

The Optimal Speed to Go over Speed Bumps in Financial Markets*

Jun Aoyagi[†]

August 11, 2020

[Click here for the Latest Version](#)

Abstract

A speed bump in financial markets is an intentional delay imposed on trade execution. Its primary purpose is to mitigate asymmetric information by slowing down high-frequency traders (HFTs). In contrast to its intended purpose, this paper shows that a speed bump can endogenously fuel HFTs' investment in high-speed technology (i.e., the crowding-in effect). I investigate when the crowding-in effect dominates a speed bump's intended effect and explain its ambiguous impact on speed acquisition and equilibrium market quality.

Keywords: speed acquisition; speed bump; high-frequency trading; adverse selection.

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

*First draft: June 2018. This paper was previously circulated under the titles "Speed Choice by High-Frequency Traders with Speed Bumps" and "The Dark Side of Regulating Fast Informed Trading." This paper was awarded the *2018 Moriguchi Prize* by the Institute of Social and Economic Research, Osaka University.

[†]Department of Economics, University of California at Berkeley. E-mail jun.aoyagi@berkeley.edu. I deeply acknowledge that Nicolae Gârleanu and Christine Parlour provided invaluable assistance. I appreciate the constructive comments from Kosuke Aoki, Kerry Back, Michael Brolley, Jesse Davis, Joel Hasbrouck, Terry Hendershott, Ryo Horii, Ohad Kadan (discussant), Ulrike Malmendier, Hayden Melton, Sophie Moinas (discussant), Emi Nakamura, Yoko Shibuya, David Sraer, Alberto Teguia, Yingge Yan, Haoxiang Zhu, and Marius Zoican, and from participants at the 16th FRA conference, the 9th Stern Annual Microstructure Conference, the ISER at Osaka University, and finance seminars at UC Berkeley, University of Tokyo, and Financial Services Agency.

1 Introduction

Creating a level playing field is the centerpiece of financial regulation. Traditionally, regulations are designed to mitigate asymmetric information by requiring firms to disclose material news so that all traders obtain information of equal quality (e.g., Sarbanes–Oxley Act of 2002). In recent years, technological innovations add a new dimension to traders’ information acquisition and regulation—speed. Traders called high-frequency traders (HFTs) exploit sophisticated communication tools to extract information at the speed of light. Their speed advantage exposes other traders to adverse selection risk, as HFTs can “snipe” liquidity at stale prices before other traders notice.

Some exchange platforms try to slow down HFT by introducing *speed bumps*.¹ A speed bump imposes an intentional delay on arrival or execution of trading orders at a market, aiming to protect slow traders and liquidity providers. For example, Aequitas NEO and TSX Alpha, both Canadian exchanges, apply a few milliseconds of random delay to slow down liquidity-taking orders. As summarized by Table 1 (see Appendix B for more details), many other exchanges have also adopted or proposed adopting a speed bump. Speed bump design differs across exchanges but, reflecting its purpose, all speed bumps in operation (except for IEX) are *random* and *asymmetric*, meaning that only liquidity-taking orders are delayed.

Nevertheless, the study of speed bumps is still in its infancy. This is the first paper that studies the effect of a speed bump on equilibrium market quality by incorporating strategic speed acquisition by HFTs. As the main contribution, this paper shows that a speed bump has a *crowding-in* effect on the HFTs’ speed acquisition, meaning that it can increase the equilibrium speed of HFTs. The crowding-in effect may undermine or even dominate the intended impact of a speed bump. Thus, depending on market structures, a speed bump can worsen adverse selection and contribute to creating an “un-

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

Table 1: Design of Speed Bumps

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

Note: This table reproduces Table 3 in [Baldauf and Mollner \(2017\)](#) and Table 1 in [Khapko and Zoican \(2019\)](#) with some updates. * indicates a random speed bump. **As of December 2019, ICE announced the removal of a speed bump from NYSE American based on their finding that a speed bump worsens the liquidity and the trading share of the exchange. As of May 2020, exchanges with a random speed bump do not announce the distribution that a random delay follows.

even playing field.” The main model takes a random asymmetric speed bump as a primary market structure, while a deterministic delay is also analyzed as a special case.

A simple benchmark model of a quote-driven market (à la [Glosten and Milgrom, 1985](#)) with a single HFT is studied to separate the key mechanism. The HFT as a liquidity taker (sniper) acquires speed by incorporating a random speed bump, and a competitive market maker sets a bid-ask spread that reflects the adverse selection cost. As the literature points out (e.g., [Biais, Foucault and Moinas, 2015](#)), the faster the HFT, the more severe adverse selection the market maker faces. Thus, a speed-up by the HFT widens the bid-ask spread and reduces her trading profit. Put differently, the bid-ask spread works as an endogenous upward-sloping cost of being faster for the HFT due to the positive price impact of speed.

A speed bump brings about two positive impacts on the HFT’s speed acquisition. Firstly, the market maker is less likely to be picked off due to a speed bump. She downplays an increase in the HFT’s speed, and the price impact of speed weakens. Even if the HFT becomes faster, her sniping profit shrinks only slightly, meaning that the endogenous marginal cost of speed declines. Secondly, a speed bump’s direct effect exogenously tightens the spread by curbing the arrival of the HFT. It

reduces the HFT's trading cost, boosts the sniping reward, and increases the marginal benefit of being faster. These are the *equilibrium* channels through which a speed bump strengthens the HFT's speed acquisition.

As the existing models argue (see Subsection 1.1), a speed bump also has a negative equilibrium effect on speed acquisition, which I call the crowding-out effect. The crowding-out effect arises because a speed bump reduces the probability that the HFT can successfully snipe a stale limit order, and paying exogenous sunk costs for speed becomes less worthwhile. I investigate when and why one of the crowding-in and crowding-out effects becomes dominant. A switch occurs depending on some observable parameters, such as the exogenous marginal cost of speed, the expected length of a speed bump, and the asset's volatility. They increase or decrease the *effective* exogenous cost of being faster (i.e., the cost per successful sniping) and magnify or shrink the crowding-out effect of a speed bump.

The second part of the model considers a more realistic environment. It analyzes speed competition among multiple HFTs—an “arms race”—where they serve both as snipers and high-frequency market makers. I show that an arms race can involve strategic complementarity. An increase in market makers' speed has the same effect as an extension of a speed bump, as both parameters reduce market makers' risk of being picked off. Therefore, a speed-up by an HFT as a market maker can promote other HFTs' speed acquisition, triggering a positive externality. The externality amplifies the effect of a speed bump on speed acquisition. A longer delay makes each HFT willing to be faster (resp. slower) when the crowding-in (resp. crowding-out) effect is dominant, but it also triggers a stiffer (resp. less stiff) arms race, amplifying the original effect.

Given the comparative statics regarding speed acquisition, I analyze the adverse selection problem (i.e., the equilibrium spread). Firstly, a speed bump exogenously reduces the spread by directly hampering the HFTs' arrival and alleviating adverse selection. Secondly, it affects the spread through the equilibrium speed acquisition of HFTs: the crowding-in effect versus the crowding-out effect.

Since an arms race involves amplification, the effect through HFTs' speed acquisition can dominate the direct impact. Therefore, my model demonstrates that a speed bump can be beneficial or detrimental to liquidity providers, depending on the design of a speed bump and the exogenous cost of speed.

The crowding-in effect in my model provides novel and somewhat optimistic implications regarding exchanges' incentive to adopt speed bumps. For example, the model shows that alleviating adverse selection and increasing HFTs' demand for speed can simultaneously be achieved by appropriately designing speed bumps (e.g., adjusting the expected length of a delay). With some parameter values, the crowding-in effect dominates the crowding-out effect on speed acquisition, but it cannot overturn the speed bump's direct impact on the spread. Then, the HFTs' optimal speed increases, while a speed bump mitigates the adverse selection problem. In reality, exchanges earn a large portion of profits by selling speed services to traders (e.g., colocation services; direct data feeds). Hence, the above argument suggests that there exist speed bumps that can achieve both the profit maximization of exchanges and the protection of liquidity providers.

1.1 Literature review

This paper contributes to the literature on HFT and market structure (see Jones, 2013; O'Hara, 2015; Menkveld, 2016 for reviews). The speed acquisition problem of HFT has been analyzed in the existing studies, such as Foucault, Roell and Sandas (2003), Liu (2009), Foucault, Kadan and Kandel (2013), Foucault, Kozhan and Tham (2016).² However, the existing models study the problem in the absence of either a strategic motive of HFTs in speed acquisition or a slow market structure. My model integrates both factors and provides an insight into how a speed bump affects equilibrium market quality.

Moreover, my model shares the same interests as the studies on slow market structures, such

²For the studies on high-frequency market making, see Aït-Sahalia and Saglam (2013), Han et al. (2014), Hoffmann (2014), and Conrad et al. (2015), Bongaerts and Van Achter (2016).

as frequent batch auctions (Budish, Cramton and Shim, 2015; Haas and Zoican, 2016) and speed bumps (Baldauf and Mollner, 2017; Brolley and Cimon, 2017; Aldrich and Friedman, 2018).³ In contrast to my model, they do not consider speed acquisition by HFTs with a continuous domain and delay-sensitive endogenous costs. They focus on the positive side of the slow market structure and conclude that it mitigates adverse selection for market makers via its direct effect and the crowding-out effect on traders' speed. My model uncovers the opposing force, i.e., the crowding-in effect. It complements the literature by demonstrating that the impact of a speed bump can be positive, negative, and ambiguous depending on other market structures.

