\right), \tag{21}$$

where ϕ_L^* is defined in Proposition 5.

Figure 5: Price discovery



Note: This figure plots the expected time until price discovery $\mathbb{E}[\tau]$ against the expected length of delay λ with different β.

Two competing effects are at play: a longer speed bump (λ) causes a delay in price discovery by prohibiting the HFS from immediately sniping, while it promotes the speed level (ϕ_L^*), thereby making cancellation and sniping more likely. Figure 5 illustrates the impact of a larger λ and finds that the second effect dominates the first one. Once again, this is because the positive externality in an arms race promotes ϕ^* significantly so that it dominates the direct impact of a speed bump.

Also, Figure 5 shows that β negatively affects $\mathbb{E}[\tau]$. Intuitively, a large β makes the market maker (and the spread) less responsive to an increase in ϕ_S due to the same logic as the impact of λ and ϕ_M (see Panel A in Figure 4). Thus, the best response function tends to be steep when λ and β are large, making the strategic complementarity strong. In this situation, ϕ_L^* becomes more reactive to λ and dominates the direct impact of λ in (21). As a result, expanding a speed bump promotes the speed of HFTs and facilitates price discovery.

4 Exogenous cost of speed

To compare my results to the existing models, the following subsection considers a model with non-strategic HFTs and the exogenous cost of being faster, $C(\phi_i) = c\phi_i$, as in Foucault et al. (2016).[19] Later, Subsection 4.2 incorporates the effects in the existing models into my model. To make the comparison clearer, the model I have discussed so far is referred to as the *strategic model*.

4.1 Non-strategic HFTs

If the strategic motive is absent, the optimal speed solves

$$W_i = \max_{\phi_i} \pi_i(\phi_i, \phi_j, \lambda)(\sigma - s) - C(\phi_i),$$

[19]The linear cost can be replaced by a quadratic cost, $\frac{c}{2}\phi_i^2$, without changing the implications of my results.

by taking s given. π_i is the sniping probability of HFT i and is given by (16).

In this case, the marginal profit of being faster and the cross-derivative become

$$\frac{\partial W_i}{\partial \phi_i} = (\sigma - s(\phi_i, \phi_M))\frac{\partial \pi_i}{\partial \phi_i} - c, \tag{22}$$

$$\frac{\partial^2 W_i}{\partial \phi_M \partial \phi_i} = (\sigma - s(\phi_i, \phi_M))\underbrace{\frac{\partial^2 \pi_i}{\partial \phi_M \partial \phi_i}}_{<0} + \underbrace{\frac{\partial \pi_i}{\partial \phi_i}\frac{\partial(\sigma - s(\phi_i, \phi_M))}{\partial \phi_M}}_{>0}. \tag{23}$$

The exogenous marginal cost c replaces the second term of (18) that represents an endogenous marginal cost. Also, effects (iii) and (iv) in (19) that show the strategic motive disappear from (23). Intuition for each component remains the same as I have discussed with (19).

At the symmetric equilibrium, the following results hold.

Proposition 7. *(i) Around the symmetric equilibrium, the best response function exhibits strategic substitution;* $\frac{dBR_i(\phi_j)}{d\phi_j} < 0$.
(ii) The equilibrium speed and spread are decreasing in λ.

Proof. See Appendix B.5. □

A marginally faster opponent in the non-strategic model affects HFT-i's decision by rendering sniping more difficult (the negative first term in [23]) and by improving the sniping profit (the positive second term in [23]). The first one is the direct effect—a reduction in the sniping probability of HFT i—whereas the second one is the indirect effect—the reduction in π_i mitigates adverse selection for the HFM, thereby tightening the spread. As in Foucault et al. (2016), the direct effect cannot dominate the indirect one, and the total effect is negative. As a result, the optimal speed for HFT i declines as her opponent becomes faster (Proposition 7).

Intuitively, the exogenous cost is sunk, and HFT i must pay it anyway. By contrast, her speed investment pays out only if her sniping attempt is fulfilled. Therefore, if HFT i thinks she is less likely to snipe due to a faster opponent (a lower π_i), the exogenous cost becomes more salient (C/π_i increases), hampering her speed investment.

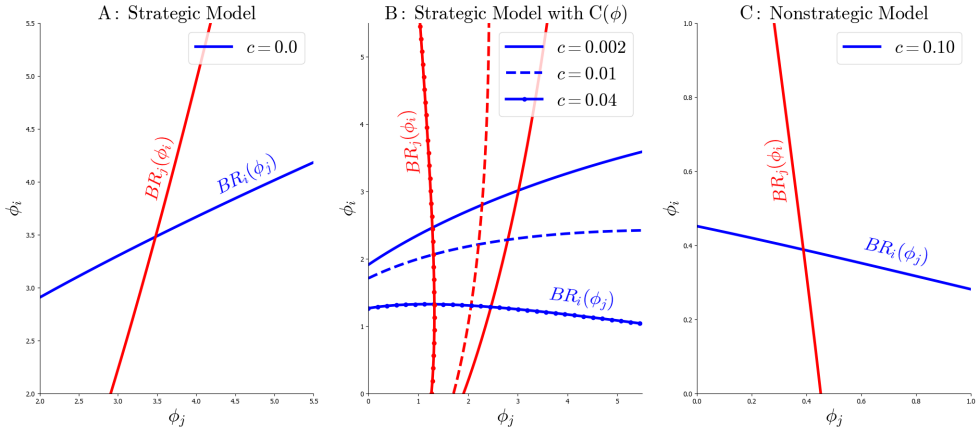
This logic can be applied to analyze a speed bump as well: it makes sniping less likely, leaving HFTs reluctant to pay the sunk cost. This is why traditional models conclude that a speed bump is effective to slow HFTs down and to mitigate adverse selection, which is replicated by point (ii) in Proposition 7.

In the following subsection, I analyze the role of the exogenous sunk cost in my strategic model. It explores when one of the two opposing results of a speed bump in the existing models (Proposition 7) and my strategic model (Propositions 2 and 5) becomes dominant.

4.2 Strategic complementarity versus substitution

Consider an extension of the *strategic* model by introducing the exogenous sunk cost of speed, $C(\phi_i)$. This extension shows that the exogenous cost tries to generate strategic *substitution*, while an endogenous cost (i.e., the bid-ask spread) promotes strategic *complementarity*. Thus, adjusting these two effects explains both complementarity and substitution in an arms race, as well as the positive, negative, and insignificant reactions of the spread to a speed bump.

Figure 6: Best Response with Exogenous Cost



Note: This figure plots the best response functions with different $C(\phi)$.

The optimization problem for HFT i is

$$\max_{\phi_i} W_i = \pi(\phi_i, \phi_j)(\sigma - s(\phi_i, \phi_j)) - C(\phi_i),$$

$$\text{s.t., } s(\phi_i, \phi_j) = \frac{\frac{\phi_i}{1+\lambda(\phi_i+\phi_j+\beta)}}{\beta + \frac{\phi_i}{1+\lambda(\phi_i+\phi_j+\beta)}}\sigma,$$

with $C(\phi) = c\phi_i$ and π given by (16).

Best response functions

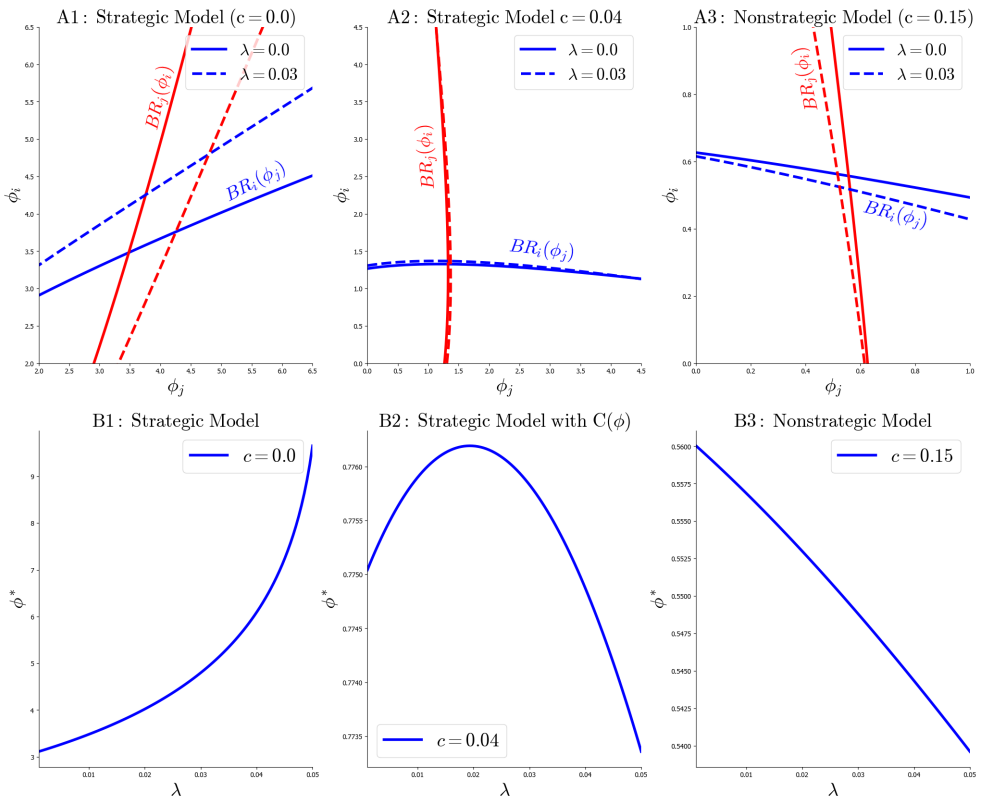
The best response functions are provided in Figure 6, where the exogenous cost C is adjusted by variating c. The strategic model in Subsection 3.1 corresponds to $c = 0$ (Penal A). As conjectured, a small exogenous cost (e.g., $c = 0$, 0.002, and 0.01) offers strategic complementarity because the strategic motive relative to the exogenous cost is still strong. If the cost becomes large (e.g., $c = 0.04$ in Panel B), the best response tends to be hump-shaped and eventually slopes downward (Panel C). This is because a faster opponent makes the exogenous sunk cost more salient, and paying it for acquiring speed becomes less attractive.

Therefore, an arms race exhibits both complementarity and substitution, i.e., the BR curves become hump-shaped. When c is small enough, the extended model is close to the strategic model with complementarity, while a large c makes it similar to the non-strategic model with strategic substitution.

Impact on speed competition and equilibrium speed

Figure 7 describes how the best-response functions and the equilibrium speed behave when the length of a delay becomes longer. The case with $c = 0$ (Panels A1 and B1) follows the analytical result in Proposition 5. If $c > 0$, the impact of increasing λ resembles the impact of increasing c because both of them make an *effective* exogenous cost (C/π) large. This negative channel arises

Figure 7: Impact of λ on BR and ϕ^*



Note: Panels A1-A3 in this figure plot the best response functions with different exogenous cost parameter, c, and see the impact of a longer delay ($\lambda = 0 \to 0.03$). Panels B1-B3 show the speed level at the (stable) symmetric equilibrium as a function of λ.

only if $c > 0$ and competes against the positive strategic effect (in Panels A1 and B1) that tries to push ϕ^* up. If c is intermediate, these two effects may cancel out, as in Panel A2 of Figure 7, whereas the negative channel can dominate the positive strategic effect if c increases even more (Panel A3). This argument has a direct implication for the equilibrium speed: it becomes hump-shaped with an intermediate cost c (Panel B2) and slopes downward when it is large (Panel B3).

