Strategic Speed Choice by High-Frequency Traders under Speed Bumps[*]

Jun Aoyagi[†]

March 10, 2019

Abstract

We study how high-frequency traders (HFTs) strategically decide their speed level in a market with a random speed bump. If HFTs recognize the market impact of their speed decision, they perceive a wider bid-ask spread as an endogenous upward-sloping cost of being faster. We find that the speed elasticity of the bid-ask spread (slope of the endogenous cost function) negatively depends on the expected length of a speed bump since a longer delay makes market makers insensitive to HFTs' speed increment. Hence, speed bumps promote the investment of HFTs in high-speed technology by reducing the marginal cost of getting faster, undermining their intended purpose of protecting market makers. Depending on the expected length of a bump and exogenous cost of speed, an arms race among HFTs exhibits both complementarity and substitution. These findings explain the ambiguous empirical results regarding speed bumps and adverse selection for market makers.

Keywords: High-frequency trading, market structure, speed bumps, adverse selection, strategic speed decision.

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

[*] First draft: June 2018.

[†] Department of Economics, University of California at Berkeley. E-mail jun.aoyagi@berkeley.edu.
I appreciate the constructive comments from Michael Brolley, Nicolae Garleanu, Terry Hendershott, Ryo Horii, Ulrike Malmendier, Hayden Melton, Emi Nakamura, Christine Parlour, Yoko Shibuya, David Sraer, Yingge Yan, Haoxiang Zhu, and seminar participants at Osaka University (ISER), 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.

1 Introduction

> "Never before in human history have people gone to so much trouble and spent so much money to gain so little speed."—*Flash Boys: A Wall Street Revolt* by Lewis (2014).

The ever-increasing speed of electronic financial markets pushes traders to be lightning fast. They are obsessed with being the first to acquire information for trading purposes, spending significant amounts of money on high-speed technologies, such as custom-built fiber-optic cables and microwave/millimeter-wave transmissions. With these 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. It is argued that the informational advantage that HFTs obtain through increasing speed exposes market makers to the cost of adverse selection in the sense of Glosten and Milgrom (1985). That is, HFTs trade with market makers only if they receive news that is not yet publicly available and find market makers' orders outdated and mispriced.[1] By exploiting their speed advantage, HFTs "snipe" stale quotes provided by market makers (Budish et al., 2015).

The speed race by HFTs has prompted some exchange platforms to slow down HFT-involved transactions by introducing *speed bumps*. A speed bump imposes a 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 Exchange and TMX Group, both Canadian exchanges, also apply a few milliseconds of random delay to non-cancellation orders.[2] Specifically, the speed bump in the latter markets aims to slow down only HFT-involved orders by classifying traders into high-frequency (latency-sensitive) and non-high-frequency categories.[3]

[1] One of the most frequently cited market benefits is liquidity provision by high-frequency market makers. However, extremum events, such as the May 2010 "flash crash," make regulators increasingly concerned that the liquidity provided by HFTs is likely to evaporate when it is most needed. See, Conrad et al. (2015) for the empirical study of this liquidity evaporation.

[2] See Appendix F of Baldauf and Mollner (2017) which provides a comprehensive summary of institutional details.

[3] Depending on the institutional details, the types of traders (or orders) to be protected may change. For

Table I: 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.

This paper analyzes the effect of speed bumps on speed decision of HFTs and on the adverse selection cost for market makers by focusing on a non-cancellation delay aimed at hampering sniping behavior of HFTs.[4] The key result is that a speed bump can *increase* the speed of HFTs and *worsen* adverse selection for market makers in contrast to its intended purpose. Specifically, once we allow HFTs to strategically choose their speed level (i.e, they are aware of the reaction of price setters), a speed bump increases the marginal benefit of being faster.

The strategic motive in the speed decision arises because major high-speed financial institutions have significant shares in the trading volume in markets.[5] For example, Table I 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. Typically, HFTs benefit from a huge number of small, short-lived transactions, and each trading decision does not impact the equilibrium price. However, when an institution decides a speed technology, she becomes aware of the market impact of her choice because a sizable number of transactions involve the same speed technology and affect the equilibrium price.

instance, the speed bump in the IEX is more likely to protect pegged orders from non-HFTs, but not market makers on a lit LOB, from being sniped by HFTs. The non-cancellation delay and the HFT-specific delay adopted by the two Canadian exchanges are more likely to save market makers from adverse selection cost.

[4] The term "speed" includes the choice of the geographical location of the firm's information server. For example, a spot in the mid-Atlantic ocean is the optimal point to exploit the price difference between the NYSExchange and the London Stock Exchange.

[5] 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).

As the literature points out, the faster the HFTs, the more severe adverse selection the market makers face. Thus, a higher speed puts positive pressure on the bid-ask spread and, in turn, reduces the sniping profit of HFTs. Therefore, the spread works as an *endogenous* upward-sloping cost of being faster. Importantly, it provides a channel through which a speed bump affects the speed decision of HFTs, because market makers' adverse selection risk, as well as the equilibrium spread, is affected by a speed bump. This channel is novel since an exogenous sunk cost of speed analyzed in the literature is independent of a bump.

First, we consider a simple benchmark structure to separate the key mechanism: homogeneous markets with a single HFT, having a random speed bump of a δ-period with $\lambda = E[\delta]$. If λ increases, market makers know that they are less likely to be picked off by the HFT. As a result, they do not care much about a marginal increase in the HFT's speed, and their pricing behavior, i.e., the bid-ask spread, becomes less responsive. This induces a lower endogenous marginal cost of speed investment for the HFT, providing her with a stronger incentive to be faster.

As an extension, we consider a speed competition among *multiple* HFTs—an "arms race"—and allow them to serve not only as snipers but also as high-frequency market makers. In the literature, such as Foucault et al. (2003), traders' speed levels interact with each other because one trader's speed affects other traders' probability of successful snipe, leading to the strategic substitution. In contrast, our endogenous cost of speed (bid-ask spread) provides a new channel for the interaction because the spread is an equilibrium variable. Specifically, depending on the relative significance of the exogenous and endogenous costs of speed, the arms race can exhibit both strategic complementarity and substitution.

If the exogenous sunk cost of speed is relatively small, an arms race creates strategic *complementarity*, because the speed-up by an HFT as a market maker reduces the sensitivity of her spread to other HFTs' speed-up. Also, a faster market maker decreases the sniping probability of snipers, making them care less about an adverse price movement caused by their speed increase. As a result, a faster market maker reduces snipers marginal cost of being faster and enhances their investment in speed.

In this situation, the introduction of a speed bump can backfire because a higher λ makes each HFT willing to be faster, triggering a fiercer speed competition and positive externality due to the complementarity. 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.

Therefore, our strategic model with the endogenous cost of speed proposes opposite results to a traditional model with an exogenous sunk cost. We can think of these as two extreme cases: By introducing exogenous cost in our model and adjusting the exogenous cost parameter, for example, our model navigates between these extremes, leading to rich equilibrium behavior.

While our model is theoretical, we propose some testable implications and policy discussions. For example, our model implies that the SEC's policy in 2017 that approved the IEX (with a bump) as a National Securities Exchange can strengthen HFTs' demand for speed technologies, thereby allowing exchange platforms to charge higher fee for the direct data feed and colocation service. Our model indicates that this effect undermines the recent attempt of the SEC to block the exchange platforms from increasing the price for their data access.[6]

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 (2018) describes the speed decision of HFTs as a model of irreversible investment with an optimal stopping time, while Bongaerts and Van Achter (2016) view it from a perspective of high-frequency market making.[7] However, the speed decision in these models is discrete (i.e., being fast or not), and they abstract away from addressing the implications of the equilibrium level of speed. Based on Foucault et al. (2003), Liu (2009) and Foucault et al. (2016) investigate a continuous choice of speed based on the monitoring intensity of traders.[8] However, traders decide on the speed level simultaneously with other

[6] 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.

[7] 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.

[8] Foucault et al. (2013) consider the optimal choice of the monitoring intensity by high-frequency snipers and market makers. It involves the exogenous cost but is not strategic. Both snipers and market makers

types of players (e.g., market makers), which requires them to focus on the exogenous cost of speed investment. Our model differs from theirs since the speed decision is continuous and bears an endogenous cost due to the strategic motive of HFTs. Our results are unique since these two modifications empower us to analyze how speed choice is affected by speed bumps.