The scope of the literature extends to empirical findings.⁴ Hu (2018) and Chakrabarty, Huang and Jain (2019) analyze the SEC approval of IEX as a National Securities Exchange (NSE) and find a net improvement in market quality measured by spreads. Shkilko and Sokolov (2016) exploit the interruptions of messaging caused by precipitation to find a reduction in quoted spreads.⁵ In contrast, Chen, Foley, Goldstein and Ruf (2017) analyze a speed bump at TSX Alpha and find an increase in quoted spreads. NYSE American (2019) also reports a decline in market quality and liquidity. Anderson, Andrews, Devani, Mueller and Walton (2018) report that a change in the market-wide effective spread and the price impact is insignificant. Consistent with my model, a recent experimental study by Khapko and Zoican (2019) finds that a marginally longer speed bump stimulates the traders' investment in speed when the execution price is endogenous.⁶

³The speed and frequency of order executions by a trading platform are also analyzed by Du and Zhu (2017), Kyle and Lee (2017), Menkveld and Zoican (2017), and Pagnotta and Philippon (2018).

⁴Other empirical studies on HFTs include, for example, Hendershott and Moulton (2011), Hasbrouck and Saar (2013), Riordan and Storkenmaier (2012), Ye et al. (2013), Frino et al. (2014), Boehmer et al. (2015), and Brogaard et al. (2015).

⁵Although the interruption by precipitation may have a similar effect to a speed bump, this phenomenon is not paid much attention by financial institutions. In contrast, traders anticipate a speed bump and take it into their decision making.

⁶If the price and trading profit are exogenously fixed, Khapko and Zoican (2019) find that a speed bump diminishes equilibrium investments into speed technologies.

2 Benchmark model

This section proposes a simple benchmark model to separate the main mechanism. Consider a one-shot exchange of an asset between three types of market participants: a single HFT, a competitive market maker, and liquidity traders (see Internet Appendix A for a sequential trading model). When the market opens at time $t = 0$, the asset has value $v = v_0 = 0$, which is common knowledge. $v = 0$ can be thought of as a situation where no mechanical arbitrage exists, e.g., SPY on NYSE has the same price as ES on CME.⁷

At stochastic time $t = t_0 \sim \text{Exp}(\kappa^{-1})$, however, there is a jump in the asset value, which renders $v = v_0 \pm \sigma$ with equal probability. A jump in v can be thought of as a new arbitrage opportunity triggered by a jump in ES on CME that is not yet reflected by SPY on NYSE (Budish et al., 2015). It could also stem from some public news, such as Fed announcements and the release of government statistics, or execution of Intermarket Sweep Orders (Baldauf and Mollner, 2017).

Following the convention of market microstructure (e.g., Glosten and Milgrom, 1985), I assume that each trader can take only a unit position. Moreover, traders use a *limit order* when they add liquidity to the limit order book (LOB) and a *market order* when they take liquidity. I abstract away from more sophisticated order types (such as “Post Only” and “Immediate or Cancel”) since they are not essential to the paper.

2.1 Traders

Market maker There is a risk-neutral competitive *market maker*. When the market opens, she posts a single-unit limit order by quoting ask and bid prices $(s, -s)$ and commits to trade at these prices. In

⁷SPY stands for the SPDR S&P 500 exchange traded fund, and ES stands for the S&P 500 E-mini futures contract. Both financial instruments track the S&P 500 index and are almost perfectly correlated in relatively long time intervals (e.g., 60 seconds). However, the correlation breaks down in a high-frequency time frame: Budish et al. (2015) estimate that the correlation coefficient becomes 0.1016 at 10 milliseconds and 0.0080 at 1 millisecond. The following discussion can be applied to any other highly correlated but distinct assets, or the same asset traded in different venue.

what follows, price s is referred to as the (half) spread.

At the same time, the market maker starts monitoring the market, anticipating the asset value jump. When the asset value jumps at $t = t_0$, the market maker cannot immediately know whether the jump is positive or negative, but her learning process (defined below) is put in motion. Upon privately learning v , the market maker revokes (or re-prices) her limit order to avoid being picked off by an informed HFT.

For the market maker, it takes stochastic time $T_M \sim \text{Exp}(\gamma)$ to learn v and cancel her quote after the jump in v . Thus, T_M can be thought of as the shortest necessary time for the market maker to modify her previous limit order. I take speed parameter γ as exogenous, but Subsection 3 considers an endogenous choice of γ by high-frequency market makers.⁸

High-frequency trader There is a risk-neutral *high-frequency trader (HFT)*. Before she enters the market, the HFT invests in a speed technology that provides speed ϕ . It could be colocation services, high-bandwidth connectivity, direct data feeds, or a sophisticated trading algorithm. The HFT also monitors the market and, conditional on the asset value jump, starts gathering and processing information on v . Upon learning it, the HFT immediately submits a market order to “snipe” a limit order on the LOB.⁹ This entire process takes stochastic time $T_H \sim \text{Exp}(\phi)$.

Liquidity traders There is a continuum of *liquidity traders (LTs)* who are exposed to a liquidity shock. The shock exogenously makes one of the LTs submit a buy or sell market order with equal proba-

⁸Since all events take place with Poisson distributions, the market maker at time $t = 0 + \epsilon$ does not have an incentive to change her initial quote, given that no events have happened until $t = 0 + \epsilon$.

⁹HFT 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.

bility, so LTs' behavior conveys no information. LTs capture slow passive investors with no material information in the real world. Their behavior stems from some exogenous reasons, such as hedging motives. Let $T_L \sim \text{Exp}(\beta)$ be the stochastic timing that the liquidity shock hits and an LT's market order arrives at the market. Parameter β captures an LT's arrival frequency, and I set $\beta > \frac{\gamma}{3}$ to make the model well defined.¹⁰ Since their trading motive is independent of the asset value and its price, they are exposed to the shock from $t = 0$ onward.

Figure 1: Flow of events

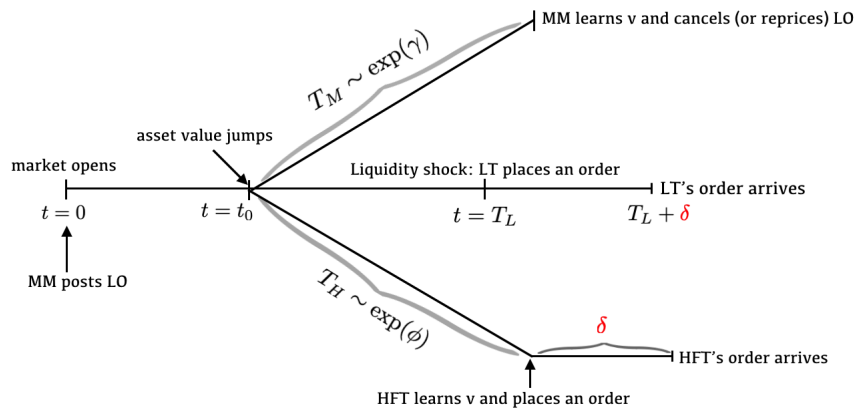


Figure 1 visualizes the timeline of the game (δ represents the speed bump defined below). In what follows, all random variables, $(t_0, v, \{T_i\}_{i=M,H,L})$, are assumed to be independent. Also, for simplicity, I suppose that there is no mechanical latency *per se* other than those mentioned above.¹¹

2.2 Speed bump

A market imposes a speed bump and delays execution of all market orders by the same length $\delta \geq 0$.¹² The model can describe both a random asymmetric speed bump and a deterministic asymmetric speed bump, but the primary focus of the model below is on the random ones. The case with a

¹⁰This inequality implies that the liquidity traders arrive at the market at a sufficiently high frequency.

¹¹Adding other sources of latency, such as the gateway-to-matching engine latency, does not qualitatively change the result.

¹²See Internet Appendix B for the case that a delay is applied in an iid manner.

deterministic delay is analyzed in Section 5 as a special case of a random speed bump, where the result turns out to be somewhat trivial.

Asymmetric speed bumps mean that a delay is imposed only on liquidity takers, whereas liquidity-providing orders (i.e., limit orders in this model) are not restricted. Unless otherwise noted, “a speed bump” refers to an asymmetric delay.¹³

Regarding a random delay, I assume that δ follows the exponential distribution with mean ε and is independent of other random variables. Internet Appendix C shows that the following results are robust to generalizing the distribution of δ with some parameter restrictions.

2.3 Equilibrium

The model is conceptualized as a sequential game with two stages, and the equilibrium concept is the subgame perfect equilibrium. In the first stage, the HFT decides the level of speed ϕ .¹⁴ When the market opens, the market maker submits a limit order with competitive prices, anticipating a confrontation with the informed HFT and liquidity traders. Given the jump in v , both of the HFT and the market maker start processing information, and the first trader who observes v and executes an order (or a liquidity trader) takes or cancels the limit order.

2.4 Equilibrium behavior of the market maker

Unless otherwise mentioned, I consider a *random* speed bump. Without loss of generality, I analyze how the ask price (s) is determined when $v = +\sigma$.¹⁵

Given the length of a random speed bump (δ) and the timing of a jump in asset value (t_0), the fastest arrival time of a liquidity trader is $t = \delta$ due to a speed bump (i.e., the liquidity shock hits at

¹³My model is not an appropriate workplace to analyze symmetric speed bumps, such as those at IEX. It delays execution or arrival of all orders (both from takers and makers) with an exemption provided to a specific type of undisplayed pegged orders. As Appendix B explains in detail, the pegged orders cannot be analyzed by limit orders in my model.