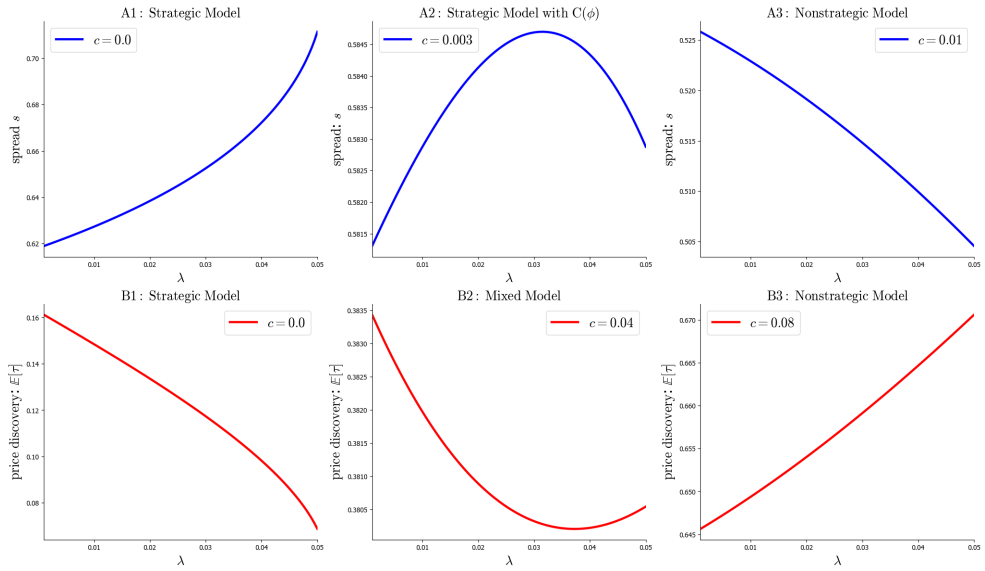
Impact on market quality

Figure 8 illustrates the impact of λ on market quality: the equilibrium spread and the measure of price discovery, $\mathbb{E}[\tau]$.

Remember that a higher λ directly mitigates adverse selection and tightens the spread. When c is sufficiently small, at the same time, it also promotes the speed of HFTs and widens the spread. As Proposition 6 shows, a strong complementarity (or externality) significantly facilitates ϕ^* so that the spread increases, as Panel A1 in Figure 8 shows. However, this effect diminishes as the exogenous cost c rises, making s hump-shaped and slope downward (Panels A2 and A3).

Panels B1-B3 plot the measure of price discovery, $\mathbb{E}[\tau]$, defined by (21). It takes the opposite reaction to ϕ^* and s. This is intuitive: the faster the HFTs, the faster the price discovery process becomes.

Figure 8: Impact of λ on market quality



Note: This figure plots the optimal speed at the symmetric equilibrium with different $C(\phi)$.

Overall, the strategic model with complementarity and the non-strategic models with substitution can be seen as two extremes. This subsection puts together these two structures and shows some intermediate cases, where a speed bump λ may have a positive, negative, or insignificant impact on market quality. Section 6 proposes empirical implications of this premise.

5 Discretionary liquidity traders

The analyses so far have established that a speed bump facilitates HFTs' speed. A speed bump exogenously reduces the arrival rate of HFTs but generates some favorable impacts on HFTs' marginal benefit by changing the market makers' pricing behavior. In this section, we argue that a change in the liquidity traders' arrival rate, β, may have the same implication as a speed bump because an increase in β reduces the relative arrival rate of the HFTs.

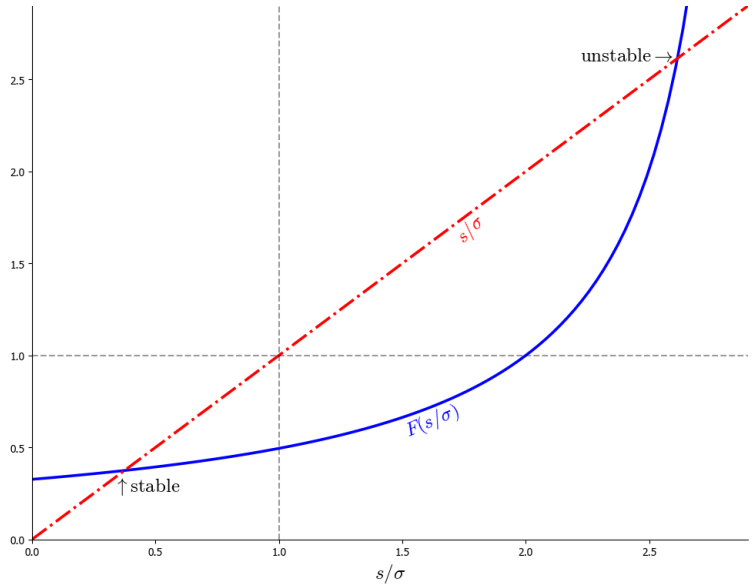
To adopt this idea, I consider price-elastic liquidity traders, i.e., discretionary liquidity traders (DLTs), as Admati and Pfleiderer (1988) and Zhu (2014) have analyzed.[20] Importantly, the price-elastic LTs amplify the dark-side effect of a speed bump; A longer expected delay (λ) endogenously induces large β because λ exogenously reduces the level of the spread s and makes it easier for the liquidity traders to participate. A large set of LTs, in turn, facilitates HFTs' investments in speed technologies.

5.1 Liquidity traders' behavior

Consider an extension of the benchmark model. There is a unit mass of potential liquidity traders (LTs). They are exposed to a liquidity shock that happens at random date $T_L \sim \exp(\hat{\beta})$. The shock

[20] I thank Sophie Moinas for suggesting this extension.

Figure 9: Spread with DLT



Note: This figure illustrate the determination of the spread with DLT. The blue solid curve represents $F(s/\sigma)$ with linear $G(\alpha) = \alpha/2$, and the red dashed line is the 45-degree line.

gives each LT a type denoted by $\alpha \in [0, \infty]$, which represents a desire for immediacy. α is drawn from density $g(\alpha)$ with $G(x) \equiv \int_0^x g(\alpha)d\alpha$. Regardless of the true value of the asset v, a liquidity trader with immediacy α obtains utility $u(\alpha) = \alpha\sigma - s$ if she buys at ask price s and $\alpha\sigma + (-s)$ if she sells at bid price $-s$. As these give the same utility, an LT with $u(\alpha) \geq 0$ tries to fulfill her need to trade by randomly buying or selling a single unit of asset.

Therefore, given the quoted spread s, the liquidity shock makes LTs with $\alpha \geq \frac{s}{\sigma}$ want to trade. Then, by incorporating the measure of *active* LTs, their behavior is described by the Poisson arrival with rate $\beta(s) \equiv \hat{\beta}\left[1 - G\left(\frac{s}{\sigma}\right)\right]$.

Equilibrium spread

The market makers competitively determine the spread s, as in the benchmark model. Thus, the (effective) equilibrium spread must satisfy

$$\frac{s}{\sigma} = \frac{\frac{\phi}{1+\lambda(\phi+\gamma)+\lambda\beta(s))}}{\beta(s) + \frac{\phi}{1+\lambda(\phi+\gamma)+\lambda\beta(s))}} \equiv F(\frac{s}{\sigma}, \phi). \tag{24}$$

Note that the arrival rate of the LTs is endogenous and affected by s. Hence, the equilibrium spread is determined by the fixed point problem (24).

It is easy to check that F is increasing in s, and (24) generates positive feedback. The feedback represents the "liquidity begets liquidity" argument: a narrow spread reduces the trading cost and attracts a large set of liquidity traders, which, in turn, mitigates adverse selection for the market makers and reduces the spread even more.

As Figure 9 shows, this feedback generates multiple equilibria. Following Proposition 5 and

discussions after that, I focus on the *stable* solution of (24). As F denotes the required spread for market makers to break even, if the realized spread is smaller than F, a market maker faces a shortfall. Thus, the equilibrium spread must be higher, i.e., s tends to go up when $s < F(s)$. The symmetric argument implies that s goes down when $s > F(s)$. Therefore, the stable solution s^* is realized at s where $F(s)$ crosses the 45-degree line from above. By incorporating the dependence of β on s, the equilibrium spread (i.e., the stable solution of [24]) is denoted as[21]

$$s^* = s(\phi, \lambda).$$

Corollary 2. *The direct effect of a speed bump on the equilibrium spread s^* is negative, i.e., $\frac{\partial s^*}{\partial \lambda} < 0$.*

Proof. It is straightforward from $\frac{\partial F}{\partial \lambda} < 0$. □

Once again, a longer speed bump directly relaxes the adverse selection problem due to the same mechanism as in the benchmark model. However, the indirect effect via the speed choice $\frac{d\phi}{d\lambda} > 0$ survives and must compete against the direct effect.

5.2 Speed choice by HFT

The sniping probability of the HFT has the same form as eq. (5), while it is affected by the spread via $\beta(s)$. Thus, I denote it by $\pi(\phi, \lambda, \beta(s))$. As well, the optimization problem by the HFT takes the same form as the benchmark model (8), where the marginal effect of being faster is now given by

$$\frac{dW}{d\phi} = \underbrace{(\sigma - s^*)\frac{\partial \pi}{\partial \phi} + \pi \frac{d(\sigma - s^*)}{d\phi}}_{\text{original MB vs. MC}} + \underbrace{\frac{ds^*}{d\phi}\frac{d\beta^*(s)}{ds}\frac{\partial \pi}{\partial \beta}}_{\text{additional terms} > 0}. \tag{25}$$

The first term captures the marginal benefit (MB) versus marginal cost (MC) that is inherited from the benchmark model (see [9]). In these effects, the impact of ϕ on s^* (or $\sigma - s^*$) is magnified by the "liquidity begets liquidity" argument, that is,

$$\frac{ds^*}{d\phi} = \frac{\frac{\partial F(s^*)}{\partial \phi}}{1 - \frac{d\beta}{ds}\frac{\partial F(s^*)}{\partial \beta}}. \tag{26}$$

The numerator is the direct impact of speed on the break-even spread: a market maker faces severe adverse selection and charges a wider spread, $\frac{\partial F(s^*)}{\partial \phi} > 0$. Besides, the denominator exhibits the positive feedback between the trading cost s^* and the LTs' behavior $\beta(s)$ (i.e., $\frac{d\beta}{ds}\frac{\partial F(s^*)}{\partial \beta} < 0$).

Moreover, the second term in (25) is the new marginal benefit of being faster that stems from the DLTs. An increase in the HFT's speed widens the spread and hampers the participation of the liquidity traders, thus leading to a higher sniping probability for the HFT.