As traders get faster, questions arise regarding the speed and frequency of executions by a trading platform. By altering the trading frequency of the Kyle-type model, Du and Zhu (2017) show that a low-frequency platform works better to reallocate assets, though it limits the ability to react to new information promptly. Pagnotta and Philippon (2018) also consider platforms' decisions regarding execution frequency and fees to attract speed-sensitive traders. Menkveld and Zoican (2017) also explore the effect of latency on HFTs' strategy and spread, citing risk aversion as a key to generating the result.[9] In their analyses, which pays little attention to the speed choice of HFTs, the frequency of transactions is determined at a market level and applies to all investors.

Our 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. 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 our model.[10]

The scope of the literature extends to other empirical findings regarding the HFT and the effect of bumps.[11] Hu (2018) analyzes the SEC approval of the IEX as a national securities

obtain a positive profit from trading due to heterogeneous private values of an asset, generating strategic complementarity in an arms race.

[9] Menkveld and Zoican (2017) obtain a hump-shaped equilibrium spread against a delay. This stems from the switch from the pure-strategy to mixed-strategy equilibrium, and it depends on the risk aversion parameter.

[10] Moreover, these models do not study the coexistence of slow and fast markets, which is analyzed in Appendix A.2. In independent work, Brolley and Cimon (2017) explore this coexistence and find a result consistent with ours, although it stems from a completely different mechanism.

[11] Hendershott and Moulton (2011) analyze the impact of the hybrid market at the NYSE and show that the faster market structure increases quoted and effective spreads and adverse selection cost. Riordan and Storkenmaier (2012) focus on the system upgrade in the Deutsche Boerse, Frino et al. (2014), Boehmer et al.

exchange, making traders route their orders under the "Order Protection Rule," and finds a net improvement in market quality measured by the spreads. Shkilko and Sokolov (2016) exploit interruptions of messaging via microwave communication caused by precipitation (i.e., rain or snow) to find a reduction in quoted spreads.[12] Chen et al. (2017) investigate the effect of a bump in the TMX Alpha, reporting an increase in quoted spreads. In our model, we can reconcile these results because, depending on the relative significance of the exogenous cost and the level of expected delay, speed bumps will affect a 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, in which a short-lived HFT tries to snipe stale limit orders. The asset has a stochastic liquidation value $v = \pm\sigma$ with equal probability. v is publicly announced at a stochastic time T, which occurs as a Poisson arrival with intensity γ. With the public announcement, the asset is liquidated. It is traded during $t < T$ due to liquidity needs or the arrival of private information, as in Glosten and Milgrom (1985) and Budish et al. (2015). Following the convention of market microstructure, we assume that each trader can hold only a unit position.

2.1 Traders

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

(2015), and Brogaard et al. (2015) study the colocation as an example of latency reduction, and Hasbrouck and Saar (2013) construct a measure of low-latency in the NASDAQ to find subsequent shrinkages in spreads. On the other hand, Ye et al. (2013) analyze the importance of the tick-size constraint and report that latency declines at the NASDAQ did not significantly alter spreads (except for the smallest stocks).

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

market once they exit. Cancellation is immediate and incurs no cost.

There is one ($N = 1$) risk-neutral high-frequency trader (HFT). Before $t = 0$, the HFT invests in a technology that provides the speed ϕ.[13] Equipped with the speed device with ϕ, she can observe private news regarding v and react to it with a Poisson probability with intensity ϕ. We denote T_H as the arrival time of this Poisson news. Upon the arrival of the news, the HFT immediately submits market orders to "snipe" stale limit orders provided by the continuum of market makers.[14,15]

In addition, there is a continuum of liquidity traders who are exposed to a liquidity shock. The shock exogenously makes them submit buy or sell market orders with equal probability. We can think of them as noise traders, and trading against them conveys no information to market makers. Let T_L be the timing of the Poisson shock which arrives with intensity $\beta (\geq \gamma)$.

Finally, 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.

2.2 Market Structure

A continuous market imposes a random speed bump on incoming orders from the HFT. Specifically, an order submitted to the market at date t arrives at $t + \delta$, where δ is a random delay. Orders from liquidity traders and cancellation requests from market makers are executed promptly.[16] Thus, during $\tau \in (t, t + \delta)$, outstanding limit orders can be illusory

[13] In the benchmark model with $N = 1$, imposing a sunk cost on the speed investment does not change our result. In the extension with $N \geq 2$, we need a positive and convex cost to hamper the strategic complementarity and to derive an equilibrium.

[14] If we give an index $i \in [0, 1]$ to each market maker, the HFT submits marketable limit orders to obtain sniping profit $\sigma - s_i$ from each i. Since all the market makers quote a competitive homogeneous spread, the HFT's aggregate gain is $\int_0^1 (\sigma - s) di = \sigma - s$.

[15] We can show that the 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.

[16] For simplicity, we assume that there are no other sources for a latency, while the primitive parameters, such as β, can be seen as the potential latency that characterizes the speed of each type of trader.

for the HFT if liquidity traders trade against them or if market makers cancel due to public news.

For notational simplicity, we assume that δ follows an exponential distribution with a parameter b, and the expected length of a delay is denoted by $\lambda \equiv E[\delta] = b^{-1}$.[17]

Alternative Market Structures

As an extension, we analyze a situation with multiple HFTs, $N \geq 2$, in Section 3, in which each HFT serves not only as a sniper but also as a high-frequency market maker. This setting sheds light on a strategic property of an "arms race." Appendix A.1 considers a case where market makers can continuously update (cancel and resubmit) their limit orders. The coexistence of slow and fast markets is analyzed in Appendix A.2. This shows that a bump triggers a shift of adverse selection from slow markets with a bump to fast markets with no bumps, consistent with the empirical result by Chen et al. (2017).

2.3 Equilibrium

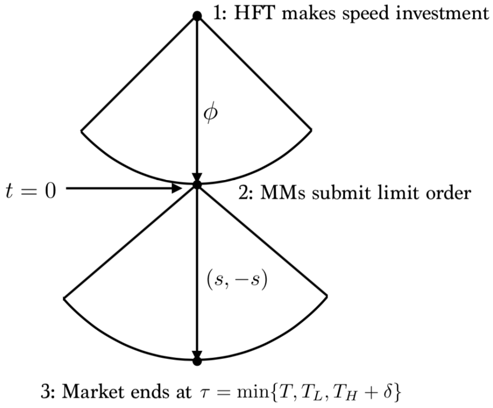
We conceptualize our model as a sequential game with two stages, as depicted in Figure I. In the first stage, the HFT decides the level of ϕ.[18] Given this, each market maker submits a competitive limit order, anticipating a confrontation with the informed HFT and liquidity traders. In the trading stage, the HFT looks for an opportunity to snipe.

The equilibrium concept is a 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. In contrast to Kyle (1985), however, the trading stage in $t \in (0, T)$ is competitive, and the HFT behaves as if her trading strategy does not have a price impact. This is because she splits her orders and sends them to an infinitely large number of market makers given the outstanding limit orders. This follows the literature

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

[18] Of course, the setting of the speed decision used in our model does not comprehensively reflect real-world conditions. As Dugast et al. (2014) suggest, some components of speed choice may occur simultaneously with market makers' behavior. However, we believe that the *ex-ante* strategic speed decision is still significant because HFTs would not invest in speed *ex-ante* if they did not exploit it in an *ex-post* trading game.

Figure I: Timeline


1: HFT makes speed investment
ϕ
$t = 0$
2: MMs submit limit order
$(s, -s)$
3: Market ends at $\tau = \min\{T, T_L, T_H + \delta\}$

and captures the real-world behavior of HTFs, who send and cancel a massive number of small orders within a very short time frame.

2.3.1 Optimal Behavior of Market Makers

In a perfect competition, a limit order sent by a market maker yields zero expected profit, as in Glosten and Milgrom (1985). Without a loss of generality, let us consider how an ask price s is determined when $v = \sigma$.[19]

Given that a market order arrives at date t, it is possible that the taker is information or liquidity driven. As a result, the spread is set so that $s = E[v|\text{buy order at } t]$, where the expectation is over δ and the timing of the trade. The key effect of a bump is to reduce the probability of being picked off by the HFT or, put differently, to increase the probability that market makers observe public news to cancel their limit orders.