¹⁴The *ex-ante* choice of speed is in line with the literature, e.g., Foucault et al. (2013) and Brolley and Zoican (2019).

¹⁵The symmetric argument gives the results for the bid-side of the market.

$t = 0$, and the order incurs a delay). By the same token, the fastest arrival time of the HFT is $t_0 + \delta$ (i.e., the HFT immediately obtains news on v at t_0 , while her order is delayed by δ).

Therefore, there are two possibilities regarding the timing of a trade: (i) $t \in [\delta, t_0 + \delta)$ or (ii) $t \in [t_0 + \delta, \infty)$. By incorporating them, the expected profit of the market maker from posting s is given by

$$V_M(s) = \mathbb{E}_{\delta, t_0} \left[s \int_{\delta}^{t_0 + \delta} \beta e^{-\beta(t-\delta)} \mathbb{I}_{\{t < t_0\}} dt + s \int_{\delta}^{t_0 + \delta} s \beta e^{-\beta(t-\delta)} e^{-\gamma(t-t_0)} \mathbb{I}_{\{t \geq t_0\}} dt + \int_{t_0 + \delta}^{\infty} (\beta s + \phi(s - \sigma)) e^{-\beta(t-\delta)} e^{-\phi(t-t_0-\delta)} e^{-\gamma(t-t_0)} dt \right], \quad (1)$$

where $\mathbb{E}_{\delta, t_0}$ denotes the expectation operator with respect to δ and t_0 , and $\mathbb{I}_X$ is the indicator function for event X .

The first two terms in (1) capture case (i), i.e., a trade takes place in $t \in [\delta, t_0 + \delta)$. In this case, the market maker knows that her trading partner is a liquidity trader. In other words, this time interval can be thought of as a “safe interval” generated by the speed bump, in which the market maker is free from the adverse selection cost.

Within this interval, a trade takes place with density $\beta e^{-\beta(t-\delta)}$ for $t < t_0$ and with density $\beta e^{-\beta(t-\delta)} e^{-\gamma(t-t_0)}$ for $t \geq t_0$, i.e., a liquidity shock hits at $t - \delta$, and a liquidity trader arrives at t before the market maker learns v . In this case, the market maker expects to earn $s - \mathbb{E}[v] = s$.

In contrast, the second line in (1) represents case (ii), i.e., a trade occurs in $t \geq t_0 + \delta$. In this case, both the HFT and liquidity traders may arrive before the market maker can revoke her limit order. Thus, it bears the adverse selection cost: the HFT buys an asset only if she learns true v and finds that a limit order is mispriced. The HFT can snipe the stale quote at time t if she becomes informed at $T_H = t - \delta$, while the market maker cannot cancel until t (i.e., $T_M > t$), and a liquidity trader arrives

after t (i.e., $T_L + \delta > t$). These conditions are satisfied with density

$$f_{HFT}(t) = \phi e^{-\phi(t-t_0-\delta)} e^{-\beta(t-\delta)} e^{-\gamma(t-t_0)},$$

and the market maker's expected profit from trading is $s - \sigma$. Similarly, the market maker may trade with a liquidity trader at t and obtains the expected profit s with density

$$f_{liq}(t) = \beta e^{-\phi(t-t_0-\delta)} e^{-\beta(t-\delta)} e^{-\gamma(t-t_0)}.$$

By incorporating the randomness in δ and t_0 , the overall expected profit from market-making is given by equation (1).

Since the market maker is competitive, her expected profit shrinks to zero, as in [Glosten and Milgrom \(1985\)](#). Thus, the break-even condition yields the following equilibrium spread.

Proposition 1. (i) With $\lambda(\varepsilon) \equiv \frac{\kappa(1+\varepsilon\gamma)+\varepsilon}{1+\kappa^{-1}\varepsilon}$, the equilibrium (half) spread is given by

$$s = \frac{\frac{\phi}{1+\lambda(\varepsilon)(\phi+\beta+\gamma)}}{\frac{\phi}{1+\lambda(\varepsilon)(\phi+\beta+\gamma)} + \beta} \sigma. \quad (2)$$

(ii) s is monotonically increasing in ϕ and decreasing in ε if ϕ is fixed.

(iii) In the equilibrium, the price impact of speed is decreasing in ε , i.e., $\frac{\partial}{\partial \varepsilon} \left(\frac{ds}{d\phi} \right) < 0$.

Proof. See Appendix A.1 for point (i). Point (ii) is obvious from (2). Point (iii) is derived by evaluating

$\frac{\partial}{\partial \varepsilon} \left(\frac{ds}{d\phi} \right)$ at the equilibrium speed in (8) below. $\square$

Note that $\lambda(\varepsilon)$ is monotonically increasing in the expected length of a delay, $\mathbb{E}[\delta] = \varepsilon$. In equation (2), a direct effect of a speed bump appears in the form of the discount on the arrival rate of the HFT, where the discount rate is $\lambda(\varepsilon)(\phi + \beta + \gamma)$. This is because a speed bump with $\varepsilon > 0$ generates the safe interval, and a higher ε increases the relative probability of trading within the safe interval. Thus, *ceteris paribus*, the market maker is less likely to be picked off and the spread shrinks, i.e.,

a speed bump mitigates adverse selection risk. However, in the equilibrium, the HFT's speed (ϕ) depends on ε through the HFT's speed acquisition, as analyzed below, and provides ε with another channel to affect s .

Point (iii) implies that a longer delay renders the spread less sensitive to the HFT's speed-up. A high ε protects the market maker from sniping by the HFT. Then, her expected profit, V_M , grows insensitive to ϕ . She behaves as if the arrival rate of the HFT is low and cares *less* about the HFT's speed-up. Consequently, even if the HFT becomes faster, the spread widens only slightly.

2.5 Speed acquisition by HFT

Consider the HFT's optimal speed acquisition problem to investigate how ϕ depends on ε .

Firstly, given δ and t_0 , the HFT can snipe the existing liquidity at time $t + \delta$ with the following probability density:

$$\pi_t(\phi, \delta) \equiv \Pr(T_H = t, T_L > t, T_M > t + \delta | \delta, t_0) = \phi e^{-\phi(t-t_0)} e^{-\beta t} e^{-\gamma(t-t_0+\delta)}. \quad (3)$$

Equation (3) involves the probability that the HFT privately observes v at t , while the arrival of a liquidity trader and the market maker's learning/cancellation do not occur until $t + \delta$.

Then, the expected (gross) trading profit of the HFT in the speed acquisition stage is

$$W_{HFT}^{gross}(\phi) \equiv \mathbb{E}_{\delta, t_0} \left[\int_{t_0}^{\infty} \pi_t(\phi, \delta) (\sigma - s) dt \right] = \pi(\phi, \varepsilon) (\sigma - s), \quad (4)$$

where $\pi(\phi, \varepsilon) \equiv \mathbb{E}_{\delta, t_0} \left[\int_{t_0}^{\infty} \pi_t(\phi, \delta) dt \right]$ denotes the expected probability that the HFT can snipe. The sniping probability is decomposed into two factors:

$$\pi(\phi, \varepsilon) = \underbrace{\frac{1}{1 + \beta\kappa} \frac{\phi}{\phi + \beta + \gamma}}_{\equiv \tilde{\pi}} \times \underbrace{\frac{1}{1 + \gamma\varepsilon}}_{\equiv h}. \quad (5)$$

The first component, denoted as $\tilde{\pi}(\phi)$, is the *potential sniping probability* when the HFT has speed ϕ .¹⁶ It represents the probability of sniping if no speed bumps are imposed, i.e., $\pi(\phi, \varepsilon = 0) = \tilde{\pi}(\phi)$. The second component, denoted as $h(\varepsilon)$, is the *discount factor* on $\tilde{\pi}$ generated by a speed bump. h is decreasing in ε , meaning that sniping becomes harder if a speed bump kicks in.

The HFT decides the optimal speed ϕ^* to maximize the following net profit.

$$W_{HFT}(\phi) \equiv W_{HFT}^{gross}(\phi) - C(\phi) = \tilde{\pi}(\phi)h(\varepsilon)(\sigma - s(\phi)) - C(\phi), \quad (6)$$

$$\text{s.t. } s = \frac{\frac{\phi}{1+\lambda(\varepsilon)(\phi+\beta+\gamma)}}{\beta + \frac{\phi}{1+\lambda(\varepsilon)(\phi+\beta+\gamma)}}\sigma.$$

In (6), $C(\phi)$ denotes the exogenous cost of speed with $C'(\phi) \geq 0$ (the functional form is specified below).