These two factors imply the following:

Corollary 3. *An increase in the price elasticity of the DLTs (i.e., $\frac{d\beta}{ds}$) promotes the HFT's investment into the speed technology.*

[21] The existence of the fixed point solution is easy to show. Firstly, $\beta(0) = \hat{\beta} < \infty$ indicates that $F(0) > 0$. Also, at $s = \sigma$, we have $\beta(\sigma) = \hat{\beta}[1 - G(1)]$ and $F(1) < 1$. These imply that (24) has at least one stable solution in $\frac{s}{\sigma} \in [0, 1]$.

5.3 Impact of a speed bump

How do the DLTs alter the impact of a speed bump? The cross-derivative of the FOC is given by

$$\frac{\partial}{\partial\lambda}\frac{dW}{d\phi} = \underbrace{\frac{\partial(\sigma - s^*)}{\partial\lambda}\frac{\partial\pi}{\partial\phi}}_{\text{effect (i)}>0} + \underbrace{(\sigma - s^*)\frac{\partial^2\pi}{\partial\lambda\partial\phi}}_{\text{effect (ii)}<0} + \underbrace{\frac{\partial\pi}{\partial\lambda}\frac{d(\sigma - s^*)}{d\phi}}_{\text{effect (iii)}>0} + \underbrace{\pi\frac{\partial}{\partial\lambda}\frac{d(\sigma - s^*)}{d\phi}}_{\text{effect (iv)}>0} + \underbrace{\frac{\partial}{\partial\lambda}\left(\frac{ds^*}{d\phi}\frac{d\beta^*(s)}{ds}\frac{\partial\pi}{\partial\beta}\right)}_{\text{additional effect}}.$$

The first four effects represent the identical channels to the benchmark model shown by (12), whereas the price elastic β adds new implications.

For example, in effect (i), λ affects the spread not only through the direct channel but also through a change in β, thereby amplifying the direct impact. Due to the same mechanism as (26), this effect is expressed as

$$\frac{\partial s^*(\phi,\lambda)}{\partial\lambda} = \underbrace{\frac{\overbrace{\frac{\partial F}{\partial\lambda}}^{\text{direct effect}}}{1 - \beta'(s)\frac{\partial F}{\partial\beta}}}_{\text{indirect effect}} < 0. \tag{27}$$

Similarly, λ has its direct impact on effects (ii)-(iv), as well as indirect impacts by reducing the spread s^* and attracting a larger set of DLTs. As these indirect impacts must go through a change in the spread (27), they cannot dominate effect (i). As a result, introducing the DLTs helps λ with pushing up the marginal profit of being faster, $\frac{dW}{d\phi}$.

5.4 Numerical results

The fixed point problem in (24) is analytically intractable. Thus, I conduct numerical analyses to see the equilibrium impact of a speed bump.

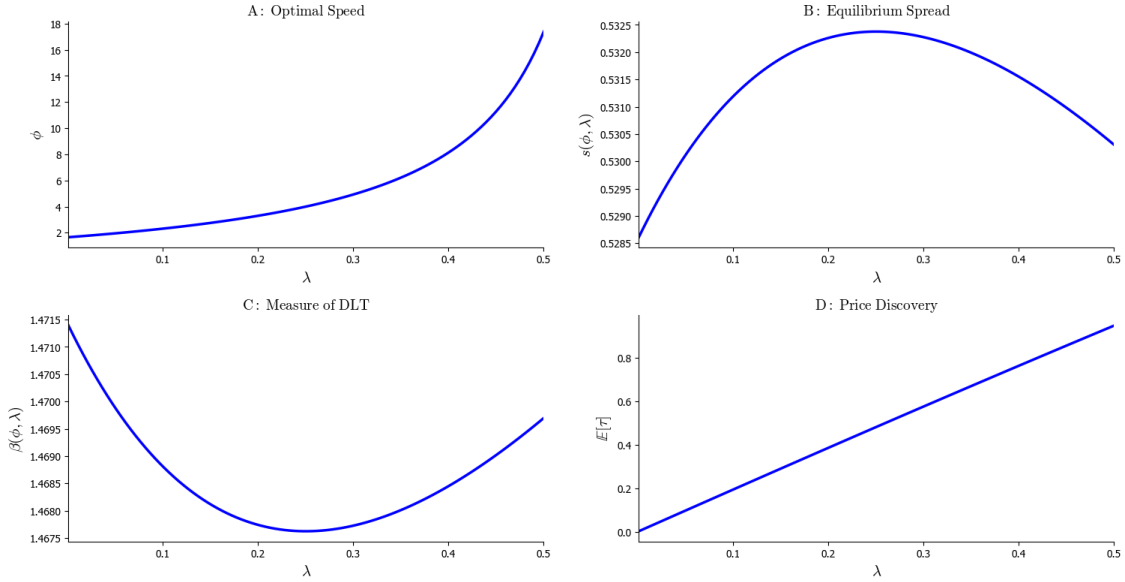
The benchmark model with DLTs

First, Figure 10 shows the impact of a longer delay (λ) on the optimal speed (Panel A), the equilibrium spread (Panel B), the measure of DLTs (Panel C), and the expected time lapse until price discovery (Panel D).

As predicted, the optimal speed for the HFT is monotonically increasing in λ (Panel A). A speed bump directly reduces the spread and attracts more liquidity traders. A large set of liquidity traders, in turn, makes market makers care less about the speed-up by the HFT, thereby generating further room for speed-up.

Interestingly, and in contrast to the benchmark result in Proposition 3, the equilibrium spread is

Figure 10: Impact of λ with DLTs



Note: This figure illustrates the impact of λ on the equilibrium speed (Panel A), spread (Panel B), the measure of DLTs (Panel C), and the measure of price discovery (Panel D). I assume linear $G(\alpha) = \alpha/2$.

hump-shaped. From (24), (26), and (27), the implicit function theorem implies that

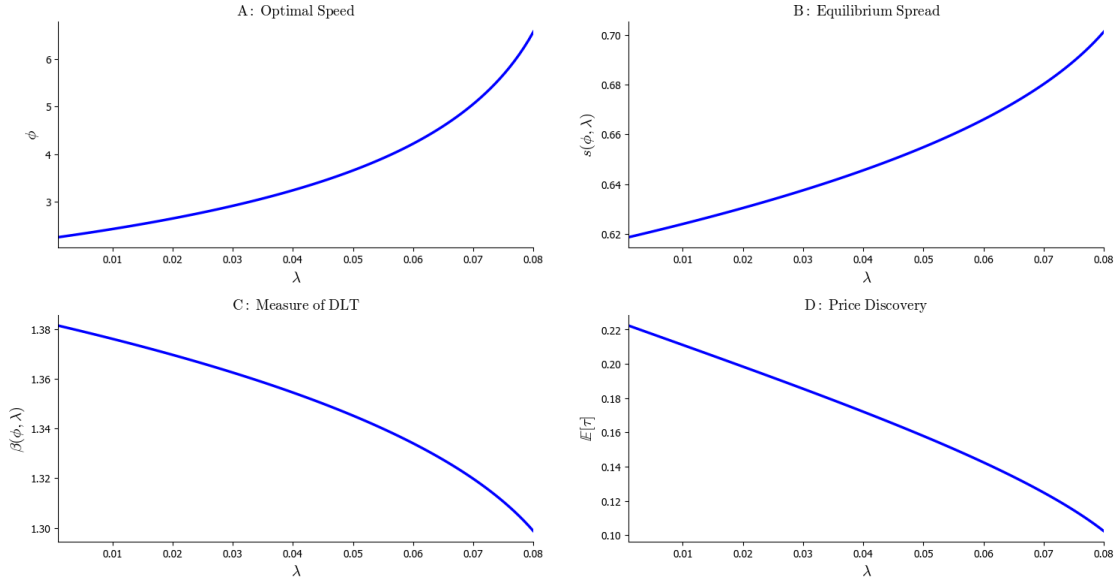
$$\frac{ds^*}{d\lambda} = \frac{\overbrace{\frac{\partial F}{\partial \lambda}}^{<0} + \overbrace{\frac{d\phi^*}{d\lambda}\frac{\partial F}{\partial \phi}}^{>0}}{1 - \beta'(s)\frac{\partial F}{\partial \beta}}. \tag{28}$$

Therefore, even though the price-elastic LTs (i.e., $\beta'(s) < 0$) quantitatively affect the slope $\frac{ds^*}{d\lambda}$, whether the spread increases with λ is determined by the numerator of (28). Namely, it is determined by the direct protection of market makers by a speed bump versus its positive impact on the HFT's speed.

In the benchmark model, there are no externalities to be incorporated by the monopolistic HFT so that she can pick ϕ^* that makes these two competing effects offset each other. In contrast, the optimal speed ϕ^* is affected by the liquidity externality, i.e., the discretionary LTs, in this extension. Therefore, in general, two effects in the numerator in (28) do not cancel out.

When the length of a delay (λ) is short, the break-even spread (F) strongly reacts to a marginal increase in the speed (ϕ) because market makers know that the arrival of the HFT remains high. At the same time, a small λ induces a small ϕ in the equilibrium (see Panel A in Figure 10). Then, the impact of increasing λ on the break-even spread tends to be weak: extending a speed bump does not pay out because the HFT is unlikely to arrive even without a speed bump. Therefore, in a small-λ region, the first term in (28) is weak and dominated by the second term, making equilibrium spread s^* increasing in λ. Due to the symmetric argument, the reaction of s^* to λ becomes hump-shaped, as in Panel B of Figure 10. Note that s^* has the direct implication for the measure of the DLTs in Panel C.

Figure 11: Impact of λ with DLTs in Multiple HFTs Model



Note: This figure illustrates the impact of λ on the equilibrium speed (Panel A), spread (Panel B), the measure of DLTs (Panel C), and the measure of price discovery (Panel D). I assume linear $G(\alpha) = \alpha/2$.

Concerning Panel D, increasing the length of a delay deteriorates the price discovery process. Even though the HFT leverages her speed technology ϕ to inject new information into the price, a large λ induces a large set of liquidity traders too. As the liquidity-driven traders do not convey any information, they slow down the price discovery. Thus, it becomes difficult for ϕ to dominate a extended delay and to improve the price discovery process.

Multiple HFTs with DLTs

The model with multiple HFTs and fast market making in Section 3 brings about the prediction that an increase in ϕ would be stronger than the benchmark model due to the competition with strategic complementarity.

Figure 11 shows the set of results with multiple HFTs with DLTs. The model is analogous to the benchmark model, but two changes must be applied: public news arriving with rate γ is replaced by the HFM's arrival with rate ϕ_M, and the equilibrium ϕ^* is determined by the interaction between the best-response functions, as in Section 3.

Due to the strategic complementarity in an arms race, the speed improves significantly when a speed bump becomes longer (Panel A of Figure 11). This speed-up by the HFTs is strong enough to dominate the direct impact of λ on the spread, thus making s^* monotonically increasing in λ (Panel B). This, in turn, has a feedback effect with the discretionary liquidity traders. Even though a longer delay tries to reduce the spread and attract more LTs, the faster HFTs deteriorate the trading cost and cause the outflow of LTs (Panel C). Finally, as in the model with fixed β, the significantly fast HFTs contribute to reducing the time for price discovery, $\mathbb{E}[\tau]$ (Panel D).

6 Discussion

This section provides empirical implications and policy discussion of my theory.