Suppose that a trade takes place at date t. If $t < \delta$, there is no fear of facing an information-driven HFT because of the speed bump. Put differently, the fastest possible arrival of the HFT occurs at δ. During this "safe interval," liquidity traders arrive before the public news with a density $\beta e^{-(\beta+\gamma)t}$. Otherwise, market makers can cancel their orders with density $\gamma e^{-(\beta+\gamma)t}$ at period t.

If a trade occurs at $t \geq \delta$, on the other hand, it bears an adverse selection cost: the HFT buys an asset only if the limit order is mispriced given the true information. The HFT gets to trade 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 $s - \sigma$. Market makers can also

[19] Results for the opposite case can be given by a symmetric argument.

trade with liquidity traders if there is a liquidity shock at t, and the HFT becomes informed after $t - \delta$. In this case, the trading profit is $s - E[v] = s$. Since δ is stochastic, the expected return for a market maker is

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

where the expectation relates to δ, and $\psi \equiv \phi + \beta + \gamma$. The first integral in (1) shows the trading profit in $t < \delta$, while the second describes the case with $t \geq \delta$.

This formulation is the result of the following probabilities: given δ,

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

which lead to the second term in (1). It is then possible to get the equilibrium spread from the break-even condition:

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

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

A few remarks on s are in order. First, 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 mitigates adverse selection risk by generating a safe interval.

If ϕ is fixed, a lower λ induces a higher spread since the expected delay becomes shorter. Also, an infinitely small expected delay ($\lambda \to 0$) makes s converge to the traditional equilibrium spread of Glosten and Milgrom (1985). Therefore, as Budish et al. (2015) and Baldauf and Mollner (2017) point out, the direct effect of a bump mitigates the adverse selection cost for market makers.

This argument is built on the assumption that ϕ is fixed, i.e., the HFT's speed decision is not influenced by the bump. When the speed choice by the HFT is considered, the speed bump affects s via the fluctuation of the optimal speed as well. The existing models argue that the incentive to be faster diminishes as the bump gets longer, i.e., a higher λ reduces

s not only by the direct effect but also by making ϕ lower, while our model proposes the opposite effect.

The following properties of the spread are useful to understand the mechanism. First, note that the price impact of the speed is positive:

$$\frac{\partial s}{\partial \phi} > 0.$$

We call this derivative the "sensitivity" of a spread (price) to a speed-up by the HFT. It turns out that this represents the slope of the endogenous cost of being faster. At the same time, we have the following:

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

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

Therefore, the longer the expected delay, the less sensitive the spread becomes. A market with a higher λ is protected by a longer (expected) safe interval, and market makers behave as if the share (arrival rate) of the HFT is small. Hence, market makers care *less* about the speed investment by the HFT, making their pricing behavior less sensitive to ϕ.

2.3.2 Profit of the HFT

When the HFT becomes informed and submits market orders at t, they will be executed at $t + \delta$ if (i) there is no liquidity shock during $(t, t + \delta)$ and (ii) no public news arrives in the same interval. This happens with

$$\pi_t(\phi, \delta) \equiv \Pr(T_H = t, \min\{T, T_L\} > t + \delta) = \phi e^{-\psi t} e^{-(\beta+\gamma)\delta}. \tag{3}$$

Thus, if the random delay is δ, the profit from sniping at t is $\pi_t(\phi, \delta)(\sigma - s)$.

The first coefficient in (3) represents the probability that the HFT obtains the information at t. The sniping probability involves an additional exponential coefficient, $e^{-(\beta+\gamma)\delta}$, which shows the disadvantage of the HFT that stems from a speed bump, i.e., front-running by liquidity traders or cancellation by market makers due to the δ-delay. Therefore, a longer delay directly reduces the expected profit of the HFT in the second stage.

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

$$W(\phi) \equiv E_\delta \left[\int_0^\infty \pi_t(\phi, \delta)(\sigma - s)dt \right] = \frac{\phi}{\psi} \frac{1}{1 + (\beta + \gamma)\lambda} (\sigma - s). \tag{4}$$

Note that the HFT always submits unit orders since she exits the market once her orders are executed, i.e., she is a short-term investor.[20]

2.3.3 Optimal Speed

We move on to the speed choice by the HFT in the first stage. To obtain an interior solution, we make the expected length of the delay relatively short:

$$\lambda < \frac{1}{\sqrt{\beta(\beta + \gamma)}}. \tag{5}$$

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

Under (5), the optimization problem of the HFT is

$$\max_\phi W(\phi) \equiv E_\delta \left[\int_0^\infty \pi_t(\phi, \delta)(\sigma - s)dt \right], \tag{6}$$
$$s.t.\ s = \frac{\frac{\phi}{1+\lambda\psi}}{\frac{\phi}{1+\lambda\psi} + \beta} \sigma.$$

This 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 market makers and saps her sniping profit. For this reason, we can think of the equilibrium spread as an *endogenous* cost of speed. Importantly, Lemma 1 suggests that the slope of this endogenous cost is affected by λ, and, in turn, affects the marginal cost of being faster for the HFT.

To analyze how λ alters the optimal decision, consider a marginal gain of being faster:

$$\begin{aligned} \frac{dW}{d\phi} &= E_\delta \left[\int_0^\infty \left\{ (\sigma - s(\phi)) \frac{d\pi_t(\phi, \delta)}{d\phi} + \pi_t(\phi, \delta) \frac{d}{d\phi}(\sigma - s(\phi)) \right\} dt \right]. \\ &= (\sigma - s(\phi)) E_\delta \left[\int_0^\infty \frac{d\pi_t(\phi, \delta)}{d\phi} dt \right] (1 - \varepsilon(\phi)), \end{aligned} \tag{7}$$

[20] See Appendix A.1 for a more general setting with continuous updating by market makers and time-dependent s_t.

where

$$\varepsilon \equiv -\frac{d \log(\sigma - s(\phi))}{d \log E_\delta \left[\int_0^\infty \pi_t(\phi, \delta) dt\right]} > 0.$$

ε is the sensitivity of the sniping profit $(\sigma - s)$ to a change in the expected sniping probability of the HFT ($E_\delta \left[\int_0^\infty \pi_t(\phi, \delta) dt\right]$). We call this the elasticity, and Appendix B.1 provides an explicit formula.

Note that obtaining a higher ϕ affects W through two competing channels and exhibits a price-liquidity tradeoff: it increases the sniping probability (the first term in [7]), while reducing the sniping profit via the adverse price movement (the second term in [7]).

When the equilibrium spread is more sensitive to ϕ than the sniping probability (i.e., $\varepsilon > 1$), being faster harms the profit of the HFT, and incentive to increase ϕ dwindles. On the other hand, if the HFT knows that the price impact of her speed choice is small, it is more likely that an improvement in sniping probability ($\frac{d\pi}{d\phi}$) dominates a decline in profit due to the wider spread ($\frac{d(\sigma - s)}{d\phi}$), luring her to be faster. In other words, for the strategic HFT, the marginal cost captured by the sensitivity of the spread matters considerably.

The following results guarantees the concavity of the problem (see Appendix B.1 for proofs):

Lemma 2. *The elasticity is increasing in the speed: $d\varepsilon(\phi)/d\phi > 0$.*

When the HFT is fast, market makers estimate that the economy is inhabited by a relatively large measure of the HFT in terms of the arrival rate. Therefore, a marginal increase in ϕ reduces market makers' expected profit, and they charge a wide spread to compensate for the expected loss. That is, as the HFT becomes faster, market makers grow more concerned about facing the HFT, and their pricing behavior is more sensitive to changes in speed. Thus, as ϕ increases, it is more likely that the steeper endogenous marginal cost of being faster will outweigh the higher marginal benefit from a higher π, making the objective function (6) concave.[21]

As a result, the optimal speed ϕ^* is derived by solving for the FOC.

[21] The condition in (5) is required to make Lemma 2 hold. If λ is sufficiently large, market makers become too insensitive to make $\frac{d\varepsilon}{d\phi} > 0$. Thus, the HFT can be infinitely fast, and we need an exogenous cost to make the problem well defined.

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

$$\phi^* = \frac{\sqrt{\beta+\gamma}(1+\lambda(\beta+\gamma))}{1-\lambda\sqrt{\beta(\beta+\gamma)}}. \tag{8}$$

(ii) ϕ^* is increasing in λ.

Proof. See Appendix B.1. □

In contrast to the traditional models, Proposition 2 demonstrates that, if the speed choice is strategic, a speed bump increases the equilibrium speed of the HFT. This modification of the speed decision is natural given that several high-frequency financial institutions control significant shares, as discussed in Section 1 and shown in Table I.