The FOC is given by

$$0 = \frac{dW_{HFT}(\phi)}{d\phi} = \left[(\sigma - s) \frac{d\tilde{\pi}}{d\phi} + \tilde{\pi} \frac{d(\sigma - s)}{d\phi} \right] h(\varepsilon) - C'(\phi). \quad (7)$$

Firstly, ϕ increases W_{HFT} because the HFT is more likely to snipe, i.e., $\frac{d\tilde{\pi}}{d\phi} > 0$ in (7). Secondly, besides the exogenous cost of speed C , a speed-up reduces the HFT's profit because trading cost s increases, i.e., $\frac{d(\sigma - s)}{d\phi} < 0$ in (7). Put differently, the equilibrium spread can be seen as an *endogenous* cost of being faster,¹⁷ and $\frac{ds}{d\phi} > 0$ constitutes the *endogenous* marginal cost of speed. For this reason, the problem is well-defined even if there is no exogenous cost or the cost is constant, i.e., $\frac{d^2W_{HFT}}{d\phi^2} = \frac{d^2W_{HFT}^{gross}}{d\phi^2} < 0$ when $C = 0$ or $C' = 0$.

¹⁶ $\beta\kappa$ affects $\tilde{\pi}$ because a more frequent arbitrage opportunity (a smaller κ) and a slow liquidity trader (a smaller β) make sniping easier for the HFT.

¹⁷See Internet Appendix E and Table 2 for a motivating fact. Also, Aquilina, Budish and O'Neill (2020) show that more than 80% of sniping races are played by top 6 high-frequency trading firms. Thus, incorporating the strategic motive of HFTs is natural.

2.6 Crowding-in effect

Now, I analyze the impact of a speed bump (ε) on speed acquisition (ϕ^*). To separate the key mechanism of the crowding-in effect, suppose that the exogenous cost is zero or constant for now. The next subsection retrieves the cost and shows that $C'(\cdot) > 0$ is the only channel through which a speed bump curbs speed acquisition. Thus, setting $C = 0$ (or $C' = 0$) is equivalent to muting the crowding-out effect.

2.6.1 Optimal speed

Proposition 2 (The crowding-in effect of a speed bump). *If there is no exogenous cost (or it is constant), the optimal speed is given by*

$$\phi^* = \begin{cases} \frac{\sqrt{\beta(\beta+\gamma)}(1+\lambda(\varepsilon)(\beta+\gamma))}{1-\lambda(\varepsilon)\sqrt{\beta(\beta+\gamma)}} & \text{if } \lambda < \sqrt{\beta(\beta+\gamma)}^{-1} \\ \infty & \text{otherwise.} \end{cases} \quad (8)$$

Moreover, ϕ^* is increasing in ε (when it is bounded).

Proof. Solving $W'_{HFT} = 0$ yields ϕ^* . The second statement is obvious, as $\lambda'(\varepsilon) \propto \kappa\gamma > 0$. $\square$

Proposition 2 demonstrates the *crowding-in* effect, i.e., a speed bump increases the equilibrium speed of the HFT. The following discussion focuses on the interior solution ($\phi^* < \infty$).

Intuition Consider the partial derivative of $\frac{dW_{HFT}}{d\phi}$ with respect to the length of a delay (ε) to understand the crowding-in effect. From (7),

$$\frac{\partial}{\partial \varepsilon} \left(\frac{dW_{HFT}}{d\phi} \right) \Big|_{\phi=\phi^*} = \left[\overbrace{\frac{\partial(\sigma-s)}{\partial \varepsilon} \frac{d\tilde{\pi}(\phi)}{d\phi}}^{\text{effect (i)} > 0} + \overbrace{\tilde{\pi}(\phi) \frac{\partial}{\partial \varepsilon} \left(\frac{d(\sigma-s)}{d\phi} \right)}^{\text{effect (ii)} > 0} \right] h(\varepsilon) + \frac{dW_{HFT}^{gross}}{d\phi} \frac{1}{h(\varepsilon)} \underbrace{\frac{dh(\varepsilon)}{d\varepsilon}}_{\text{effect (iii)} < 0}. \quad (9)$$

Increasing the expected length of a speed bump (ε) has two positive effects and one negative effect.

1. A speed bump exogenously mitigates adverse selection, reduces the spread, and increases the sniping profit $\sigma - s$ (Proposition 1). As a result, being faster and increasing the sniping probability become more worthwhile, as effect (i) in equation (9) shows.
2. The spread becomes less responsive to the speed-up by the HFT, as Proposition 1 attests. This reduces the marginal cost of being faster for the HFT. Effect (ii) in (9) captures this channel.
3. A speed bump makes it difficult for the HFT to snipe, and $h(\varepsilon)$ declines. It reduces the marginal benefit of being faster, as shown by effect (iii) in equation (9).

In the absence of the exogenous cost of speed, the FOC implies that

$$\frac{dW_{HFT}^{gross}(\phi^*)}{d\phi} = C'(\phi) = 0.$$

Therefore, effect (iii) disappears, which corresponds to the envelope theorem. As a result, the first two positive effects survive, and a longer expected delay (ε) stimulates the HFT to invest in the speed technology (ϕ). Thus, speed bumps can promote speed acquisition rather than thwarting it.

2.6.2 Adverse selection and liquidity

Hereafter, I use the equilibrium half spread s as a measure of adverse selection and market liquidity. A speed bump directly mitigates adverse selection by delaying the HFT's arrival (Proposition 1) but it promotes speed acquisition by the HFT (Proposition 2), thereby worsening adverse selection. In the equilibrium, these two competing effects cancel out each other.

Proposition 3. *If there is no exogenous costs or it is constant, the equilibrium spread is independent of the expected length of a delay, i.e., $\frac{ds}{d\varepsilon} = 0$.*

Proof. Substituting ϕ^* in (8) for ϕ in (2) yields the result. □

Although the crowding-in effect (Proposition 2) fuels speed acquisition, the strategic HFT knows

that a higher ϕ exacerbates adverse selection and reduces her trading profit. Since she is a monopolistic HFT facing no externalities, and there is no exogenous cost to adjust ϕ , she can freely choose the level of ϕ such that the liquidity-reducing impact of ϕ offsets the direct impact of a speed bump.¹⁸

The irrelevance theorem (Proposition 3) holds only in the benchmark model with $C = 0$ or $C' = 0$. Specifically, it rests on the environment with a single HFT, no exogenous costs, and no externalities. Thus, the following discussion with more general models overturns the result. Nevertheless, it works as the simplest building block of the subsequent discussions and helps us understand why somewhat counterintuitive results can arise.

2.7 Introducing crowding-out effect

As proposed by the literature, the negative impact of a speed bump on speed acquisition exists in my model. Once the exogenous costs C (with $C'(\phi) > 0$) is introduced, the FOC implies that

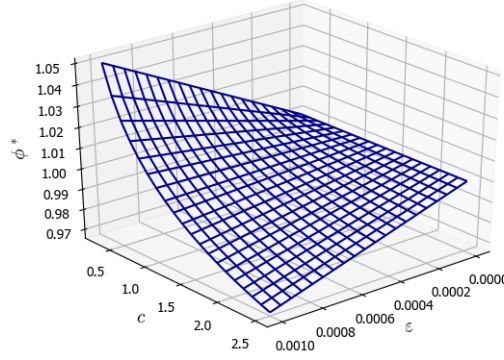
$$\frac{dW_{HFT}^{gross}(\phi^*)}{d\phi} = C'(\phi^*) > 0. \quad (10)$$

As a result, effect (iii) in equation (9) does not disappear and imposes a negative effect on speed acquisition. Intuitively, increasing the speed level involves the exogenous sunk cost C , but it pays out only if the HFT can snipe liquidity. Since a longer delay makes it difficult to snipe, it becomes less worthwhile to increase the speed. I call this channel the crowding-out effect of a speed bump.

To obtain some analytical results, I consider a quadratic cost function, $C(\phi) = c\phi^2$. With the crowding-out effect, the results in Propositions 2 and 3 are modified as follows:

¹⁸Proposition 3 resembles the result in Kyle (1985) that shows that an increase in the volatility of noise trading does not affect the market depth. In Kyle (1985), the monopolistic insider knows that more volatile noise trading reduces the price impact of the order flow (Kyle's λ), and she tries to exploit her informational advantage more intensively. At the same time, however, she knows more intensive informed trading increases the price impact. Thus, she picks the trading intensity so that more aggressive informed trading just offsets the impact of more volatile noise trading.

Figure 2: Reaction of the optimal speed to a delay with different c



Note: this figure illustrates the impact of ϵ on ϕ^* with various c . The level of $\phi^*(\epsilon, c)$ is normalized by $\phi^*(0, c)$ to mute the level effect of c and visualize the impact of ϵ . Other parameters are: $\beta = \gamma = \kappa = 1.0$.

Corollary 1. (i) There are $\underline{c}$ and $\bar{c}(> \underline{c})$, such that,

(i-a) If $c < \underline{c}$, then the crowding-in effect dominates the crowding-out effect, i.e., $\frac{d\phi^*}{d\epsilon} > 0$.

(i-b) If $c > \bar{c}$, then the crowding-out effect is dominant, i.e., $\frac{d\phi^*}{d\epsilon} < 0$.

(ii) The equilibrium spread is decreasing in the length of a speed bump, i.e., $\frac{ds}{d\epsilon} < 0$.

Proof. See Appendix F.1. □

Corollary 1 shows that whether the crowding-in effect dominates the crowding-out effect (negatively) depends on the exogenous cost, specified by parameter c . This result is intuitive because effect (iii) in equation (9) and equation (10) show that C is the driving force of the crowding-out effect.

Although it is difficult to derive cutoff c that separates the region into $\frac{d\phi^*}{d\epsilon} > 0$ and $\frac{d\phi^*}{d\epsilon} < 0$, Figure 2 plots the optimal speed ϕ^* and shows that the impact of c on $\frac{d\phi^*}{d\epsilon}$ is robust for the entire space of $c \geq 0$. That is, the reaction of ϕ^* to $\mathbb{E}[\delta] = \epsilon$ is smooth and switches only once.