Empirical implications

My model provides some testable implications for the strategic nature of an arms race among HFTs and for the effect of speed bumps on a spread, adverse selection for market makers, and price discovery.

As Subsection 4.2 demonstrates, an arms race exhibits strategic complementarity or substitution depending on the exogenous cost of speed and the expected length of a speed bump. A long (resp. short) expected speed bump or a significant (resp. slight) exogenous cost of speed creates strategic substitution (resp. complementarity), and the introduction of a speed bump and a marginally longer delay would effectively reduce (resp. widen) a spread, meaning that a speed bump mitigates adverse selection.

In the real world, the exogenous speed costs for HFTs (i.e., $C(\phi)$) involve two factors. The first cost is a large sunk cost to develop a high-speed communication technology or to investigate the optimal location of information servers to exploit an arbitrage between segmented markets. For example, it is well known that Spread Networks LLC invested about $300 million to *re*construct a fiber-optic network between Chicago and New York exchanges.

In contrast, as of 2018, the most common way to acquire speed is to pay subscription fee to get access to these communication technologies developed by telecom companies or to colocate an information server to an exchange platform. Subscription fees can be relatively small compared to the aforementioned cost. For example, Mckay Brothers and Quincy Data provide fast data access via their lowest-latency communication technology at about $3,000∼$10,000/month. To obtain the fast data feed on 8 securities (such as SPY and IWM), for instance, one has to pay $3,300 per month, or $19.64 (≈$3,300/8*21) per trading day and security. If there are 800 arbitrage opportunities per day on average (as Budish et al., 2015 estimate regarding SPY), the cost for speed per transaction amounts to $C = \$0.024$. This is small relative to the profit of taking arbitrage, which is estimated to be about $\sigma - s = \$98.02$ (Budish et al., 2015).

Depending on the condition of financial markets, my model suggests different implications. If HFTs take the first investment approach, the traditional model fits better: an arms race involves strategic substitution and a speed bump is effective. In contrast, if the exogenous cost is relatively small compared to the total profit, my strategic model is more appropriate, suggesting the complementarity and detrimental effects of a speed bump. Of course, my model has a ϕ-dependent cost (i.e., $C = c\phi$) and cannot apply these back-of-envelope calculations directly. However, I believe that the scale of the exogenous cost of speed normalized by the profit (σ) should be sufficiently small, such as $C/\sigma = \$0.00025$ ($\approx \$0.024/98.02$), to justify the validity of the strategic model.

This comparison can also be applied to the market power of a high-frequency financial institution because the endogenous cost of speed stems from the strategic motives of HFTs. In other words, when N is large, the adverse price impact of increasing ϕ_i diminishes, leaving the model close to the traditional one.

Also, to test the implications of λ on market quality, the length of a speed bump or the distri-

bution of a random delay can be investigated. If a speed bump is expected to be long, a marginally longer delay can be effective and slows HFTs down due to the strategic substitution in the traditional model. On the other hand, if a platform tries to avoid a delay cost by keeping the length of a delay minimal, the introduction of a speed bump can aggravate adverse selection, as HFTs become faster. The lengths of speed bumps in each platforms (i.e., bumps at the ANEO and TMX are longer than those at the IEX or Chicago Stock Exchange) can be easily compared, and my model shows that the *level* of λ matters. Deriving the critical (c, λ) involves numerical calculation, and estimating it requires further data on the cost of the speed technologies, speed levels, profit, and the market power of high-frequency financial institutions. Thus, I leave this as a topic for the future research.

Policy discussion

Recently, exchange platforms have experienced declines in their revenue from transaction fees (e.g., listing fees and maker-taker fees). Instead, they charge an increasingly high price to the fast access to their data, such as direct data feed and colocation of data servers. Importantly, one of the suppliers of the speed technologies, which I did not specify in the model, can be exchange platforms. The SEC is concerned about the skyrocketing price for fast data feed and has issued a proposal to block exchanges to raise fees in March 2018.

Another HFT-related policy that the SEC has adopted is the approval of the IEX (with a speed bump) as a National Securities Exchange (NSE) in 2017. Due to the Reg. NMS and Order Protection Rule, this approval effectively makes traders affected by a speed bump.

In a nutshell, the SEC tries to curb the price of speed technologies supplied by exchange platforms, while it prompts the introduction of a speed bump. Importantly, my model suggests that the introduction of a delay does not necessarily conflict with a provision of expensive speed technologies by an exchange platform. That is, a speed bump can increase HFTs' demand for fast information access, allowing an exchange platform to charge a higher price. Thus, aforementioned policies adopted by the SEC can be self-negating.

In this situation, we can analyze whether "the market will fix the market," as Budish et al. (2018) put forth. They argue that a platform does not have an incentive to introduce frequent batch auctions, as long as competing platforms can copy the innovation at a small cost. In my model, a speed bump does not always mitigate adverse selection, while an exchange platform may have an incentive to introduce it to facilitate demand for speed technologies. More detailed analyses can be my future research topic, but the mechanism in my model can provide another explanation for why "the market cannot fix the market."

7 Conclusion

A speed bump, which seeks to mitigate adverse selection for market makers, can backfire. When HFTs strategically choose their speed level by considering the impact of their speed decision on market behavior, a bid-ask spread charged by market makers works not only as a trading cost but also as an endogenous cost of speeding up, as it widens as HFTs becomes faster. The bid-ask spread tends to be insensitive to the HFTs' speed-up if a speed bump is long. This is because market makers with a long speed bump behave as if the share of HFTs is small and do not care much about HFTs'

speed-up. Then, a speed bump or a longer delay diminishes the marginal cost of being faster, leading to a higher equilibrium speed of HFTs.

My model shows that an arms race between HFTs can exhibit strategic complementarity and involves positive externality. As a result, the positive impact of a speed bump on the HFTs' speed tends to be strong. Hence, regulating fast informed trading by a speed bump deteriorates the adverse selection problem, while the price discovery process is promoted.

This paper describes how fast HFTs obtain private information and act on it. Although this follows the literature (Budish et al., 2015; Foucault et al., 2016; Haas and Zoican, 2016; Brolley and Zoican, 2019) and provides a tractable framework, it ignores the other dimension of information, namely, the precision of a signal. Several studies, such as Huang and Yueshen (2018) and Dugast and Foucault (2018), argue that the precision of private information and the information processing speed may have a negative relationship. Hence, analyzing the impact of a speed bump on the information choice, both in terms of speed and precision, would be a good extension of my model.

Moreover, I have not analyzed the problem of the market design, i.e., what is the optimal structure of speed bumps to improve market quality? As well, while the speed regulation is a good representation of my model, its insight is broader and more general. It can be applied to other forms of regulations, such as policies on fair disclosure, transparency, and sanctions on insider trading. Regulation, in general, can backfire if it aims to mitigate asymmetric information or adverse selection and renders market makers less responsive to informed traders' information acquisition. As my model indicates that restricting informed trading facilitates the information acquisition, it goes counter to the literature that has argued for the "crowd-out" effect of information disclosure (i.e., fair information disclosure disincentivizes acquisition of private information).[22] I leave this as a topic for future research.

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.

Aït-Sahalia, Yacine and Mehmet Saglam, "High frequency traders: Taking advantage of speed," Technical Report, National Bureau of Economic Research 2013.

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

Baldauf, Markus and Joshua Mollner, "High-frequency trading and market performance," *SSRN Electronic Journal*, 2017.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, "Equilibrium fast trading," *Journal of Financial Economics*, 2015, *116* (2), 292–313.

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

Bongaerts, Dion and Mark Van Achter, "High-frequency trading and market stability," *SSRN Electronic Journal*, 2016.

[22] See, for example, a survey by Goldstein and Yang (2017).

Brogaard, Jonathan, Björn Hagströmer, Lars Nordén, and Ryan Riordan, "Trading fast and slow: Colocation and liquidity," *The Review of Financial Studies*, 2015, *28* (12), 3407–3443.

Brolley, Michael and David A Cimon, "Order Flow Segmentation, Liquidity and Price Discovery: The Role of Latency Delays," *SSRN Electronic Journal*, 2017.

__ **and Marius Zoican**, "Liquid speed: On-demand fast trading at distributed exchanges," *arXiv preprint arXiv:1907.10720*, 2019.

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**, "Will the Market Fix the Market? A Theory of Stock Exchange Competition and Innovation," *Manuscript in Preparation*, 2018.

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

Conrad, Jennifer, Sunil Wahal, and Jin Xiang, "High-frequency quoting, trading, and the efficiency of prices," *Journal of Financial Economics*, 2015, *116* (2), 271–291.

Delaney, Laura, "Investment in high-frequency trading technology: A real options approach," *European Journal of Operational Research*, 2018, *270* (1), 375–385.

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

Dugast, Jérôme and Thierry Foucault, "Data abundance and asset price informativeness," *Journal of Financial Economics*, 2018, *130* (2), 367–391.

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.

__ **, 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.

Frino, Alex, Vito Mollica, and Robert I Webb, "The impact of co-location of securities exchanges' and traders' computer servers on market liquidity," *Journal of Futures Markets*, 2014, *34* (1), 20–33.

Glosten, Lawrence R 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.

__ **and Talis J Putnins**, "Welfare costs of informed trade," *Columbia Business School Research Paper*, 2016, pp. 16–58.

Goldstein, Itay and Liyan Yang, "Information disclosure in financial markets," *Annual Review of Financial Economics*, 2017, *9*, 101–125.

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

Hasbrouck, Joel and Gideon Saar, "Low-latency trading," *Journal of Financial Markets*, 2013, *16* (4), 646 – 679.

Hendershott, Terrence and Haim Mendelson, "Crossing networks and dealer markets: Competition and performance," *The Journal of Finance*, 2000, *55* (5), 2071–2115.

__ **and Pamela C Moulton**, "Automation, speed, and stock market quality: The NYSE's hybrid," *Journal of Financial Markets*, 2011, *14* (4), 568–604.

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

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

Huang, Shiyang and Bart Z Yueshen, "Speed Acquisition," *Available at SSRN 2845864*, 2018.

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 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.

Lewis, Michael, *Flash boys: a Wall Street revolt*, WW Norton & Company, 2014.

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.

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.

Riordan, Ryan and Andreas Storkenmaier, "Latency, liquidity and price discovery," *Journal of Financial Markets*, 2012, *15* (4), 416–437.

Rosu, Ioanid, "Dynamic Adverse Selection and Liquidity," in "Paris December 2018 Finance Meeting EUROFIDAI-AFFI" 2018.

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

Ye, Mao, Chen Yao, and Jiading Gai, "The externalities of high frequency trading," *SSRN Electronic Journal*, 2013.

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

A Different market structures

A.1 Speed bump on liquidity traders

Suppose that the liquidity shock can occur even before the jump at $t = 0$. This is natural because the liquidity traders are motivated by the trading needs rather than the trading profit. To this end, let T_L has density $\beta e^{-\beta(t+\kappa)}$ where $\kappa \geq 0$ captures the starting point of the random process.