When the HFT knows how her speed investment affects the pricing behavior of market makers, an equilibrium spread generates an *endogenous* cost of being faster. This not only guarantees a bounded solution even without an exogenous cost of speed (Lemma 2), but also overturns the traditional result regarding speed bumps (Proposition 2).

The key mechanism is the negative impact of a speed bump on the sensitivity of the price in Lemma 1. Since a bump intends to slow down the HFT and protect market makers, the spread becomes insensitive to a change in the speed. That is, an intentional delay endogenously reduces the marginal cost of being faster. Hence, a speed bump does not prevent a speed race but promotes it, as Proposition 2 attests.

This surprising finding highlights the main difference of our results from the literature, such as Budish et al. (2015), Haas and Zoican (2016), and Baldauf and Mollner (2017). As in their models, if we assume that the HFT does not care about the effect of her speed choice on the spread, the second effect in (7) disappears. In this case, some exogenous costs of speed are required to make ϕ^* bounded, and the effect of λ on ϕ^* becomes reversed. We compare our model to the traditional ones in more detail in Subsection 3.4.

2.4 Effect on Adverse Selection

We are interested in how a speed bump affects adverse selection cost for market makers. It is straightforward that there are two competing effects. First, as the literature suggests, a speed bump reduces adverse selection cost because it dampens the probability for market makers of confronting the HFT. However, our strategic model adds an opposing channel:

a speed bump promotes speed investment by the HFT since it endogenously reduces the marginal cost of being faster. In the following, we take the equilibrium half spread s as a measure of adverse selection and investigate its equilibrium behavior.

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.

With ϕ fixed, a speed bump reduces the profit of the HFT since the first-δ periods become safe intervals for market makers and the HFT cannot snipe. To compensate for this disadvantage, the strategic HFT gets faster. Anticipating the price impact of her speed investment, she chooses the level of ϕ that eliminates the cost from the speed bump, thereby muting its effect. As a result, the two competing consequences of a speed bump cancel each other out. Put differently, the speed-up due to a longer delay is an indirect effect of a change in the price sensitivity, $\frac{ds}{d\phi}$. Since the reduction of a spread by λ is a direct effect, the speed-up cannot predominate.

In the following sections, however, we show that this finding regarding adverse selection is specific to the benchmark setting and significantly changes if we consider more general market structures.

3 Multiple HFTs and High-Frequency Market Making

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

Assume that there are two HFTs ($i = 1, 2$), both of whom provide limit orders at $t = 0$ as market makers at a competitive price. At a random date, $T_i \sim \exp(\phi_i)$, HFT i obtains private news about v. When the news arrives, it is optimal for HFT i to immediately send market orders to snipe the stale limit orders of her opponent (HFT j) and to simultaneously cancel her limit orders.

The behavior of liquidity traders is the same as in the benchmark, but we ignore public news at $T \sim \exp(\gamma)$, since it only adds complexity. The other structures of the game stay the same as in the benchmark. Note that the results with $N = 2$ can be easily extended to

$N \geq 3$ with an additional parameter N that measures the (inverse of) market power. For technical reasons, assume that $\beta \geq 1$ and focus on the symmetric equilibrium.

3.1 Optimal Behavior of HFTs

Consider HFT j as a high-frequency market maker (HFMM). Her behavior is the same as that of ordinary market makers in the benchmark model except that she can cancel her limit orders at $T_j \sim \exp(\phi_j)$. Thus, the break-even condition provides the following equilibrium spread:

$$s_j = \frac{\phi_i}{\phi_i + \beta(1 + \lambda(\phi_i + \phi_j + \beta))}\sigma.$$

This spread has the same structure as s in (2): it reflects an expected value of v conditional on the trade. Note that the symmetric equilibrium makes the spreads set by both HFTs the same; $s = s_1 = s_2$.

We turn to the optimal speed decision of HFT i as a sniper. Since the competition drives her total profit from market making to zero, her gains come only from the sniping profit. Thus, the optimization problem is analogous to (4):

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

Since this is exactly the same as the benchmark case if we substitute ϕ_j for γ, the best response function of HFT i is a modified (8):

$$BR_i(\phi_j) = \frac{\sqrt{\beta + \phi_j}[1 + \lambda(\beta + \phi_j)]}{1 - \lambda\sqrt{\beta(\beta + \phi_j)}}, \tag{9}$$

as long as $1 > \lambda\sqrt{\beta(\beta + \phi_j)}$. Otherwise, $\phi_i = \infty$ is the best response. In this section, we focus on bounded responses, while Subsection 3.2 analyzes all possible symmetric equilibria.[22] The following property of the best response function helps explain the mechanism:

Proposition 4. *The best response function exhibits strategic complementarity, i.e.,* $\frac{dBR_i(\phi_j)}{d\phi_j} > 0$.

[22] Technically, we can avoid the unbounded equilibrium if we introduce a positive exogenous sunk cost for the speed.

The intuition behind this proposition should be clear if we analyze the marginal gain of being faster for HFT *i*:

$$w_i \equiv \frac{\partial W_i}{\partial \phi_i} = (\sigma - s_j)\frac{\partial \pi_i}{\partial \phi_i} + \pi_i \frac{\partial(\sigma - s_j)}{\partial \phi_i}, \tag{10}$$

where

$$\pi_i \equiv \frac{\phi_i}{[1 + \lambda(\beta + \phi_j)]\psi}$$

represents her sniping probability.

The first term is the marginal improvement in the sniping probability, and the second stands for a decline in the profit. These terms can be seen as the marginal benefit and cost of being faster. The structure of the marginal gain of increasing ϕ_i is the same as in the benchmark case, while it depends on the speed of the competitor.

Furthermore, we need a cross derivative to obtain the reaction of BR_i.

$$\frac{\partial w_i}{\partial \phi_j} = \left[(\sigma - s_j)\frac{\partial^2 \pi_i}{\partial \phi_j \partial \phi_i} + \frac{\partial \pi_i}{\partial \phi_i}\frac{\partial(\sigma - s_j)}{\partial \phi_j}\right] + \overbrace{\left[\frac{\partial \pi_i}{\partial \phi_j}\frac{\partial(\sigma - s_j)}{\partial \phi_i} + \pi_i \frac{\partial^2(\sigma - s_j)}{\partial \phi_j \partial \phi_i}\right]}^{\oplus}. \tag{11}$$

When the opponent (HFT *j*) increases her speed, it affects both the marginal benefit and cost of being faster for HFT *i*. The first component of (11) is a change in the marginal benefit that stems from a marginal improvement in π_i. A faster opponent (i) increases or decreases the marginal improvement in the sniping probability and (ii) raises the sniping profit, making it more worthwhile to have a higher π_i. These are the first and second terms in the first brackets in (11). At the same time, a faster opponent reduces the (endogenous) marginal cost of being faster for HFT *i*. Intuitively, (iii) since a faster opponent makes HFT *i* less likely to snipe, she does not need to care much about the adverse price movement of being faster. Moreover, (iv) a faster opponent becomes more insensitive to HFT *i*'s speed-up due to the same logic that states that a higher λ makes *s* less sensitive to ϕ in the benchmark case. As a result, the second term is positive, as is the total effect of ϕ_j on w_i, i.e., a faster opponent renders speeding up more profitable for HFT *i*.

Moreover, "tit for tat" due to complementarity can be strong enough when the opponent is sufficiently fast.

Lemma 3. *There is a unique $\phi_j = \phi_0$ such that*

$$\frac{d^2 BR_i(\phi_j)}{d\phi_j^2} > 0 \Leftrightarrow \phi_j > \phi_0. \tag{12}$$

Proof. See Appendix B.3. □

If ϕ_j is sufficiently high, the negative effect of ϕ_i on the expected profit becomes minimal. This is because a very fast opponent makes it extremely difficult for HFT i to snipe. Thus, she barely cares about the negative impact of ϕ_i on the price. In addition, a fast opponent as a market maker tends to be highly insensitive to a change in ϕ_i because she estimates that being sniped by HFT i is not likely to happen. Both of these effects strongly prompt an incentive of HFT i to be faster, making the best response function convex.

3.2 Equilibrium Speed

To see if (9) has a symmetric solution, we first observe that $BR_i(0) > 0$, 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 (12), this implies that multiple equilibria can arise. We focus on the symmetric equilibria.