Also, point (ii) is straightforward from Proposition 3. By the chain rule, we have

$$\frac{ds^*(\phi^*, \epsilon)}{d\epsilon} = \underbrace{\frac{\partial s^*}{\partial \epsilon}}_{<0} + \underbrace{\frac{d\phi^*}{d\epsilon} \frac{\partial s^*}{\partial \phi}}_{\geq 0}. \quad (11)$$

The first term is the negative direct impact of a speed bump (Point [ii] in Proposition 1), and the

second term is the impact of ε through HFT's speed acquisition, the sign of which is determined by the crowding-in effect versus the crowding-out effect. With $c = 0$, the second term is positive due to the absence of the crowding-out effect, and it completely offsets the negative first effect. Because the exogenous cost with $c > 0$ generates the crowding-out effect, it holds that

$$\frac{d\phi^*}{d\varepsilon}|_{c=0} > \frac{d\phi^*}{d\varepsilon}|_{c>0}.$$

Since the cost parameter c does not have any direct impacts on s , the first effect in (11) dominates the second effect via speed acquisition if $c > 0$.

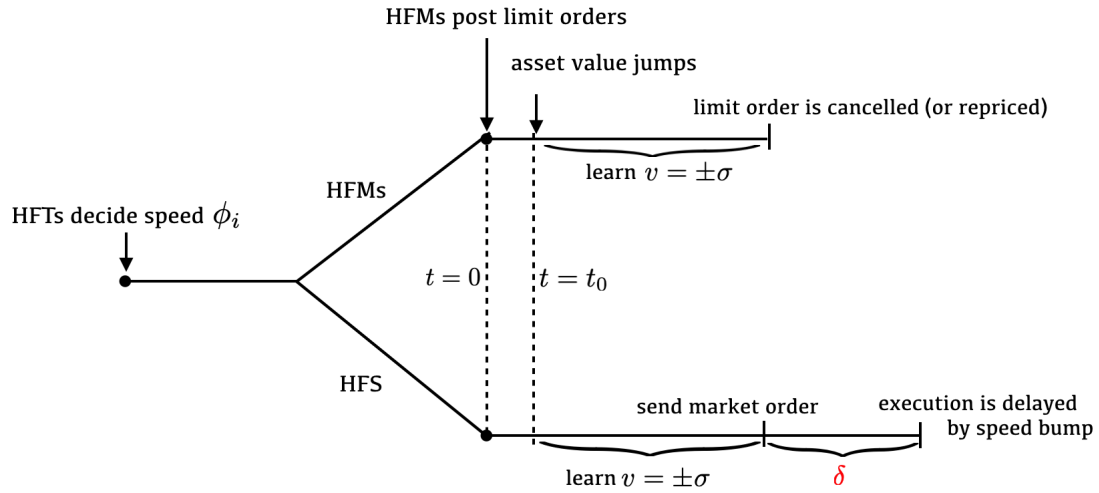
Therefore, the baseline model with a single HFT proposes the crowding-in effect but demonstrates that a speed bump is still effective to reduce the spread, as long as increasing ϕ takes some exogenous costs. However, the following sections with an arms race generalize this result, showing that a speed bump can backfire in terms of the spread in some parameter regions.

3 Multiple HFTs and high-frequency market making

In reality, HFTs serve not only as takers (snipers) but also as liquidity providers. I modify the benchmark model to capture this fact. To keep deviation from the benchmark model as minimal as possible, I adopt the following structure with $N = 3$ HFTs, indexed by $i = 1, 2, 3$. See Internet Appendix D for a model with general $N \geq 3$.

It is worth noting that the number of trading firms (N) that participate in a sniping race in reality tends to be small. For example, Aquilina, Budish and O'Neill (2020) show that more than 80% of sniping races are played by the top 6 high-frequency trading firms, meaning that a race is highly concentrated (see also Table 2 in Internet Appendix E).

Figure 3: Flow of events for HFTs



The flow of events for HFTs in the generalized model is summarized in Figure 3, and the structure in each step is explained in the following:

1. HFTs determine the speed level ϕ_i before the market opens.
2. At $t = 0$, one of the HFTs is randomly selected as a high-frequency sniper (HFS; $i = S$), while the other two HFTs become high-frequency market makers (HFMs). The HFMs post a single-unit limit order with (half) spread s_i . At this stage, they engage in *strategic* competition for liquidity provision à la [Dennert \(1993\)](#).
3. At $t = t_0 \sim \text{Exp}(\kappa^{-1})$, asset value v jumps to $v = v_0 \pm \sigma$ with equal probability.
4. With stochastic latency $T_i \sim \text{Exp}(\phi_i)$, HFT i learns the realized value of v . If HFT i is a sniper, she instantly places a market order to snipe *all* available limit orders (i.e., two units) provided by other HFTs $j (\neq i)$. If HFT i is an HFM, she revokes (or reprices) her limit order to avoid being picked off by an HFS. Also, at time $T_L \sim \text{Exp}(\beta)$, the liquidity shock hits a liquidity trader, and she posts a single unit buy or sell market order to fulfill her trading needs.

Note that the assumption on a unit trading size is relaxed for an HFS: she takes all available liquidity

on the LOB (i.e., two units) as long as they are profitable (see [Dennert, 1993](#); [Baruch and Glosten, 2019](#)). In contrast, a liquidity trader trades only one unit due to an exogenous capacity constraint. Other structures of the game, including a speed bump, stay the same. Once again, the problem is solved by backward induction.

3.1 Strategic liquidity provision by high-frequency market makers

Note that the market-making sector is no longer competitive. Suppose that HFT i and j become HFMs. This subsection analyzes HFM i 's decision on ask price s_i at the liquidity-provision stage.

HFM i knows that an HFS tries to snipe her limit order, as the sniper takes all available liquidity. In contrast, a liquidity trader takes HFM i 's limit order if it is on the top of the book. Thus, she can compensate for the loss from adverse selection only if she posts a better price than her rival (HFM j). In the case of a tie ($s_i = s_j$), I assume that HFM i trades with a liquidity trader with probability $m \in (0, 1)$.¹⁹

Since the structure of the trading game is the same as the benchmark model, HFM i 's expected profit from market-making, given her rival's quote s_j , is written as follows.

$$V_{M,i}(s_i) = \mathbb{E}_{\delta, t_0} \left[\beta \theta(s_i, s_j) s_i \left(\int_{\delta}^{t_0 + \delta} e^{-\beta(t-\delta)} \mathbb{I}_{t < t_0} dt + \int_{\delta}^{t_0 + \delta} e^{-\beta(t-\delta)} e^{-\phi_i(t-t_0)} \mathbb{I}_{t \geq t_0} dt \right) + \int_{t_0 + \delta}^{\infty} (\theta(s_i, s_j) \beta s_i + \phi_S(s - \sigma)) e^{-\beta(t-\delta)} e^{-\phi_S(t-t_0-\delta)} e^{-\phi_i(t-t_0)} dt \right] \quad (12)$$

$$\propto \frac{\phi_S}{1 + \lambda(\varepsilon)(\phi_i + \phi_S + \beta)} (s_i - \sigma) + \theta(s_i, s_j) \beta s_i \quad (13)$$

with

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

$\theta(s_i, s_j)$ represents the probability that HFM i can trade with a liquidity trader when the posted prices are s_i and s_j (conditional on the liquidity shock hits). Equation (13) indicates that the arrival rate of

¹⁹The tie-breaking rule, m , does not affect the following results.

the HFS (ϕ_s) is discounted by λ , as in the benchmark model, while the probability of trading with a liquidity trader is also discounted by θ due to competition for liquidity provision (e.g., [Dennert, 1993](#)).

Moreover, I introduce the following variable. For $k \in \{i, j\}$,

$$s_{k,0} \equiv \frac{\frac{\phi_s}{1+\lambda(\epsilon)(\phi_k+\phi_s+\beta)}}{\frac{\phi_s}{1+\lambda(\epsilon)(\phi_k+\phi_s+\beta)} + \beta} \sigma. \quad (14)$$

$s_{k,0}$ represents the break-even spread for HFM k when $\theta = 1$. If market makers engage in perfect competition, as in the benchmark model, the equilibrium spread is given by $(s_{i,0}, s_{j,0})$. However, HFTs in my model are strategic and know the impact of her rival's quote on her profits. As a result, the following result holds.