If the liquidity traders are subject to a speed bump, the density of their arrival at t (without other events happening) is $\beta e^{-\beta(t-\delta+\kappa)}$. By using this, the expected profit of market makers is modified as follows:

$$V = \mathbb{E}_\delta \left[\int_0^\delta s\beta e^{-(\beta+\gamma)t} e^{-\beta(\kappa-\delta)} dt + \int_\delta^\infty (\beta s + \phi(s-\sigma)) e^{-\psi(t-\delta)} e^{-(\beta+\gamma)\delta} e^{-\beta(\kappa-\delta)} dt \right],$$

which yields the following break-even spread,

$$s = \frac{\frac{\phi}{1+\lambda\psi a(\lambda)}}{\beta + \frac{\phi}{1+\lambda\psi a(\lambda)}} \sigma$$

with $a(\lambda) \equiv \frac{\lambda}{(1-\beta\lambda)(\beta+\gamma)}$. Therefore, if I rewrite $\hat{\lambda} \equiv \lambda a(\lambda)$, the result in this setting becomes the same as the benchmark model.

In this setting, the expected profit of the HFT is given by

$$W(\phi) = \pi(\phi, \lambda)(\sigma - s)$$

with

$$\pi(\phi, \lambda) = e^{-\beta\kappa} \frac{\phi}{\psi} \frac{1}{1+\lambda\gamma}.$$

Thus, the structure of the optimization problem does not change. As well, by replacing λ by $\hat{\lambda}$, the optimal speed of the HFT is same as the benchmark model:

$$\phi = \frac{\sqrt{\beta+\gamma}(1+\hat{\lambda}(\beta+\gamma))}{1-\hat{\lambda}\sqrt{\beta(\beta+\gamma)}}.$$

Since $a(\lambda)$ is monotonically increasing in λ, analyzing the model by using $\hat{\lambda}$ instead of λ provides the same results.

A.2 Continuous update by market makers

I modify the benchmark case by allowing each market maker to update (cancel and resubmit) limit orders continuously before HFTs move. Other structures of the model is the same as the benchmark.[23]

Consider a market maker who updates her limit order by resubmitting competitive s_t. The competition

[23] The possibility that an informed HFT splits her orders across the time can be eliminated because executions in a part of the markets let other market makers know the arrival of the HFT and true information. This event triggers the cancellation of outstanding limit orders. I also abstract away from the possibility of a mixed strategy since each order from the HFT does not have price impact. As mentioned in Footnote 9, I can show that the mixed strategy is not an equilibrium because waiting is not credible.

drives $s_t = E[v|\text{trade at } t]$. Since δ is stochastic,

$$\begin{aligned} s_t &= \int_0^\infty be^{-b\delta} E[v|\text{trade at } t,\delta]d\delta \\ &= \int_t^\infty be^{-b\delta} E[v|\text{trade at } t,\delta]d\delta + \int_0^t be^{-b\delta} E[v|\text{trade at } t,\delta]d\delta \qquad (29) \\ &= 0 \times \int_t^\infty be^{-b\delta}d\delta + \int_0^t be^{-b\delta}\frac{\phi}{\phi+\beta}\sigma d\delta \qquad (30) \\ &= (1-e^{-bt})\frac{\phi}{\phi+\beta}\sigma. \end{aligned}$$

In (29), the first term represents the case that t is in the "safe interval," i.e., $0 < t < \delta$. Conditional on trade occurs, the market maker expects that $E[v] = 0$ because the trade must be against a liquidity trader, and it does not convey any information. This is why the first term in (30) bears 0. The second is the case that t is outside of the "safe interval," leading the conditional expected return to be the probability of the HFT arrival (times σ) in (30).

A speed bump has the same effect on the endogenous marginal cost as in the benchmark (the proof is omitted as it is straightforward):

Proposition 8. *(i)* $\frac{\partial s_t}{\partial b} > 0$, *and (ii)* $\frac{\partial}{\partial b}\left(\frac{\partial s_t}{\partial \phi}\right) > 0$.

Since $b = \lambda^{-1}$ represents the inverse of the expected length, a longer delay (i) directly mitigates adverse selection for market makers, but (ii) it makes the marginal cost (spread) less sensitive to speed-up by the HFT.

I impose an exogenous sunk cost to make the model well-defined. The optimization problem of the HFT regarding the speed is given by

$$\begin{aligned} \max_\phi W(\phi) &= E_\delta\left[\int_0^\infty \phi e^{-\psi t}e^{-\eta\delta}(\sigma - s_{t+\delta})dt\right] - \frac{c}{2}\phi^2, \\ &\text{s.t., } s_t = (1-e^{-bt})\frac{\phi}{\phi+\beta}\sigma. \end{aligned}$$

The sniping profit from sending market orders at t is given by $\sigma - s_{t+\delta}$ since they possibly arrive and executed at $t+\delta$. Note that, given that the trade occur, there is no price uncertainty since s_t is deterministic.

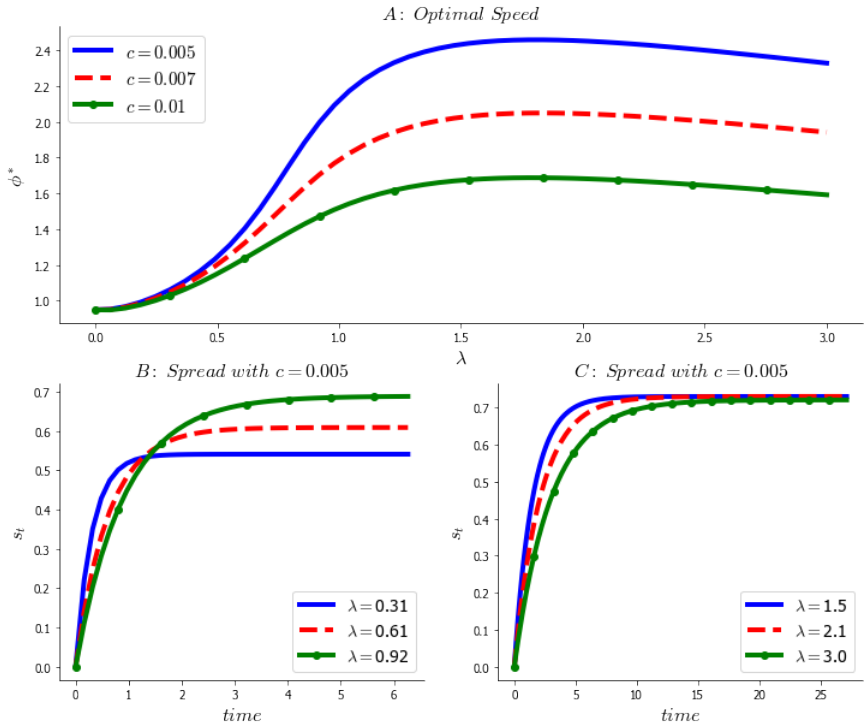
The behavior of the optimal speed and spread is hard to show analytically. However, the numerical solutions in Figure 12 can be discussed by using the ingredients I have already analyzed. In the single HFT economy, a speed bump positively affects the optimal speed ϕ^* through (i) decline in the marginal cost and (ii) increase in the sniping profit, while (iii) it reduces ϕ^* by magnifying the exogenous sunk cost. When λ is small, the execution risk is relatively small, leaving effect (iii) less significant compared to (i) and (ii), while a longer delay in the high-λ region makes (iii) more salient. As a result, ϕ^* takes the hump-shaped curve as depicted by Panel A in Figure 12. Given λ, a larger cost slows HFT down as a parallel shift in the curves suggests.

As (29) suggests, the spread is time dependent and increasing in t. This is because the market makers expect that they are less likely to be in the safe interval as t increases. The effect of $\lambda = b^{-1}$ is

$$\begin{aligned} \frac{ds_t}{d\lambda} &= -\frac{1}{\lambda^2}\frac{\partial s_t}{\partial b} + \frac{d\phi}{d\lambda}\frac{\partial s_t}{\partial \phi} \\ &= -\frac{1}{\lambda^2}te^{-bt}\frac{\phi}{\beta+\phi} + \frac{d\phi}{d\lambda}(1-e^{-bt})\frac{\beta}{(\phi+\beta)^2}. \qquad (31) \end{aligned}$$

First, a longer delay (higher λ) reduces the spread since market makers are directly protected by the longer safe interval. This is represented by the first term in (31). However, as a higher λ may push ϕ up or down, it increases or decreases the spread, as the second term in (31) suggests. When $\frac{d\phi}{d\lambda} > 0$, the second effect competes

Figure 12: Effect of λ on ϕ^* and s_t



Note: The panel A plots the optimal speed against λ with different values of c. The panels B and C plot the dynamics of the spread, s_t with different values of λ. The panel B is the case with the increasing optimal speed, $\frac{d\phi^*}{d\lambda} > 0$, and the panel C is the decreasing optimal speed, $\frac{d\phi^*}{d\lambda} < 0$.

with the first effect: the safe interval gets longer, while the HFT becomes faster. The result is provided by Panel B in Figure 12.

Whether the first effect dominates the second effect depends on t. From (31), the first negative effect is increasing in $t \leq b^{-1}$ and then starts decreasing. On the other hand, the second one is monotonically increasing in t and concave. There is a unique $t = \tau$, such that $\frac{ds_t}{d\lambda} > 0$ if and only if $t > \tau$, i.e., a longer speed bump increases the spread and worsens the adverse selection problem in the long-run.

The intuition is straightforward. When the current period t is $t > \tau$, the probability that t is in the safe interval is relatively small. Then, market makers think that an increase in λ has only a small effect to mitigate the adverse selection, while it increases the speed of the HFT.

When $\frac{d\phi}{d\lambda} < 0$, the result is given by Panel C in Figure 12. Since a bump reduces the speed in this case, the second effect helps the first effect reduce the spread and adverse selection cost. Once again, whether or not a speed bump increases the equilibrium speed depends on the exogenous cost, c, competing with the endogenous cost effect. Hence, the effect on the spread is governed by the market structure c and the time frame t.

A.3 Coexistence of fast and slow markets

In the real economy, the major market structure is still the continuous limit market with no speed bumps. Thus, the introduction of a speed bump inevitably makes these market structures coexist. However, analyses provided by the literature deal only with homogeneous markets. I extend my model to relax this limitation.

A.3.1 Environment

Consider the benchmark economy in Section 2. $q \in (0,1)$ fraction of market makers are in the market with a delay $\delta > 0$, which is stochastic, and the rest of them are in the market with no delay, $\delta = 0$. I call the first market the *slow market* and the latter one the traditional *fast market*. Each market is competitive. Market makers in the slow market submit limit orders with the (half) spread s_λ, while those who in the traditional fast market provide s_0. As in the benchmark, λ is the expected length of the delay. In contrast to the literature (Biais et al., 2015), I impose no restrictions on the venue choice by the HFT. Transactions information in each market becomes public right after an order execution, i.e., the markets are perfectly transparent.

A.3.2 Strategy of HFT

Consider a strategy of the HFT who becomes informed of $v = \sigma$ at date t. There are two possible (pure) trading strategies for the HFT. First, if she submits orders into the fast market, they are immediately executed, fulfilling $1-q$ of her total buying attempts. This market activity is publicly observable, allowing all market makers to realize that the transactions are information driven.[24] Based on this premise, market makers in the slow market can cancel their limit orders in the interval $\tau \in (t, t+\delta)$ which is protected by the speed bump. I call this the "strategy one" and denote it by $A = 1$.