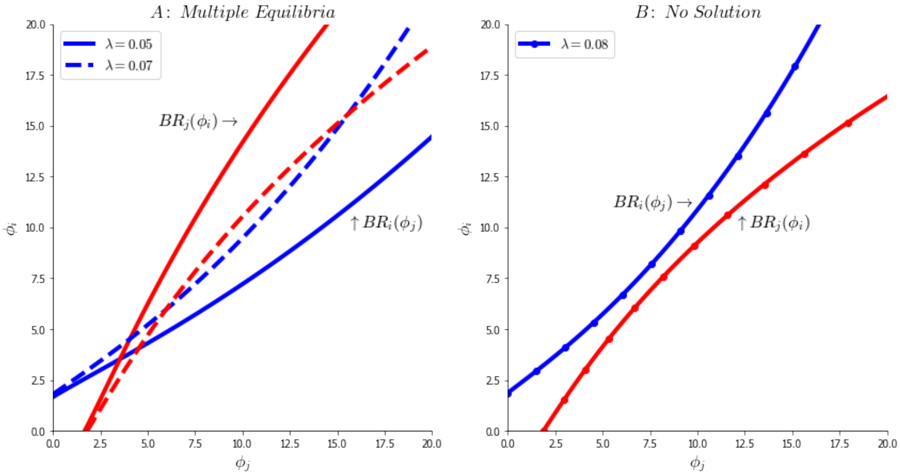
Proposition 5. *(i) There is a unique $\lambda = \lambda_0$. If $\lambda > \lambda_0$, no bounded solution exists. If $\lambda \leq \lambda_0$, there are two bounded solutions to $BR(\phi) = \phi$. The low-ϕ solution is stable and the high-ϕ solution is unstable.*
(ii) In the stable equilibrium, $\phi^* \equiv \phi_i = \phi_j$ is increasing in λ.

Proof. See Appendix B.3. □

Note that a higher λ has the same implication as a higher ϕ_j for the sensitivity of s to ϕ_i and for the improvement in the sniping probability π_i. Thus, due to the same logic as in Lemma 3, a sufficiently high λ makes the complementarity strong enough to eliminate a bounded solution, i.e., $\phi = \infty$ is always optimal. On the other hand, when λ is small, we obtain bounded symmetric equilibria.

Following the convention (Hendershott and Mendelson, 2000; Zhu, 2014), we use stability as an equilibrium selection criterion. The unstable equilibrium is not robust to a small

Figure II: Best Response Functions



perturbation in a parameter value, whereas the stable one does not diverge even if a parameter changes slightly. Thus, our focus is on the small-ϕ solution.

Figure II provides the best response functions for different values of λ. In multiple equilibria, a small-ϕ solution is stable, while a higher ϕ makes "tit for tat" strong and the solution can explode.

As Figure II indicates, a longer delay has the same effect on the optimal speed as in the benchmark, i.e., it increases the marginal benefit of being faster. Thus, the best response function shifts upward, leading to a higher speed in the stable equilibrium.

3.3 Effect of Speed Bumps on the Spread

The effect of a bump on the spread and adverse selection can be derived analytically. Due to "fast market making," adverse selection risk is mitigated by fast market makers, and it helps λ protect them. However, in the symmetric equilibrium, the increase in market makers' speed occurs identically for snipers. Then, the above-mentioned effect of fast market making is offset by the increase in the snipers' speed.

Since the strategic complementarity is sufficiently strong, this arms race outweighs the direct protection of the speed bump, expanding the spread.

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

Proof. See Appendix B.4. □

The introduction of a speed bump or a longer expected delay can backfire not only in terms of speed but also of adverse selection. In the benchmark model, we have $\frac{ds}{d\lambda} = 0$ because the speed-up by the single HFT is an indirect consequence of the speed bump and cannot offset the direct protection of market makers.

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, shifting them upward, as shown in Figure II. This triggers an arms race with positive externality, amplifying the first 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.

3.4 Comparison with Traditional Models with an Exogenous Cost

The results in the previous subsection run counter to traditional models with an exogenous cost of speeding up. To illustrate this, consider a model with non-strategic HFTs; Instead of an endogenous cost, we introduce an exogenous sunk cost of being faster denoted by $C(\phi_i) = \frac{c}{2}\phi_i^2$, as in Foucault et al. (2016). To make the comparison clearer, we call our model in Subsection 3.1 the *strategic model*.

If the strategic motive is absent, the FOC in (10) and cross derivative in (11) are modified as follows:

$$\begin{aligned} w_i \equiv \frac{\partial W_i}{\partial \phi_i} &= (\sigma - s_j)\frac{\partial \pi_i}{\partial \phi_i} - c\phi_i, \\ \frac{\partial w_i}{\partial \phi_j} =& (\sigma - s_j)\frac{\partial^2 \pi_i}{\partial \phi_j \partial \phi_i} + \frac{\partial \pi_i}{\partial \phi_i}\frac{\partial(\sigma - s_j)}{\partial \phi_j}. \end{aligned} \tag{13}$$

The second term of (10) that represents an endogenous marginal cost is replaced by the exogenous marginal cost, $c\phi_i$, and the effect of the opponent's speed via the strategic motive, denoted by the second set of brackets in (11), disappears from (13).

We focus on a symmetric equilibrium and obtain the following results.

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. □

The third column (panels A3, B3, and C3) of Figure III shows these results. As we have established, if HFTs are strategic, HFT j's speed-up improves w_i through the second brackets in (11), while this effect is absent in the traditional models.

To understand this intuition, note that the profit function of an HF sniper is roughly given by

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

where π_i is the sniping probability of HFT i. In this formulation, the interaction between HFT i and j occurs only through π_i, i.e., probability of successful sniping.

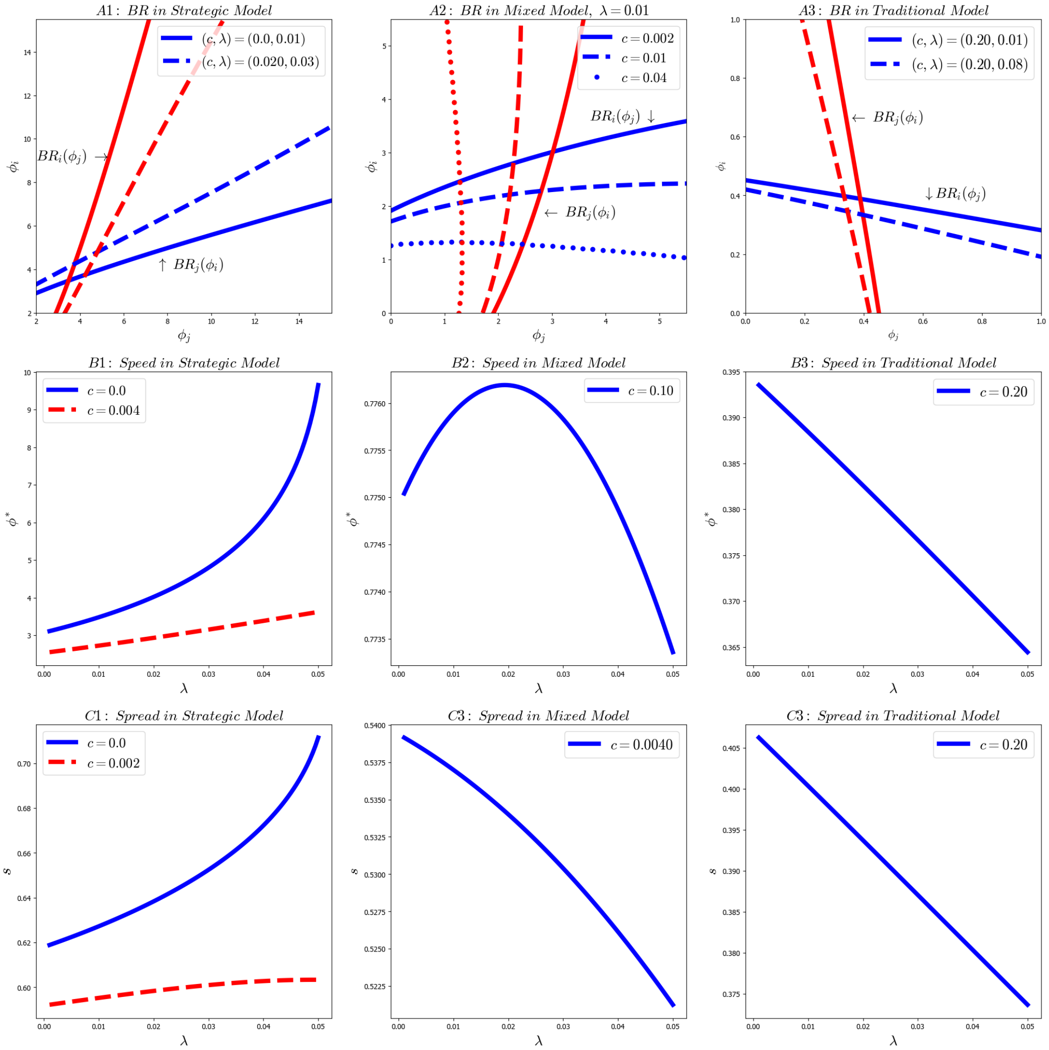
A marginally faster opponent in a traditional model affects HFT-i's decision by making sniping more difficult, i.e., ϕ_j changes V_i only by reducing π_i. Since the exogenous cost is sunk, HFT i must pay it anyway. By contrast, her speed investment pays out only if her sniping attempt is fulfilled. Therefore, 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 applies to a speed bump as well: a bump 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 mitigate adverse selection, which is replicated by Proposition 7.