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

Proof. See discussions below. □

Figure 4: HFM i 's profits

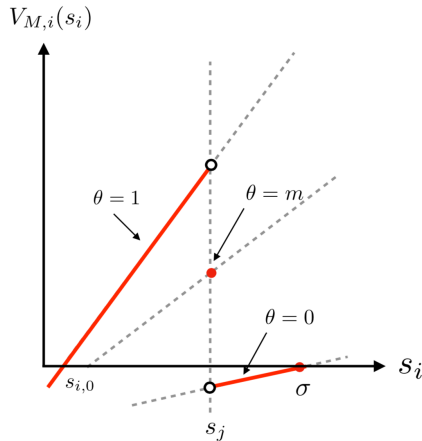
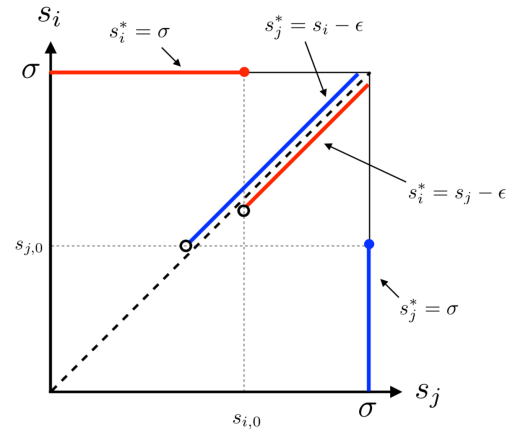


Figure 5: Best response and equilibrium



The result is analogous to [Dennert \(1993\)](#) and [Baruch and Glosten \(2013, 2019\)](#), and the mechanism is explained by Figures 4 and 5. Figure 4 draws HFM i 's profit from market-making, $V_{M,i}$, as a function of her strategy, s_i . Figure 5 shows the best-response function of HFM k ($k = i, j$), denoted as

s_k^* , to her rival's quote.²⁰

Figure 4 shows that it is optimal for HFM i to slightly undercut s_j as long as $s_j > s_{i,0}$ because a better price attracts a profitable order flow from a liquidity trader. Thus, in Figure 5, the best-response price of HFM i for $s_j > s_{i,0}$ is $s_i^* = s_j - \epsilon$ with $\epsilon \rightarrow +0$. Once s_j hits $s_{i,0}$, however, quoting $s_i \in [0, \sigma)$ generates negative profits. Since placing $s_i = \sigma$ always grants zero profit, the best response price jumps to $s_i^* = \sigma$. Symmetric arguments provide the best response of HFM j , denoted as s_j^* in Figure 5.

Figure 5 shows that price competition between strategic HFMs does not result in equilibrium in pure strategies. This is because HFMs comprehend how prices (s_i, s_j) affect their profit and try to exploit discontinuity at $s_i = s_j$.

Hence, I turn to mixed strategy equilibrium.

Proposition 4. *There is a unique mixed strategy equilibrium, in which HFM $k(= i, j)$ randomizes her quote s_k over the support $[s_{l,0}, \sigma]$, with $l \neq k$, according to the following cumulative distribution function.*

$$F_k(s) = \Pr(s_k \leq s) = 1 - \frac{1}{\beta} \frac{\phi_s}{1 + \lambda(\epsilon)(\phi_k + \phi_s + \beta)} \frac{\sigma - s}{s}, \quad (15)$$

where ϕ_s is the sniper's speed and $s_{l,0}$ is given by (14). Moreover, HFMs earn zero profits in the equilibrium.

Proof. Suppose that HFM i earns expected profit $V_{M,i}^*$ when HFM j randomizes quote s_j over $[\underline{s}_j, \sigma]$ with distribution F_j . The goal is to find three unknowns $(V_{M,i}^*, \underline{s}_j, F_j)$ that consist of a mixed strategy equilibrium. Firstly, θ in (13) is replaced by

$$\theta(s_i, F_j) = 1 - F_j(s_i) \quad (16)$$

²⁰The best response functions are not symmetric because it is possible that $\phi_i \neq \phi_j$ and $m \neq \frac{1}{2}$.

leading to

$$V_{M,i}^* \propto \frac{\phi_S}{1 + \lambda(\epsilon)(\phi_i + \phi_S + \beta)}(s_i - \sigma) + \underbrace{\beta(1 - F_j(s_i))}_{\text{Prob(trade w / liq.trader)}} s_i.$$

For the mixed strategy to be an equilibrium, $V_{M,i}^*$ must be independent of s_i . Since $s_i = \sigma$ is feasible, we have $V_{M,i}^* = 0$. By solving the above equation with $V_{M,i}^* = 0$ for F_j , we obtain (15). Finally, the lower bound of the distribution $\underline{s}_j$ is derived from $F_j(\underline{s}_j) = 0$, leading to $\underline{s}_j = s_{i,0}$. For uniqueness, see Appendix A.2. $\square$

The following expected spread works as an endogenous trading cost for liquidity takers.

Lemma 2. (i) With the break-even spread $s_{i,0}$ in (14), the expected spread posted by HFM i is given by

$$\mathbb{E}[s_i] = -\sigma \frac{1}{\beta} \frac{\phi_S}{1 + \lambda(\epsilon)(\phi_i + \phi_S + \beta)} \log s_{i,0}.$$

(ii) The expected spread is monotonically increasing and concave in ϕ_S and is decreasing in ϵ (with ϕ fixed).

Proof. Equation (15) leads to point (i). See Appendix A.3 for point (ii). $\square$

3.2 Speed acquisition

Consider HFT i 's speed acquisition problem before her role is determined. Since the indifference condition drives her expected profit from market-making zero, the *ex-ante* expected gains come only from sniping. Thus, she decides ϕ_i in the expectation that she becomes an HFS and the other two HFTs serve as HFMs. Since I focus on the symmetric equilibrium in the speed acquisition stage, HFT i solves the following problem anticipating that the other two HFTs take the symmetric speed level, i.e., $\phi_j = \phi_k$ for $j, k \neq i$.

The *ex-ante* expected trading profit of HFT i is given by

$$W_i^{\text{gross}} = \frac{1}{3} \sum_{l=j,k} \pi(\phi_i, \phi_l) \left(\sigma - \int_{s_{-l,0}}^{\sigma} s dF_l(s) \right) + \frac{2}{3} \mathbb{E}[V_{M,i}(s_i)]$$

where $-l \equiv \{j, k\} / \{l\}$. The first term represents the case that HFT i becomes an HFS, and the second term is the profits from market making. Note that $\mathbb{E}[V_{M,i}(s_i)] = 0$ from Proposition 4. Due to the symmetric speed of HFMs, $\phi_j = \phi_k$, the speed acquisition problem of HFT i is given by the following:

$$\max_{\phi_i} W_i = \overbrace{\frac{2}{3} \pi(\phi_i, \phi_j)(\sigma - \mathbb{E}[s_j]) - C(\phi_i)}^{=W_i^{gross}}, \quad (17)$$

$$\text{s.t., } \pi(\phi_i, \phi_j) = h(\varepsilon, \phi_j) \tilde{\pi}_i(\phi_i, \phi_j) \equiv \frac{1}{1 + \varepsilon \phi_j} \frac{\phi_i}{\phi_i + \phi_j + \beta} \frac{1}{1 + \beta \kappa}, \quad (18)$$

$$\mathbb{E}[s_j] = -\frac{1}{\beta} \frac{\phi_i}{1 + \lambda(\varepsilon)(\phi_i + \phi_j + \beta)} \log s_{j,0}.$$

Since the coefficient does not affect the result, I ignore $2/3$ on W^{gross} in the following.

The FOC is given by

$$0 = \frac{dW_i}{d\phi_i} = h(\varepsilon, \phi_j) \left(\frac{d\tilde{\pi}_i}{d\phi_i} (1 - \mathbb{E}[s_j]) - \tilde{\pi}_i \frac{d\mathbb{E}[s_j]}{d\phi_i} \right) - C'(\phi_i), \quad (19)$$

and it captures the marginal benefit and cost of being faster, where the cost emanates from the endogenous adverse price movement, $\frac{d\mathbb{E}[s_j]}{d\phi_i} > 0$, and the exogenous cost of speed, C' . I denote the solution of (19) as $\phi_i^* = BR_i(\phi_j)$ and call it the best response speed of HFT i to her rivals' (symmetric) speed ϕ_j .

3.3 Crowding-in effect

Once again, as the first step, I analyze the case with $C = 0$ (or $C' = 0$) to check the existence of the crowding-in effect.

3.3.1 Optimal speed

Although the explicit formula for $\phi_i^* = BR_i$ is hard to obtain, the following proposition analytically shows that the result in the benchmark model is robust. Moreover, it uncovers the strategic nature of

an arms race.

Proposition 5. *If the exogenous cost is absent or constant;*

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

Proof. See Appendix A.3. □

Intuition for point (i) of Proposition 5 is the same as the crowding-in effect in the benchmark model (i.e., equation [9]). As for point (ii), consider the cross-derivative of the FOC (19) with respect to the rival's speed ϕ_j .

$$\begin{aligned} \frac{\partial^2 W_i}{\partial \phi_j \partial \phi_i} = & \left[\overbrace{\frac{\partial(\sigma - \mathbb{E}[s_j])}{\partial \phi_j} \frac{\partial \tilde{\pi}_i}{\partial \phi_i} + \tilde{\pi}_i \frac{\partial^2(\sigma - \mathbb{E}[s_j])}{\partial \phi_j \partial \phi_i}}^{\text{effects (i) + (ii)} > 0} \right] h(\varepsilon, \phi_j) + \frac{\partial W_i^{\text{gross}}}{\partial \phi_i} \frac{1}{h} \overbrace{\frac{dh(\varepsilon, \phi_j)}{d\phi_j}}^{\text{effect (iii)} < 0} \\ & + \left[\overbrace{\frac{\partial \tilde{\pi}_i}{\partial \phi_j} \frac{\partial(\sigma - \mathbb{E}[s_j])}{\partial \phi_i}}^{\text{effect (iv)}} + \overbrace{\frac{\partial^2 \tilde{\pi}_i}{\partial \phi_j \partial \phi_i} (\sigma - \mathbb{E}[s_j])}^{\text{effect (v)}} \right] h(\varepsilon, \phi_j). \end{aligned} \quad (20)$$

As ϕ_j increases, HFM j becomes less likely to be picked off. It affects HFT i 's expected sniping profit through three channels: (a) by changing the behavior of the spread, (b) by lowering the discount factor $h(\varepsilon, \phi_j)$ of the sniping probability, (c) and by reducing the potential sniping probability $\tilde{\pi}_i(\phi_i, \phi_j)$. The third channel arises because, even without a speed bump (i.e., $h = 0$), faster HFT j makes it more difficult for HFT i to snipe.