Second, the HFT who becomes informed at t can immediately send market orders for q shares into the slow market, anticipating the execution with the δ-delay. She refrains from sending orders to the fast market at t and waits until the orders sent to the slow market are executed. By observing the execution in the slow market, she sends orders to the traditional fast market at $t+\delta$, which incur no delay by the construction. In this case, all of her orders arrive at the markets at the same time, $(t+\delta)$, and she can conceal her identity. Hence, none of the market makers can cancel their quotes. This "wait-and-grasp-all" strategy is denoted as $A = 2$.[25] Overall, taking $A = 2$ bears the execution risk, though the return from it is larger than $A = 1$ if accomplished.

The mixed strategy is the probability distribution over the set of actions $A \in \mathcal{A} = \{1,2\}$, and let $\theta_t \in [0,1]$ be the probability that the HFT takes the action $A = 2$. For $A \in \mathcal{A}$, let $w_A(t)$ be the expected profit from taking $A \in \mathcal{A}$ when the information arrives at date t. Figures 13 and 14 illustrate the timing of the executions when the HFT becomes informed at t.

First, $A = 1$ does not bear the execution risk because the HFT can immediately snipe limit orders in the fast market. However, this behavior becomes public immediately, allowing market makers in the slow market to cancel their orders. Thus, q fraction of quotes disappear, and the expected profit is given by

$$w_1(t) = (1-q)(\sigma - s_0).$$

On the other hand, $A = 2$ can snipe all outstanding liquidity at the same time, while it bears the execution risk that stems from the δ-delay. If there is a liquidity shock or the public news during $(t, t+\delta)$, the HFT cannot exploit her information and speed. Given that the HFT gets informed at date t, she obtains the profit with probability $\Pr(T_L > \delta, T > \delta) = e^{-(\beta+\gamma)\delta}$. Moreover, since the HFT is a price taker regarding her trading behavior, her expected return is

$$\begin{aligned} w_2(t) &= \int_0^\infty b e^{-(b+\beta+\gamma)\delta} \left[q(\sigma - s_\lambda) + (1-q)(\sigma - s_0)\right] d\delta, \\ &= \frac{q(\sigma - s_\lambda) + (1-q)(\sigma - s_0)}{1 + \lambda(\beta+\gamma)}. \end{aligned}$$

[24] This is because orders from liquidity traders will be fulfilled at the slow and the fast markets simultaneously.

[25] Note that making other lengths of strategic time lag is not optimal for the HFT, as any other intentional delay than δ tells that the orders are not from liquidity traders but from the HFT.

Figure 13: Strategy 1

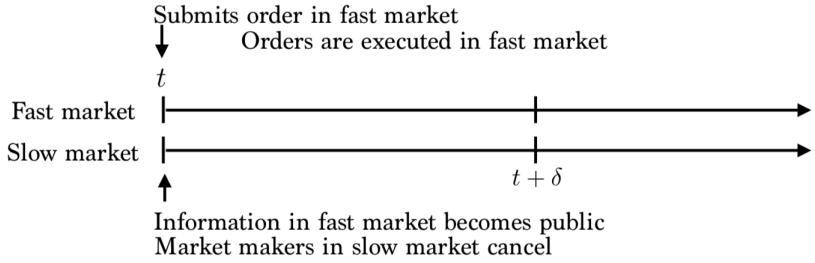
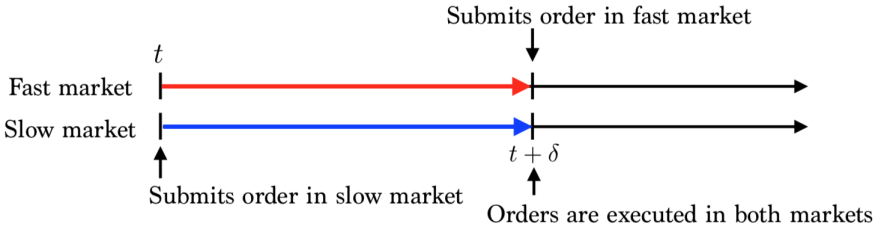


Figure 14: Strategy 2



Note that both of $\{w_j(t)\}_{j\in\mathcal{A}}$ are time independent due to the memoryless property of the exponential distribution. This implies that the optimal decision of $A \in \mathcal{A}$ is also time independent. No matter when the HFT becomes informed, a timing of private news does not matter—only the delay can be her concern.

A.3.3 Behavior of market makers

I take q an exogenous parameter in my model and assume that each market maker is randomly assigned the structure of the market. Given an assigned market, each market maker earn zero profit and does not have an incentive to move to another market with a different structure.

In the fast market

The market makers in the fast market suffer from adverse selection cost no matter what strategy the HFT takes. In other words, They are not given any chances to cancel orders by observing market-based information before the HFT snipes them. The expected return is

$$\begin{aligned} V_0 =& E_\delta \left[\theta \left(\int_0^\delta s_0 \beta e^{-(\beta+\gamma)t} dt + \int_\delta^\infty e^{-\psi(t-\delta)-(\beta+\gamma)\delta} (\beta s_0 + \phi(s_0-\sigma)) dt \right) \right. \\ & \left. + (1-\theta) \int_0^\infty e^{-\psi t} (\beta s_0 + \phi(s_0-\sigma)) dt \right]. \end{aligned} \tag{32}$$

The first line is the case that the HFT takes $A = 2$. In this case, the fast market is under the protection of the speed bump even though the speed bump is not applied to the fast market.[26] This is because the HFT takes "wait-and-grasp-all" strategy, and she does not snipe the fast market until she accomplishes her trading attempt in the slow market. The expected profit, in this case, is identical to those in the benchmark with

[26] Liquidity traders arrive at t if (i) the HFT becomes informed after t or (ii) the HFT becomes informed before t and takes $A = 2$. Given that the quote remains alive at t_-, $\Pr(T_H < t, A_{T_H} = 2 | \text{quote is alive at } t_-) = \Pr(T_H < t)$. Since, otherwise, the HFT arrives immediately at T_h and snipes the stale quotes. Therefore, $\Pr(\text{Liq. trade at } t) = \beta e^{-\psi t} + \beta e^{-(\beta+\gamma)t} \int_0^t \phi e^{-\phi\tau} d\tau = \beta e^{-(\beta+\gamma)t}$.

homogeneous markets. In the second line, the possibility of $A = 1$ is characterized. In this case, the speed bump does not matter for the fast market, and the conditional expected return is the same as a model with $\lambda = 0$.

In the slow market

In contrast to the fast market, the strategy of the HFT determines whether or not market makers in the slow market are protected. If $\theta = 0$, the slow market is perfectly protected by the speed bump: they can cancel their limit orders to avoid the HFT for sure. On the other hand, if $\theta \neq 0$, it is possible that the HFT arrives at the slow market to trade.

The expected return is

$$\begin{aligned} V_\lambda =& E\left[\theta\left(\int_0^\delta s_0\beta e^{-(\beta+\gamma)t}dt + \int_\delta^\infty e^{-\psi(t-\delta)-(\beta+\gamma)\delta}(\beta s_0 + \phi(s_0-\sigma))dt\right)\right. \\ &\left.+(1-\theta)\int_0^\infty s\beta e^{-\psi t}dt\right]. \end{aligned} \tag{33}$$

The intuitions behind the first line are the same as those in (32). As mentioned above, when the HFT takes $A = 1$, there is no chance for the HFT to snipe in the slow market. On the other hand, liquidity traders arrive at the slow market at t if $T_L = t$, $T_H > t$, and $T > t$, which gives the integrand in the second line.

A.3.4 Equilibrium in the trading stage

Let $Q \equiv (1-q)/q$. I first solve for the equilibrium spread given θ:

Proposition 9. *The equilibrium spread in fast and slow markets are given by*

$$s_j = \begin{cases} \dfrac{\phi}{\phi+\beta+\beta\psi\theta\frac{\lambda}{1+\lambda(\beta+\gamma)(1-\theta)}} & \text{for } j = 0, \\ \dfrac{\phi\theta}{\beta+\phi\theta+\lambda\beta(\beta+\gamma)\left(1+\frac{\phi\theta}{\beta+\gamma}\right)} & \text{for } j = \lambda. \end{cases} \tag{34}$$

Proof. Solving $V_j = 0$ yields the result. □

These formulae show the following:

Corollary 4. *θ affects s_0 and s_λ in an opposite way, that is,*

$$\frac{\partial s_0}{\partial \theta} < 0, \frac{\partial s_\lambda}{\partial \theta} > 0.$$

With ϕ fixed, s_0 is decreasing in θ, while s_λ is increasing in θ. For market makers in the fast market, a higher θ (i.e., the probability of $A = 2$) implies that the fast market is more likely to be protected by the speed bump in the slow market due to the HFT's "wait-and-grasp-all" strategy. This mitigates the adverse selection risk for the market makers in the fast market, making s_0 lower.

On the other hand, a higher θ (or $\phi\theta$) has a negative impact on market makers in the slow market. This is because a higher probability of $A = 2$ reduces the chance for market makers to observe sniping activity in the fast market to cancel the quote. Thus, a higher θ exposes the slow market to more severe adverse selection and pushes the spread s_λ up.

Now, the (mixed) strategy is characterized by the following, in which θ must satisfy the indifference condition, $w_1 = w_2$.

Proposition 10. *The optimal trading strategy for the HFT is*

$$\theta = \begin{cases} 0 & \text{if } \lambda Q(\beta+\gamma) > \frac{\phi+\beta}{\beta}, \\ \theta^* \in [0,1] & \text{if } \lambda Q(\beta+\gamma) \in [1, \frac{\phi+\beta}{\beta}], \\ 1 & \text{if } \lambda Q(\beta+\gamma) < 1, \end{cases} \tag{35}$$

with

$$\theta^* = \frac{(\phi+\beta) - \beta(\beta+\gamma)Q\lambda}{1+\lambda Q(\beta+\gamma) - \beta\lambda(1-\lambda\eta Q)} \frac{1+\lambda\eta}{\phi}. \tag{36}$$

Proof. See Appendix B.6. □

With ϕ fixed, the strategy θ of the HFT in the second stage game crucially depends on (i) the expected length of delay λ and (ii) the share of the slow market q. As proposed by (35), a higher λ and smaller q make the HFT reluctant to take $A = 2$ because both negatively impact the expected profit of $A = 2$ by imposing a higher execution risk and lower profit in the slow market, respectively. Thus, as λ or Q increases, θ^* declines and converges to 0. On the other hand, the HFT sticks to the strategy $A = 2$ (i.e., $\theta \to 1$) when λ or Q is sufficiently small.

A higher speed ϕ has two effects on the behavior of θ^*. First, it is straightforward that a higher ϕ widens the region for $\theta = \theta^* \in (0,1)$. Also, under the mixed strategy, the following result arises:

Corollary 5. *Ceteris paribus,* $\frac{\partial\theta^*}{\partial\phi} > 0$.