Once HFTs become aware of the price impact of their speed choice, they perceive the price as $s = s(\phi_i, \phi_j, \lambda)$, generating a new channel through which ϕ_j affects V_i. In the previous subsection, we showed that this modification overturns the result of the benchmark. This is because the pricing behavior of market makers becomes insensitive to an increase in ϕ_i (the endogenous marginal cost declines) when the opponent as a market maker becomes faster or a bump is expected to be longer. In our strategic model, this endogenous cost channel works against the traditional exogenous cost.

3.5 Strategic Complementarity or Substitution

As Proposition 7 and the following discussion suggest, an exogenous sunk cost tends to make an arms race exhibit strategic substitution, while an endogenous cost in our model promotes complementarity (Proposition 4). These results imply that introducing an exogenous sunk cost in our strategic model allows us to explain both complementarity and substitution in an arms race, as well as the positive and negative reaction of the spread to a

Figure III: Effect of Exogenous Cost



Note: The first row plots the best response functions with different $(c\lambda)$, the second and third rows plot the optimal speed and the equilibrium spread, respectively, against the expected length of a speed bump with different values of c. In each row, $c = 0$ corresponds to the strategic model with no exogenous costs. The figures in the third column replicate the traditional results.

bump, by changing the parameter values.

Consider a model that is the same as in Subsection 3.1, except that HFT i solves

$$\max_{\phi_i} W_i(\phi) = \frac{1}{1+\lambda(\beta+\phi_j)}\frac{\phi_i}{\psi}(\sigma - s_j) - C(\phi_i),$$

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

where $C(\phi) = \frac{c}{2}\phi^2$ is the exogenous sunk cost of speed.

The best response functions, equilibrium speed, and spread are provided in Figure III.

Figure IV: Summary

Strategic Model (this paper) ← → **Traditional Model**

Arms Race	Strategic Complementarity	Complementarity/ Substitution	Strategic Substitution
Speed, Adverse Selection, and Spread	Increasing	Hump-Shaped	Decreasing
Speed Bump	Backfire	Ineffective	Effective

The strategic model in Subsection 3.1 corresponds to $c = 0$. As expected, a small exogenous cost (panels A1 and A2 with $c = 0.002$ and 0.01) offers strategic complementarity and an increasing ϕ^* against λ (panel B1). This leads to the increasing s against λ (panel C1).

As long as $c > 0$, ϕ^* becomes flatter as λ rises and can be hump shaped (panels B1 and B2). This is because a higher λ makes C more salient. That is, a longer delay makes sniping more difficult, and HFTs become more reluctant to pay the sunk cost. In this situation, an arms race exhibits both complementarity and substitution (Panel A2 with $c = 0.04$), i.e., the BR curves become hump shaped. When λ is small enough, the model is close to our strategic model with complementarity, while a large λ makes it similar to the traditional model with substitution. As a result, the direct protection of market makers by the bump dominates the speed increase in the high-λ region, and s slopes downward, as shown in Panels C1 with $c = 0.002$ and C2.

Finally, if the exogenous cost is sufficiently large, the economy reverts to the traditional world with no strategic speed choice (panels A3, B3, and C3). The competition exhibits global substitution, and the equilibrium speed and the spread are downward sloping against λ, i.e., a speed bump mitigates adverse selection.

Overall, we can think of the strategic model with complementarity and the traditional models with substitution as two extremes. By changing the significance of the exogenous cost, c, we can explain intermediate cases. Moreover, we have established that a speed bump, λ, also works to adjust the relative significance of the exogenous cost because a longer delay makes it more salient. This premise is summarized in Figure IV and has some empirical implications.

4 Empirical Implications and Policy Discussion

Our 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 and adverse selection for market makers.

As Subsection 3.5 demonstrates, an arms race exhibits strategic complementarity or substitution depending on the exogenous cost of speed investment and the expected length of a speed bump. As summarized in Figure IV, 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 bump and a marginally longer delay would effectively reduce (resp. widen) a spread.

In the real world, the speed cost for HFTs involves two factors. The first is a large sunk cost to develop a high-speed communication technology or investigate the optimal location of an 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 the Chicago and New York exchanges. The second cost can be a relatively small subscription fee to access these technologies developed by communication service companies or to colocate an information server to an exchange platform.[23]

Depending on the condition of the financial markets, our model suggests different implications. If HFTs take the first investment approach, the traditional model fits better: an arms race involves strategic substitution and a bump is effective. In contrast, if the exogenous cost is relatively small compared to the total profit, our strategic model is more appropriate, suggesting the complementarity and detrimental effect of a bump. This comparison can also be applied to the market power of a high-frequency financial institution because the endogenous cost 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, we can vary the length of a speed bump or the distribution of a random delay to test the implications of λ on a spread. If a bump is expected to be long, a marginally longer delay can be effective, slowing HFTs down due to the substitution in the traditional model. On the other hand, if a platform tries to avoid a delay cost by keeping the length

[23] For example, monthly payments to plug into NSYSE and NASDAQ amount to $10,000∼$22,000.

of a bump minimal, the introduction of a bump can aggravate adverse selection since HFTs become faster. Although we can compare the length of bumps in each platforms (i.e., bumps in the ANEO and TMX are longer than those in the IEX or Chicago Stock Exchange), our model shows that the *level* of λ also matters. Deriving the critical (c, λ) involves numerical calculation in our model, 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.

4.1 Policy Implications

Recently, exchange platforms have experienced declines in their revenue from transaction 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. Namely, one of the suppliers of the speed technology, which we did not specify in the model, can be the exchange platforms.

The SEC is concerned about this skyrocketing price of fast data feed and issued a proposal to block the exchanges to raise the fee in March 2018. Another HFT-related policy that SEC 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 all orders being affected by the speed bump. In a nutshell, the SEC tries to curb the price of speed technologies supplied by the exchange platforms, while they prompt the introduction of a speed bump.

Importantly, our model suggests that the introduction of a delay does not necessarily conflict with a provision of the 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 FBA, as long as competing platforms can copy the innovation with a low cost. In our model, a bump does not always mitigate adverse selection, while an exchange platform may have an incentive to introduce a bump to obtain the higher demand for the speed technology. More detailed analyses can be our future research topics, but aforementioned mechanism can provide another explanation for why "the market cannot fix the market."

5 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, since it widens as HFTs get faster. This endogenous cost tends to be insensitive to an increase in speed by HFTs when a speed bump is expected to be longer. This is because market makers behave as if the share of HFTs is small under a longer expected speed bump and do not care much about a speed increase. Then, if a bump is introduced or the length of a delay becomes longer, the marginal cost of speed diminishes, leading to a higher equilibrium speed of HFTs.

We also consider an arms race among multiple HFTs who serve both as snipers and as market makers. When the significance of the exogenous speed cost is not high compared to the endogenous cost, or when the expected length of a delay is relatively short, a speed competition can show strategic complementarity. In this situation, a longer speed bump triggers a positive externality among HFTs, leading to a very fast equilibrium speed. As a result, the increase in HFTs' trading speed dominates a direct reduction of an arrival rate of HFTs by a speed bump. That is to say, a longer bump exacerbates adverse selection and widens the equilibrium spread. The opposite holds in the case of substitution.