The first two channels, (a) and (b), bring about the same effect as an extension of a speed bump, ε , because increasing ϕ_j reduces the sniping risk for HFMs. Hence, equation (20) has effects (i)-(iii) that replicate equation (9) in the benchmark model. The envelope condition eliminates effect (iii), as $\frac{\partial W_i^{\text{gross}}}{\partial \phi_i} = 0$. Furthermore, channel (c) generates effects (iv) and (v) in equation (20) with the following intuition.

4. Effect (iv) implies that HFT i does not need to care about the adverse price movement caused by her speed-up. Even if she becomes faster, and the trading profit worsens, her *expected* profit is less affected because it is less likely to materialize due to the speed-up by HFT j (ϕ_j).
5. Effect (v) captures the ambiguous effect of HFT j 's speed on the slope of $\tilde{\pi}_i$. When HFT j is equipped with very high speed ϕ_j , a speed-up does not really improve the sniping probability of HFT i because an increase in ϕ_i is dwarfed by a large ϕ_j . Hence, $\frac{\partial^2 \tilde{\pi}_i}{\partial \phi_j \partial \phi_i} < 0$ if ϕ_j is large, and vice versa.

Although the sign of effect (v) is ambiguous by itself, effects (iv) and (v) are positive in total.

Lemma 3. *In the equilibrium, the sum of effects (iv) and (v) is positive.*

Proof. Computing the sum by using the FOC in (19) yields the result. □

Intuitively, effect (v) cannot be a large negative to dominate effect (iv) in the symmetric equilibrium, where both HFTs have the same speed. Overall, effects (iv) and (v) help the first two effects with generating strategic complementarity.

3.3.2 Equilibrium speed

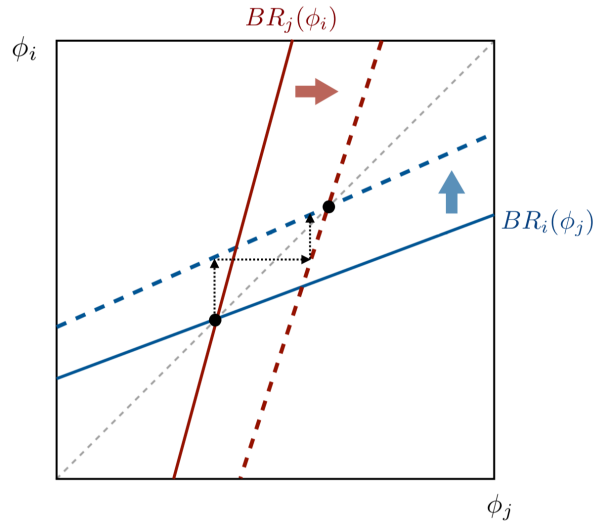
The solution of $\phi^* = BR(\phi^*)$ determines the symmetric equilibrium.

Proposition 6. *If $C = 0$ or $C' = 0$, the equilibrium speed ϕ^* is monotonically increasing in the expected length of a speed bump, ε .*

Proof. See Appendix A.4. □

Figure 6 illustrates the symmetric equilibrium and the impact of increasing ε . As in the benchmark model, the crowding-in effect shifts up BR_i . On top of that, the effect is amplified by the strategic complementarity, which is captured by the upward-sloping best response functions, generating a positive feedback loop.

Figure 6: Impact of a speed bump on BR



Note: This figure illustrates the best-response speed of HFTs. The dashed line in the middle represents the 45-degree line. The shifts from the solid lines to the upward-sloping dashed lines show the impact of increasing $\varepsilon = \mathbb{E}[\delta]$. The equilibrium evolves from the lower black dot to the upper black dot according to the dotted arrows.

3.3.3 Market quality

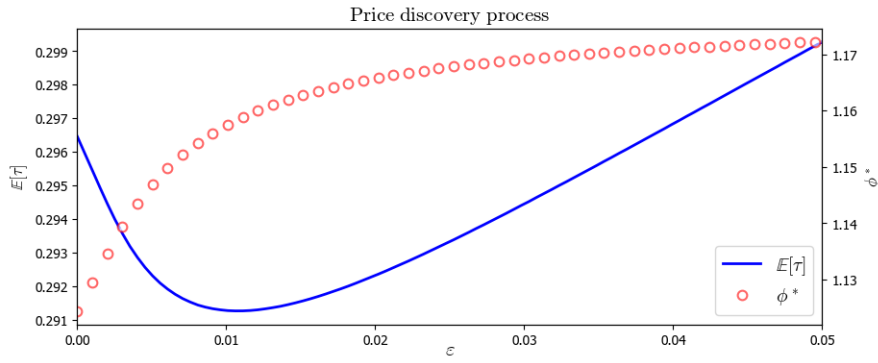
Adverse selection In the benchmark model with $C = 0$, the irrelevance result holds ($\frac{ds}{d\varepsilon} = 0$) because the HFT can exploit her monopolistic power to set ϕ^* . By contrast, an arms race between multiple HFTs generates positive externality that amplifies the crowding-in effect. Thus, the equilibrium speed, ϕ^* , becomes substantially high and dominates the direct impact of ε on the spread.

Proposition 7. *When $C = 0$ or $C' = 0$, a speed bump widens the expected spread, i.e., $\frac{d\mathbb{E}[s]}{d\varepsilon} > 0$.*

Proof. See Appendix A.4. □

The introduction of a speed bump or a longer expected delay can backfire, increasing the equilibrium speed of HFTs and worsening adverse selection. Since the primary purpose of a speed bump is to mitigate adverse selection, Proposition 7 indicates that it may lead to unfavorable consequences due to the crowding-in effect.

Figure 7: Expected time until price discovery



Note: This figure is illustrated by using $\beta = t_0 = 1.00$.

Price discovery How fast does the equilibrium price incorporate new information? For information to be reflected by the price, a stale quote must be removed from the market or repriced. This task is played by an informed market maker canceling her quote (which is equivalent to repricing in my model) or an informed high-frequency sniper taking a limit order.

The above event is triggered at stochastic time $\tau \equiv \min\{T^* + \delta, T^*, T^*\}$ with $T^* \sim \text{Exp}(\phi^*)$ where ϕ^* denotes the symmetric equilibrium speed. By exploiting the property of the exponential distribution, the expected length of time until price discovery from a jump in v is computed as²¹

$$\mathbb{E}[\tau] = \frac{1}{3\phi^*} \left(1 + \frac{\varepsilon\phi^*}{1 + 2\varepsilon\phi^*} \right). \quad (21)$$

Figure 7 plots $\mathbb{E}[\tau]$ against $\varepsilon = \mathbb{E}[\delta]$ in comparison with ϕ^* . Initially, a longer speed bump remarkably expedites the price discovery process, while it starts to slow down as ε becomes longer. On the one hand, a speed bump crowds in the speed level (ϕ^*). Therefore, the price can impound information more swiftly, leading to a lower $\mathbb{E}[\tau]$. On the other hand, a speed bump hampers the price discovery process because it directly restricts the sniper's trading behavior and undermines the crowding-in effect on speed acquisition (i.e., ϕ^* becomes concave). As a result, $\mathbb{E}[\tau]$ takes a U-shaped

²¹ τ is defined as the required time for the price to reflect v after the jump at t_0 . Thus, t_0 or κ does not affect the definition of τ and $\mathbb{E}[\tau]$.

curve.

3.4 Introducing the crowding-out effect

I retrieve the exogenous cost $C(\phi) = c\phi^2$ to study the competition between the crowding-in effect and the crowding-out effect in the multiple-HFTs model. The FOC implies that $\frac{dW_i^{gross}}{d\phi_i} = C'(\phi) > 0$. Thus, effect (iii) in equation (20) plays its role to curb speed acquisition.

As in the single-HFT model, the reaction of the arms race to a speed bump depends on the cost parameter, c .