When the HFT becomes faster, the spreads in the fast and slow markets are differently affected. Since the slow market is more likely to be protected by the speed bump, a higher ϕ has a stronger effect on s_0 than s_λ. Moreover, as mentioned earlier, the slow market will face the HFT only if she takes $A = 2$ with probability θ. Hence, as (34) suggests, the effect of ϕ on s_λ is discounted by θ (i.e., ϕ affects s_λ via $\phi\theta$). Thus, when ϕ is high, the profit from the fast market shrinks more compared to the profit from the slow market. This induces the stronger incentive for the HFT to shift her priority towards the gain from the slow markets. Therefore, she tends to refrain from taking $A = 1$, and θ increases.

A.3.5 The optimal speed choice

Given the equilibrium in the trading stage, the HFT decides her speed level. Her objective function is denoted by

$$W(\phi) = \int_0^\infty \phi e^{-\psi t} w_A(\phi) dt$$

subject to

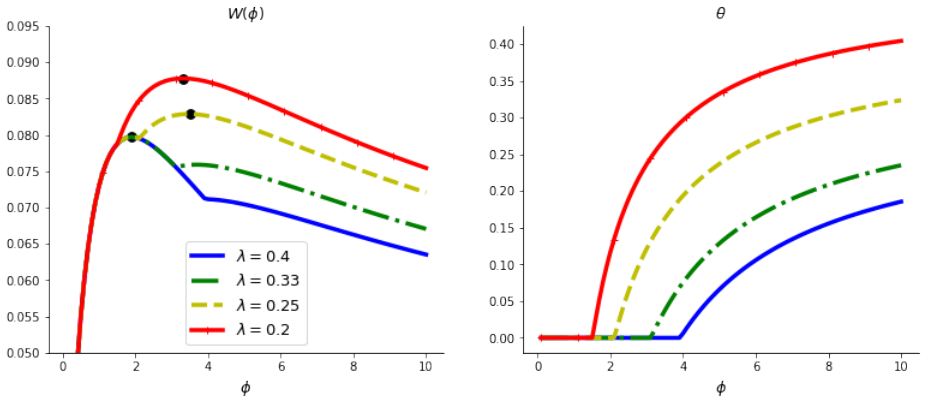
$$w_A(\phi) = \begin{cases} w_1(t) = (1-q)(\sigma - s_0(\phi,\theta)) & \text{if } \theta \in [0,1) \\ w_2(t) = E_\delta\left[e^{-(\beta+\gamma)\delta}(\sigma - s_\lambda(\phi,\theta))\right] & \text{if } \theta = 1, \end{cases} \tag{37}$$

the spreads in (34) as functions of (ϕ,θ), and the equilibrium strategy θ given by (35). Note that the mixed strategy $\theta^* \in (0,1)$ makes it indifferent for the HFT to take $A = 1$ and $A = 2$, leading to the first line in (37). Furthermore, when $\theta = 1$, the economy converges to the benchmark case since the effect of the speed bump in the slow market encompasses the fast market too. Thus, $s_0 = s_\lambda$, and I obtain w_2 in (37).

A.3.6 Short expected delay

As (35) has established, a sufficiently short expected delay, such that $\lambda < (\beta+\gamma)^{-1}Q^{-1}$, does not hamper the incentive of the HFT to take $A = 2$. She always takes "wait-and-grasp-all" strategy, resulting in $\theta = 1$. This makes the economy, as well as the equilibrium results, same as the benchmark case in Section 2. Therefore,

Figure 15: $W(\phi)$ with different λ



a longer expected speed bump increases the speed level ϕ^*, which completely offsets the reduction in the adverse selection cost due to the longer safe interval (Proposition 3). This region is depicted by the left region of the shaded area in Figure 16.

A.3.7 Long expected delay

When a delay is sufficiently long, the HFT becomes reluctant to take $A = 2$ because of the higher execution risk. She starts adopting the mixed strategy ($\theta = \theta^*$) or immediately snipes in the fast market ($\theta = 0$). The switch between these two cases occurs at

$$\hat{\phi} \equiv \beta \left[\lambda(\beta + \gamma)Q - 1\right]. \tag{38}$$

When $\phi < \hat{\phi}$ (resp. $\phi > \hat{\phi}$), the strategy of the HFT is $\theta = 0$ (resp. $\theta = \theta^*$). As already discussed, this threshold is increasing in λ and decreasing in q since both of them reduce the expected profit from sniping in the slow market.

Lemma 2. *When $\theta = 0$, the optimal speed level is given by $\phi_0^* = \sqrt{\beta(\beta+\gamma)}$. The speed and the spread are independent of the expected length of the speed bump λ.*

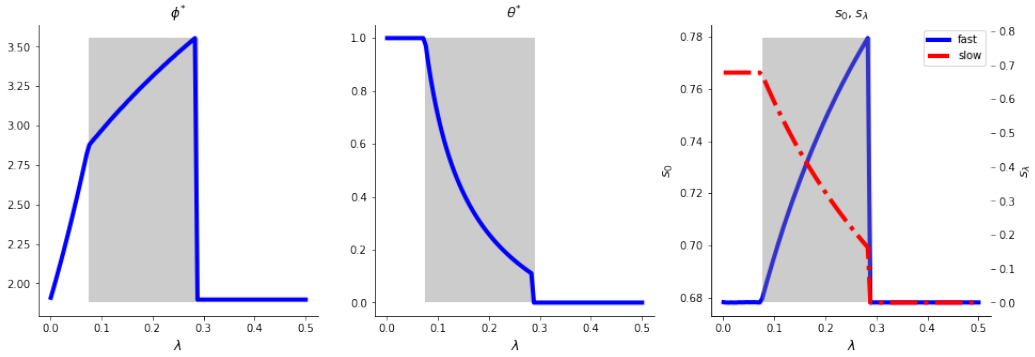
Proof. Plugging $\theta = 0$ into (37) and taking derivative immediately derive the result. □

Since the objective function W switches at $\hat{\phi}$, there exist a couple of candidates for ϕ^* depending on $\hat{\phi} \gtrless \phi_0^*$, and this is crucially affected by the values of λ and Q.

Figure 15 plots the objective function W against ϕ with various parameter values for λ, in which the effect of ϕ on θ is taken into account. This function is not smooth due to the switch at $\phi = \hat{\phi}$. When λ is relatively small, I have $\theta = \theta^* \in (0, 1)$, and the optimal ϕ is higher than $\hat{\phi}$. Then the speed is positively affected by λ, i.e., the longer the expected delay, the faster the HFT. As shown by Corollary 5, this pushes θ^* up. However, because the longer expected delay escalates the execution risk and the expected return starts waining, $W(\theta^*)$ dips below $W(\theta = 0)$ at some λ. Thus, there is a λ that makes $\theta = \theta^*$ and $\theta = 0$ indifferent.

As a result, the optimal speed plummets as λ increases when θ switches from $\theta = \theta^*$ to $\theta = 0$. Intuitively, the optimal speed must incorporate the execution risk by the speed bump only if the HFT snipes in the slow market with a strictly positive probability. Otherwise (if $\theta = 0$), the speed bump has nothing to do with the HFT's expected profit. The speed decision cares only about the endogenous cost at the fast market, which is more sensitive to the change in ϕ compared to the endogenous cost that stems from the slow market. As a result, the optimal speed with $\theta = \theta^*$ is too fast if $\theta = 0$ is the optimal strategy, leading to a dive of ϕ^* at the switch. See Figure 16 for the visual illustration of the effect of λ.

Figure 16: Effect of λ



A.3.8 Adverse selection

Figure 16 shows the level of the optimal speed ϕ^*, the mixed strategy θ^*, and the spreads in both markets as functions of λ. First, if the delay is sufficiently small, so that $\lambda Q(\beta+\gamma) < 1$, the execution risk for the HFT is sufficiently low, and she takes $A = 2$ for sure ($\theta^* = 1$). The result is the same as Section 2, and the optimal speed is increasing in λ, while the adverse selection cost, measured by the half spread, is constant. Note that there is no difference between the slow and the fast markets.

Second, if $\lambda Q(\beta+\gamma) > 1$ but λ is intermediate, the HFT finds it not attractive to take $A = 2$ with 100% probability because of the relatively high execution risk. Thus, she starts to mix $A = 2$ with $A = 1$, so that she stochastically snipes at the timing of information revelation. This is represented by the shaded area in Figure 16. In this case, if she keeps ϕ constant, the welfare declines as λ increases. However, a longer expected delay makes the endogenous cost (i.e., the spread) insensitive to an increase in the speed because market makers set s as if the HFT arrives with a lower probability. This promotes the investment in the speed. In contrast to the speed, the probability of taking $A = 2$ declines, although a higher ϕ^* has a positive impact on θ^*. This is due to the dominating negative effect of λ on θ^*: a longer expected delay makes $A = 2$ a less attractive choice for the HFT.

Finally, if λ becomes sufficiently large, as in the right region of the shaded area, the HFT no longer figures that taking $A = 2$ pays out because the execution risk becomes sufficiently high. Thus, she switches to taking $A = 1$ with 100% probability, making the optimal level of the speed insensitive to λ. The optimal level ϕ^* jumps down from the middle region of λ as mentioned in the previous subsection.

Regarding the adverse selection cost in both markets, the first region with a small λ provides the constant and same level of s. This is natural because both markets face the same risk of the HFT arrival. The intermediate region of λ makes them behave differently. A higher λ directly mitigates adverse selection for market makers, while it increases the optimal speed and the probability of the HFT-confrontation. The fast market bears the risk of the HFT no matter what strategy the HFT takes, though $A = 2$ is discounted by the delay λ. On the other hand, the slow markets are exposed to the HFT only if she takes $A = 2$, and this is protected by λ. As shown by Corollary 4, this asymmetry makes s_λ decreasing and s_0 increasing—a longer speed bump protects the slow markets at the expense of the traditional fast markets.

Once the delay becomes sufficiently long (right side of the shaded area), the risk of HFT completely diminishes in the slow market because the HFT takes $A = 1$ for sure. Hence $s_\lambda = 0$. In the fast market, the spread drops as well because the speed of the HFT is humbled. The fast market still bears the risk of the HFT and keeps the spread strictly positive.

B Appendix: Proofs

B.1 Proof of Proposition 2

Let $\eta \equiv \beta + \gamma$. The explicit formula for W is given by

$$W(\phi) = \frac{1}{1+\lambda\eta}\frac{\phi}{\psi}\frac{\beta(1+\lambda\psi)}{\phi+\beta(1+\lambda\psi)},$$

where

$$p \equiv E_\delta\left[\int_0^\infty \pi_t(\phi,\delta)dt\right] = \frac{1}{1+\lambda\eta}\frac{\phi}{\psi},$$
$$\sigma - s = \frac{\beta(1+\lambda\psi)}{\phi+\beta(1+\lambda\psi)}.$$

Therefore,

$$\begin{aligned} W'(\phi) &= p'(\sigma-s) + p(\sigma-s)' \\ &= p'(\sigma-s)(1-\varepsilon) \end{aligned}$$

with

$$\varepsilon \equiv -\frac{p}{p'}\frac{(\sigma-s)'}{\sigma-s} = \frac{(1+\lambda\eta)\psi}{\eta(1+\lambda\psi)}\frac{\phi}{\phi+\beta(1+\lambda\psi)}.$$

It is obvious that $d\varepsilon/d\phi > 0$. This implies that the optimization problem satisfies the SOC.