Thus, our model, which incorporates strategic speed choice and the endogenous cost (i.e., the equilibrium spread), generates results that are opposite to the traditional models, in which an arms race exhibits strategic substitution, the speed and adverse selection decrease with the length of the delay, and a bump is effective. By adding a traditional exogenous cost to our strategic model, we can derive a rich description of the characteristics of an arms race among HFTs and can explain the somewhat ambiguous effects of speed bumps on spreads and adverse selection.

References

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.

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.

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, Thierry Foucault et al., "False news, informational efficiency, and price reversals," *Banque de France, Working Paper*, 2014, (513).

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.

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.

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

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.

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 Continuous Updating by Market Makers

We 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.[24]

Consider a market maker who updates her limit order by resubmitting competitive s_t. The competition drives $s_t = E[v|\text{trade at } t]$. Since δ is stochastic,

$$
\begin{aligned}
s_t &= \int_0^\infty be^{-b\delta} E[v|\text{trade at } t{,}\delta] d\delta \\
&= \int_t^\infty be^{-b\delta} E[v|\text{trade at } t{,}\delta] d\delta + \int_0^t be^{-b\delta} E[v|\text{trade at } t{,}\delta] d\delta \qquad (14) \\
&= 0 \times \int_t^\infty be^{-b\delta} d\delta + \int_0^t be^{-b\delta} \frac{\phi}{\phi+\beta} \sigma d\delta \qquad (15) \\
&= (1-e^{-bt}) \frac{\phi}{\phi+\beta} \sigma.
\end{aligned}
$$

In (14), 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 (15) 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 (15).

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.

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

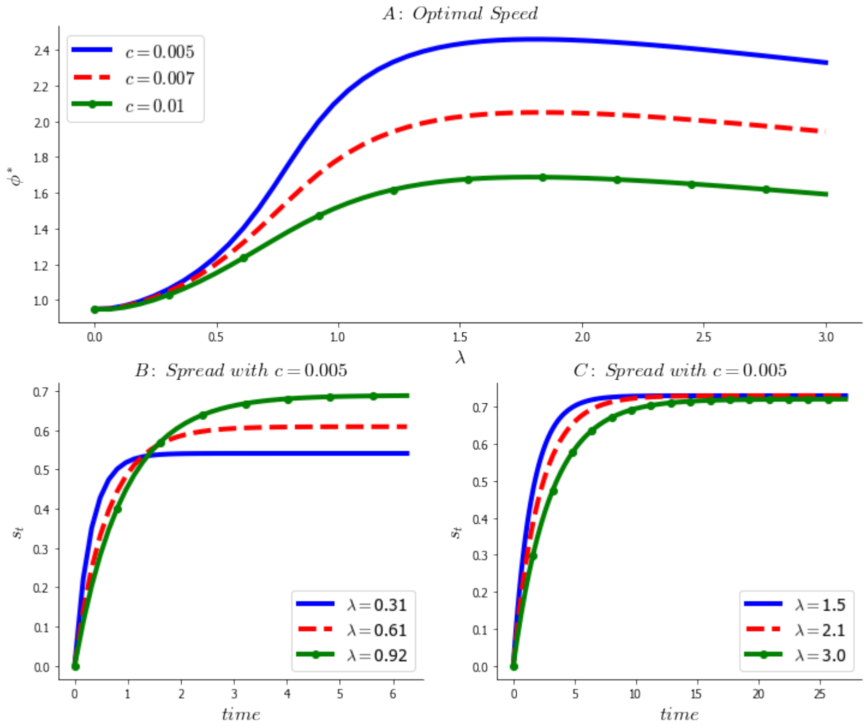
$$
\begin{aligned}
\max_{\phi} W(\phi) &= E_\delta \left[\int_0^\infty \phi e^{-\psi t} e^{-\eta\delta} (\sigma - s_{t+\delta}) dt \right] - \frac{c}{2}\phi^2, \\
\text{s.t., } s_t &= (1-e^{-bt}) \frac{\phi}{\phi+\beta} \sigma.
\end{aligned}
$$

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 V can be discussed by using the ingredients we 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.

[24] We can eliminate the possibility that an informed HFT splits her orders across the time 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. We 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 15, we can show that the mixed strategy is not an equilibrium because waiting is not credible.

Figure V: 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, $\frac{d\phi^*}{d\lambda} < 0$.

As a result, ϕ^* takes the hump-shaped curve as depicted by Panel A in Figure V. Given λ, a larger cost slows HFT down as we can see from a parallel shift in the curves.

As (14) 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

$$\begin{aligned}\frac{ds_t}{d\lambda} &= -\frac{1}{\lambda^2}\frac{\partial s_t}{\partial b} + \frac{d\phi}{d\lambda}\frac{\partial s_t}{\partial \phi} \\ &= -\frac{1}{\lambda^2} t e^{-bt} \frac{\phi}{\beta + \phi} + \frac{d\phi}{d\lambda}(1 - e^{-bt})\frac{\beta}{(\phi + \beta)^2}. \end{aligned} \tag{16}$$

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 (16). However, as a higher λ may push ϕ up or down, it increases or decreases the spread, as the second term in (16) suggests. When $\frac{d\phi}{d\lambda} > 0$, the second effect competes with the first effect: the safe interval gets longer, while the HFT becomes faster. The result is provided by Panel B in Figure V.

Whether the first effect dominates the second effect depends on t. From (16), 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 V. 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.2 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. We extend our model to relax this limitation.

A.2.1 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, $\delta = 0$. We 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), we 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.2.2 Strategies 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.[25] 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. We 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 $A = 2$.[26] 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 VI and VII 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

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

[26] 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 VI: Strategy 1

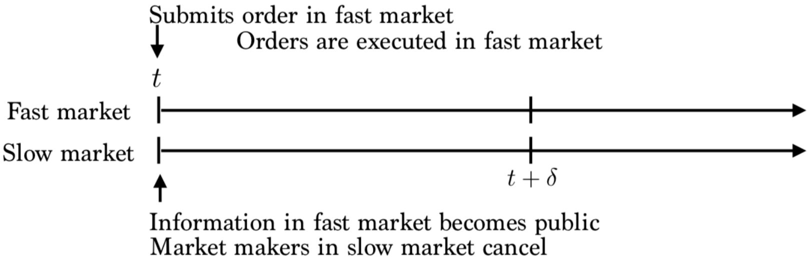
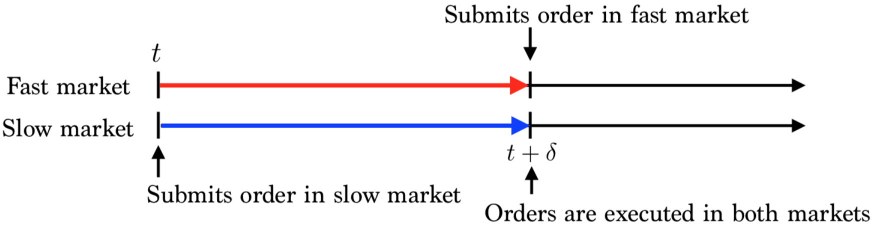

Submits order in fast market
Orders are executed in fast market
t
Fast market
Slow market
$t+\delta$
Information in fast market becomes public
Market makers in slow market cancel

Figure VII: Strategy 2


Submits order in fast market
t
Fast market
Slow market
$t+\delta$
Submits order in slow market
Orders are executed in both markets

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

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

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.2.3 Behavior of Market Makers

We take q an exogenous parameter in our 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

$$\begin{aligned} V_0 =& E_\delta \left[\theta \left(\int_0^\delta s_0 \beta e^{-(\beta+\gamma)t} dt + \int_\delta^\infty e^{-\psi(t-\delta)-(\beta+\gamma)\delta} (\beta s_0 + \phi(s_0 - \sigma)) dt \right) \right. \\ & \left. + (1-\theta) \int_0^\infty e^{-\psi t} (\beta s_0 + \phi(s_0 - \sigma)) dt \right]. \end{aligned} \tag{17}$$

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.[27] 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 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

$$\begin{aligned} V_\lambda =& E \left[\theta \left(\int_0^\delta s_0 \beta e^{-(\beta+\gamma)t} dt + \int_\delta^\infty e^{-\psi(t-\delta)-(\beta+\gamma)\delta} (\beta s_0 + \phi(s_0 - \sigma)) dt \right) \right. \\ & \left. + (1-\theta) \int_0^\infty s \beta e^{-\psi t} dt \right]. \end{aligned} \tag{18}$$

The intuitions behind the first line are the same as those in (17). 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.2.4 Equilibrium in the Trading Stage

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

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

$$s_j = \begin{cases} \frac{\phi}{\phi+\beta+\beta\psi\theta\frac{\lambda}{1+\lambda(\beta+\gamma)(1-\theta)}} & \textit{for } j = 0, \\ \frac{\phi\theta}{\beta+\phi\theta+\lambda\beta(\beta+\gamma)\left(1+\frac{\phi\theta}{\beta+\gamma}\right)} & \textit{for } j = \lambda. \end{cases} \tag{19}$$

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

These formulae show the following:

[27] 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, $\Pr(\text{Liq. trade at } t) = \beta e^{-\psi t} + \beta e^{-(\beta+\gamma)t} \int_0^t \phi e^{-\phi\tau} d\tau = \beta e^{-(\beta+\gamma)t}$.

Corollary 1. *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*

$$\theta = \begin{cases} 0 & \text{if } \lambda Q(\beta+\gamma) > \frac{\phi+\beta}{\beta}, \\ \theta^* \in [0,1] & \text{if } \lambda Q(\beta+\gamma) \in [1, \frac{\phi+\beta}{\beta}], \\ 1 & \text{if } \lambda Q(\beta+\gamma) < 1, \end{cases} \tag{20}$$

with

$$\theta^* = \frac{(\phi+\beta) - \beta(\beta+\gamma)Q\lambda}{1+\lambda Q(\beta+\gamma) - \beta\lambda(1-\lambda\eta Q)} \frac{1+\lambda\eta}{\phi}. \tag{21}$$

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 (20), 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, we have the following:

Corollary 2. *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 we can see from (19), 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.2.5 Strategic 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

$$w_A(\phi) = \begin{cases} w_1(t) = (1-q)(\sigma - s_0(\phi,\theta)) & \text{if } \theta \in [0,1) \\ w_2(t) = E_\delta \left[e^{-(\beta+\gamma)\delta}(\sigma - s_\lambda(\phi,\theta)) \right] & \text{if } \theta = 1, \end{cases} \tag{22}$$

the spreads in (19) as functions of (ϕ, θ), and the equilibrium strategy θ given by (20). 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 (22). 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 we obtain w_2 in (22).

A.2.6 Short expected delay

As we have established in (20), 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, 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 IX.

A.2.7 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

$$\hat{\phi} \equiv \beta \left[\lambda(\beta+\gamma)Q - 1 \right]. \tag{23}$$

When $\phi < \hat{\phi}$ (resp. $\phi > \hat{\phi}$), the strategy of the HFT is $\theta = 0$ (resp. $\theta = \theta^*$). As we have 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 4. *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 (22) and taking derivative immediately derive the result. □

Since the objective function W switches at $\hat{\phi}$, we have 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 VIII 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, we 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 2, 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 IX for the visual illustration of the effect of λ.

Figure VIII: $W(\phi)$ with different λ

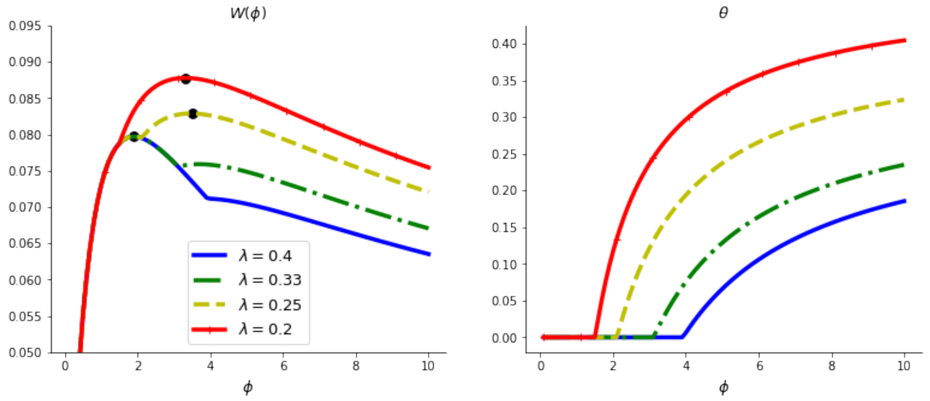
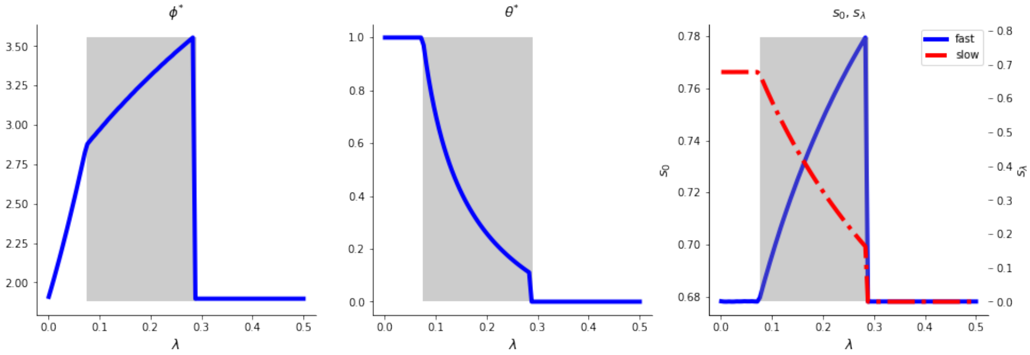


Figure IX: Effect of λ



A.2.8 Adverse Selection Cost

Figure IX 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 IX. 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 1, 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 Lemma 2 and Proposition 2

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

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

where

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

Therefore,

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

with

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

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. We can easily check that this condition is expressed as (5). If this is not satisfied, we have $\phi^* = \infty$.

When (5) holds, the $\phi^* > 0$ solves $1 = \varepsilon(\phi)$, and some tedious calculations show that the solution is given by (8). The second statement is obvious from (8).

B.2 Proof of Proposition 3

By taking a derivative, we have

$$\frac{ds}{d\lambda} \sim -\phi\psi + \frac{d\phi}{d\lambda}(1+\lambda\eta). \tag{24}$$

Moreover, by the implicit function theorem,

$$\frac{d\phi}{d\lambda} = \frac{\eta g^2 - \phi^2\gamma}{\eta g^2 + \beta\psi^2(1+\eta\lambda)^2} \tag{25}$$

where $g \equiv \phi + \beta(1+\lambda\psi)$. By substituting (25) for the one in (24),

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

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

B.3 Proof of Lemma 3 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

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

We can check that $r\frac{R'}{R} > 1$ 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 we are 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, we can define ϕ_0 be the solution of $r = r^*$ and 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:

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

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

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

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. We can check that

$$\frac{ds}{d\lambda} \sim \phi\psi[\psi(1+\lambda\psi) - 2\eta(1+\lambda\eta)] - \lambda\eta\phi\psi(1+\lambda\beta). \tag{26}$$

Since the symmetric equilibrium solves

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

we know that $\phi = \sqrt{\beta\eta}(1+\lambda\psi) \geq \beta \geq 1$. By using these conditions, we can check that the RHS of 26 is positive.

B.5 Proof of Proposition 7

First of all, the traditional model satisfies the SOC: By letting $\Gamma \equiv \phi_i + \beta(1+\lambda\psi)$,

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

Then, the FOC to solve is

$$c\phi_i = \frac{\beta}{\psi\Gamma}\frac{Y(\eta)}{Y(\psi)} \equiv K(\phi_i, \phi_j, \lambda), \tag{27}$$

with $\psi \equiv \sum_i \phi_i + \beta$ and $\eta \equiv \phi_j + \beta$. We can easily check that the RHS of (27) is decreasing in ϕ_i. Since K is decreasing in ϕ_j and λ around the symmetric equilibrium, we can prove that $\frac{dBR_i}{d\phi_j} < 0$ and $\frac{d\phi}{d\lambda} < 0$. Since 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

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

These functions have the following properties:

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

Thus, if $\lambda\eta Q < 1$, we have $L(1) < R(1)$, 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 $L(0) \gtreqless R(0)$. 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 (21). Finally, if $\lambda\eta Q > (\beta + \phi)/\beta$, we have $R < L$ for all $\theta \in [0, 1]$. Thus, $\theta = 1$ is the optimal strategy.