The solution is derived by solving the FOC, which is reduced to

$$1 = \varepsilon(\phi).$$

Note that $\varepsilon(0) = 0, \varepsilon'(\phi) > 0$, and

$$\lim_{\phi\to\infty}\varepsilon(\phi) = \frac{1+\lambda\eta}{\eta\lambda(1+\beta)}.$$

Thus, as long as $\lim_{\phi\to\infty}\varepsilon(\phi) > 1$, there is a unique solution. It can be easily checked that this condition is expressed as (7). If this is not satisfied, I have $\phi^* = \infty$.

When (7) holds, the $\phi^* > 0$ solves $1 = \varepsilon(\phi)$, and some tedious calculations show that the solution is given by (11). The second statement is obvious from (11).

B.2 Proof of Proposition 3

By taking a derivative,

$$\frac{ds}{d\lambda} \sim -\phi\psi + \frac{d\phi}{d\lambda}(1+\lambda\eta). \tag{39}$$

Moreover, by the implicit function theorem,

$$\frac{d\phi}{d\lambda} = \frac{\eta g^2 - \phi^2\gamma}{\eta g^2 + \beta\psi^2(1+\eta\lambda)^2} \tag{40}$$

where $g \equiv \phi + \beta(1+\lambda\psi)$. By substituting (40) for the one in (39),

$$\begin{aligned} \frac{ds}{d\lambda} &\sim \psi\beta(1+\lambda\eta)(1+\lambda\psi) - \phi g \\ &= \eta\beta(1+\lambda\psi)^2 - \phi^2. \end{aligned}$$

Therefore, at the optimal speed $\phi^* = \frac{\sqrt{\eta}(1+\lambda\eta)}{1-\lambda\sqrt{\beta\eta}}$, it becomes $ds/d\lambda = 0$.

B.3 Proof of Lemma 1 and Proposition 5

Let $r \equiv \sqrt{\beta + \phi_j}$. The second order derivative of BR_i is

$$\frac{d^2 BR_i(\phi_j)}{d\phi_j^2} = \frac{dr}{d\phi_j}\frac{R}{2r^2}\left(r\frac{R'}{R} - 1\right),$$

with

$$R \equiv \frac{1 + \lambda r^2}{(1 - \lambda\sqrt{\beta} r)^2} + \frac{2\lambda r}{1 - \lambda\sqrt{\beta} r}.$$

Then, $r\frac{R'}{R} > 1$ is identical to $Z(r) < 0$ with

$$Z(r) \equiv 2\beta\lambda^3 r^3 - \lambda(1 + \lambda\sqrt{\beta})r^2 - \lambda\sqrt{\beta}(3 + 2\lambda)r + 1.$$

Note that I am focusing on the bounded solution, that is $1 > \lambda\sqrt{\beta} r$. Since $Z(\frac{1}{\lambda\sqrt{\beta}}) < 0$ and $Z(0) > 0$, there is a unique r^* such that $r > r^* \Leftrightarrow Z(r) < 0$. Then, ϕ_0 is defined as the solution of $r = r^*$, and we obtain the result.

The symmetric equilibrium is given by solving $\phi = BR(\phi)$, which is rewritten as $X(r, \lambda) = 0$ with $r \equiv \sqrt{\beta + \phi}$ and

$$X(r, \lambda) = \lambda(1 + \sqrt{\beta})r^3 - r^2 + (1 - \lambda\beta\sqrt{\beta})r + \beta.$$

This function has the following properties:

$$\frac{\partial X(r, \lambda)}{\partial \lambda} > 0, \ \forall r > 0,$$
$$X(r, 0) = -r^2 + r + \beta, \lim_{\lambda \to \infty} X(r, \lambda) = \infty.$$

Therefore, as λ increases, X shifts up from $X(r, 0)$ and eventually explodes for all r. At $\lambda = 0$, $X = 0$ has a unique solution in the positive r region. By the continuity of X regarding λ, if $\lambda \searrow 0$, then $X = 0$ attains three solutions, two in the positive region (a larger one can be greater than $\frac{1}{\lambda\sqrt{\beta}}$). Let r^+ and r_- be these two solutions. Since $\frac{\partial X(r^+, \lambda)}{\partial r} > 0$ and $\frac{\partial X(r_-, \lambda)}{\partial r} < 0$, the implicit function theorem implies $\frac{dr^+}{d\lambda} < 0$ and $\frac{dr_-}{d\lambda} > 0$, which means that the stable solution is increasing in λ. By the monotonicity of X regarding λ, there is a unique $\lambda = \lambda_0$ such that $r_-(\lambda_0) = r^+(\lambda_0)$, and $X(r, \lambda) > 0$ for all r if $\lambda > \lambda_0$, i.e., there are no solutions.

B.4 Proof of Proposition 6

First, by letting $\psi \equiv \sum_i \phi_i + \beta$ and $\eta \equiv \phi_j + \beta$, the FOC for HFT i can be expressed as

$$1 = \frac{\phi_i}{\phi_i + \beta(1 + \lambda\psi)}\frac{Y(\psi)}{Y(\eta)},$$

with

$$Y(x) = \frac{x}{1 + \lambda x}. \tag{41}$$

Under the symmetric equilibrium, it reduces to

$$1 = s(\phi, \lambda)\frac{Y(\psi)}{Y(\eta)},$$

with $\psi \equiv 2\phi + \beta, \eta \equiv \phi + \beta$, and ϕ is the equilibrium speed. Then,

$$\frac{ds}{d\lambda} \sim \phi\psi[\psi(1+\lambda\psi) - 2\eta(1+\lambda\eta)] - \lambda\eta\phi\psi(1+\lambda\beta). \tag{42}$$

Since the symmetric equilibrium solves

$$\phi^2 = \beta\eta(1+\lambda\psi)^2,$$

it is that $\phi = \sqrt{\beta\eta}(1+\lambda\psi) \geq \beta \geq 1$. By using these conditions, the RHS of 42 is positive.

B.5 Proof of Proposition 7

I use the same notations as in Appendix B.4. The traditional model satisfies the SOC: By letting $\Gamma \equiv \phi_i + \beta(1+\lambda\psi)$,

$$\begin{aligned}\frac{d}{d\phi_i}\left(\frac{\partial W_i}{\partial \phi_i}\right) &= (\sigma - s)\frac{\partial^2 \pi_i}{\partial \phi_i^2} + \frac{\partial(\sigma - s)}{\partial \phi_i}\frac{\partial \pi_i}{\partial \phi_i} \\ &\sim -\Gamma\beta(1+\lambda\psi) - \psi(\Gamma - \phi_i(1+\beta\lambda)) \\ &< 0.\end{aligned}$$

Then, by using Y in (41), the FOC is rewritten as

$$C'(\phi_i) = \frac{\beta}{\psi\Gamma}\frac{Y(\eta)}{Y(\psi)} \equiv K(\phi_i, \phi_j, \lambda), \tag{43}$$

with $\psi \equiv \sum_i \phi_i + \beta$ and $\eta \equiv \phi_j + \beta$. By taking the partial derivative of K with respect to ϕ_j around the symmetric stable equilibrium, I have

$$\frac{\partial K}{\partial \phi_j} \sim -\beta - \lambda\psi\left[\eta\frac{\phi + 2\beta(1+\lambda\psi)}{\phi + \beta(1+\lambda\psi)} - \frac{\phi}{1+\lambda\eta}\right] < 0.$$

By the same token, K is decreasing in λ around the symmetric equilibrium. Together with the SOC, the implicit function theorem implies that $\frac{dBR_i}{d\phi_j} < 0$ and $\frac{d\phi}{d\lambda} < 0$. As the form of the equilibrium spread is identical to the strategic model, the opposite effect of $\frac{d\phi}{d\lambda}$ in Proposition 6 shows that $\frac{ds}{d\lambda} < 0$.

B.6 Proof of Proposition 10

The comparison is

$$w_1 \gtrless w_2 \Leftrightarrow \lambda\eta Q(\sigma - s_0) \gtrless \sigma - s_\lambda.$$

By plugging the formulae for the equilibrium spreads into the inequality above,

$$\begin{aligned}L(\theta) &\equiv \lambda\eta Q(\sigma - s_0) = \lambda\eta Q\frac{K(\theta)}{\phi + \beta K(\theta)}, \\ R(\theta) &\equiv \sigma - s_\lambda = \frac{J(\theta)}{\phi + \beta J(\theta)},\end{aligned}$$

with

$$K(\theta) = 1 + \frac{\psi}{\eta}\theta\frac{\lambda\eta}{1 + (1-\theta)\lambda\eta}, J(\theta) = 1 + \lambda\eta\left(1 - \theta + \theta\frac{\psi}{\eta}\right).$$

Table 2: Design of Speed Bumps

Exchange	Date	Targets of delay	Length of delay
IEX	Octber 2013	All (except the NBBO)	350 microseconds
NYSE American	July 2017	All	350 microseconds
Thomson Reuters	June 2016	All but cancel	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
Chicago Stock Exchange	Proposed	Liquidity takers	350 microseconds
NASDAQ OMX PHLX	Proposed	Liquidity takers	5 microseconds
ICE Futures	Proposed	Liquidity takers	3 microseconds
Interactive Brokers	Proposed	Liquidity takers	10-200 milliseconds

Note: This table reproduces Table 3 in Baldauf and Mollner (2017) and Table 1 in Khapko and Zoican (2019).

These functions have the following properties:

$$\frac{dL}{d\theta} > 0, L(0) = \frac{\lambda\eta Q}{\phi+\beta}, L(1) = \frac{\lambda\eta Q(1+\lambda\psi)}{\phi+\beta(1+\lambda\psi)},$$
$$\frac{dR}{d\theta} < 0, R(0) = \beta^{-1}, R(1) = \frac{1+\lambda\psi}{\phi+\beta(1+\lambda\psi)}.$$

Thus, if $\lambda\eta Q < 1$, then $L(1) < R(1)$, indicating that $R > L$ for all $\theta \in [0,1]$. Therefore, $\theta^* = 1$ is the optimal. When $\lambda\eta Q \geq 1$, the result depends on $L(0) \gtrless R(0)$. If $\lambda\eta Q < (\beta+\phi)/\beta$, then $R(0) > L(0)$, which implies that there is a unique interior solution θ^* that solves the indifference condition. The solution solves $L(\theta) = R(\theta)$, and tedious calculation gives (36). Finally, if $\lambda\eta Q > (\beta+\phi)/\beta$, I have $R < L$ for all $\theta \in [0,1]$. Thus, $\theta = 1$ is the optimal strategy.

C Speed bump implementations

The design of the speed bumps varies across exchange platforms. It is firstly categorized into *symmetric* and *asymmetric* speed bumps. The symmetric speed bumps impose an intentional delay on all orders, while the asymmetric speed bumps delay only a certain type of orders, such as liquidity taking orders. Moreover, the length of a speed bump can be randomized. The purpose of randomization is to refrain traders from calculating and forecasting the pattern of delays, and some exchange platform has proposed randomization by AI. Table 2 reproduces the list of speed bumps in Baldauf and Mollner (2017) and Khapko and Zoican (2019). It shows that asymmetric speed bumps tend to be more popular than symmetric ones, thus motivating the analyses in my model.