Corollary 2. *There exist $\underline{c}$, $\bar{c}(> \underline{c})$, $\underline{\varepsilon}$ and $\bar{\varepsilon}(> \underline{\varepsilon})$ such that,*

- (i) *If $c < \underline{c}$ and $\varepsilon < \underline{\varepsilon}$, then the best response function of HFT i is increasing in ε and exhibits strategic complementarity: $\frac{dBR_i}{d\varepsilon} > 0$ and $\frac{dBR_i}{d\phi_j} > 0$.*
- (ii) *If $c > \bar{c}$ and $\varepsilon > \bar{\varepsilon}$, then the best response function of HFT i is decreasing in ε and exhibits strategic substitution: $\frac{dBR_i}{d\varepsilon} < 0$ and $\frac{dBR_i}{d\phi_j} < 0$.*

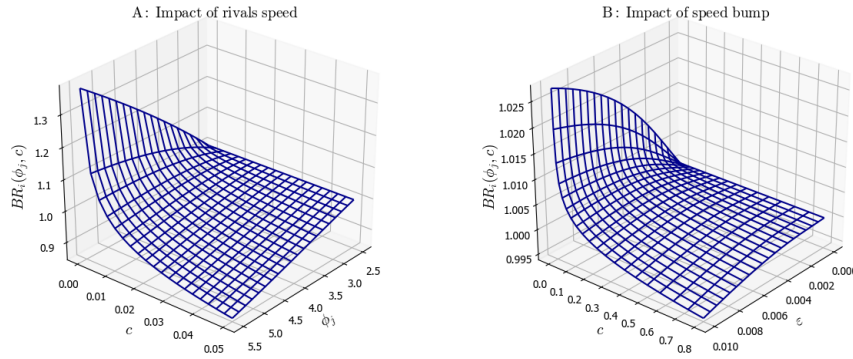
Proof. See Appendix F.1. □

Intuition is the same as that in the single-HFT model (Corollary 1). Namely, the crowding-out effect arises because both ϕ_j and ε make sniping more difficult. As the odds move against HFT i , the exogenous (sunk) cost C becomes more salient, reducing her investment in ϕ_i . As c increases, this effect becomes stronger and is more likely to dominate the crowding-in effect.

Figure 8 illustrates the best response function BR_i with different c to check the robustness of Corollary 2. Both figures plot *normalized* variables, e.g., Panel A plots $BR_i(\phi_j, c)/BR_i(0, c)$, to mute the level effect of c and visualize comparative statics regarding ε .

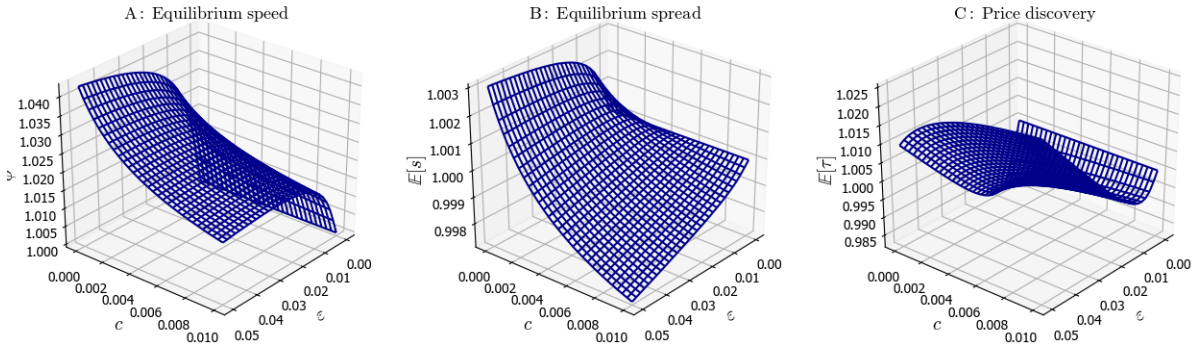
Panel A shows that an arms race involves strategic complementarity ($\frac{dBR_i}{d\phi_j} > 0$) when c is small, while strategic substitution arises as c increases. The best response function becomes hump-shaped if c is intermediate, as these two competing effects offset each other. Panel B establishes that analogous results hold for the reaction of BR_i to a speed bump, $\mathbb{E}[\delta] = \varepsilon$.

Figure 8: Best response function: $BR_i(\phi_j)$



Note: This figure is illustrated by using the following parameter value. $\beta = t_0 = 1.0$, $\varepsilon = 10^{-3}$ (Panel A), and $\phi_j = 1.0$ (Panel B). All values are normalized by $BR_i(\phi_j, c)|_{\varepsilon=0}$.

Figure 9: Equilibrium speed, spread, and price discovery



Note: This figure is illustrated by using the following parameter value: $\beta = t_0 = 1.0$. All variables are normalized by their value at $\varepsilon = 0$, e.g., Panel A plots $\phi^*(\varepsilon, c)/\phi^*(0, c)$.

Panels A, B, and C in Figure 9 show the impact of a speed bump $\mathbb{E}[\delta] = \varepsilon$ on the equilibrium speed ϕ^* , the expected spread $\mathbb{E}[s^*]$, and the price discovery process $\mathbb{E}[\tau]$ (defined by [21]).

The case with $c = 0$ follows the analytical result in Propositions 6 and 7, i.e., ϕ^* and $\mathbb{E}[s^*]$ are increasing in ε . If $c > 0$, a higher ε pushes up the *effective* exogenous cost, defined by the exogenous cost divided by the sniping probability, C/π . Thus, it tries to hamper the HFTs' incentive to be faster, leading to competition between the crowding-out effect and the crowding-in effect. As a result, with intermediate c , the reaction of the equilibrium speed becomes hump-shaped (Panel A).

The reaction of the expected spread $\mathbb{E}[s^*]$ to ε in Panel B strongly reflects the behavior of ϕ^* in Panel A. However, a higher ϕ^* does not necessarily promote $\mathbb{E}[s^*]$ because the negative direct

impact of ε on $\mathbb{E}[s^*]$ is also at play (see equation [11]). Thus, the spread becomes hump-shaped with intermediate c and monotonically decreasing in ε if c becomes sufficiently large. Finally, Panel C indicates that the U-shaped configuration of the price discovery process, measured by $\mathbb{E}[\tau]$, survives even if the exogenous cost is introduced.

Overall, considering the crowding-in effect with multiple HFTs generates the non-monotonic reaction of equilibrium variables to a speed bump due to the crowding-in effect versus the crowding-out effect. It leads to novel implications discussed in Section 5.

4 Welfare analysis

This section introduces *endogenous* liquidity traders to conduct welfare analyses.²² I also check the robustness of the results in the previous sections to endogenizing liquidity traders. In what follows, I consider the model with a single HFT, while Appendix A.5 analyzes the model with multiple HFTs.

4.1 Discretionary liquidity traders

Consider a unit mass of discretionary liquidity traders (DLTs). They replace the standard exogenous liquidity traders in the previous sections. A set of DLTs are exposed to a liquidity shock that happens at $T_L \sim \text{Exp}(\beta)$ with $T_L \geq 0$. I focus on the very short time interval such that the liquidity shock may happen at most once.

Given the occurrence of a shock, each DLT draws her type $\alpha \in [0, \bar{\alpha}]$ from the uniform distribution $U[0, \bar{\alpha}]$ with $\bar{\alpha} > 1$.²³ Type α represents the private value of asset holding or liquidity needs. Net utility from trading is independent of the asset value v but is price-dependent: a DLT with type α

²²See, for example, Admati and Pfleiderer (1988), Zhu (2014), and Han, Tang and Yang (2016).

²³The uniformity assumption is to avoid the slope effect of the distribution. As Proposition 8 attests, $\bar{\alpha} > 1$ is a sufficient condition for the existence of the equilibrium spread.

obtains

$$u(\alpha, s) = \begin{cases} \alpha\sigma - s & \text{if she buys at ask price } s \\ \alpha\sigma + (-s) & \text{if she sells at bid price } -s. \end{cases} \quad (22)$$

Note that σ in (22) is not the *realized* value of the asset but the volatility parameter, meaning that gains from trade are linear in the asset's volatility. Although this is for normalization and not essential for my results, it can stem from margin constraint or unmodeled risk aversion (see Brunnermeier and Pedersen, 2009 and Zhu, 2014). Since selling and buying the asset give the same utility, DLTs with $\mathbb{E}[u(\alpha, s)|\alpha, T_L] \geq 0$ (defined below) try to fulfill their trading needs by randomly placing a single unit buy or sell market order.

Execution price For this section, I assume that trading occurs twice: each of the HFT and a DLT trades once with the market maker. Except that, the behavior of the HFT and the competitive market maker below is the same as the benchmark model.

Regarding the execution price, there are two possibilities. First, if a liquidity taker (either a DLT or the HFT) could trade as the first trader, the execution price is $s_0 = s \in (0, \sigma)$, where s is the equilibrium spread determined in Subsection 4.2 below. In contrast, if a taker trades as the second trader (or she trades after the market maker learns v), the execution price becomes $s = s_1 = \sigma$. This is the equilibrium price for the second trade, because the market maker is posting the second limit order, knowing that a trade occurs twice only if the HFT trades either as the first or the second trader, which reveals that $v = \sigma$. See Internet Appendix A for a generalized model with n -rounds of trade.

Note that $s_1 = \sigma$ does not matter in the previous sections because the HFT earns zero profits even if she can snipe at this price. Thus, posting a new limit order at $s_1 = \sigma$ is equivalent to keeping the LOB empty after the market maker cancels the previous limit order. In contrast, I need to specify the new limit order at $s_1 = \sigma$ in this section because it matters for DLTs who derive utility from trading even at this price.

4.1.1 Behavior of discretionary liquidity traders

Consider the expected trading profits of type- α DLTs conditional on the liquidity shock at T_L . I denote the price of the asset on the LOB at time T_L as $\tilde{s}(T_L) \in \{s_0, s_1\}$. Note that $\tilde{s}$ is observable, and DLTs incorporate this price into their participation decision.

1. If $T_L < t_0$, then $\tilde{s} = s_0$, and a DLT anticipates two possibilities regarding the timing of order execution and its price. Namely, trading before or after the market maker learns v and reprices her quote from $s_0 = s$ to $s_1 = \sigma$. Note that the DLT does not fall behind the HFT because $T_H + \delta > t_0 + \delta > T_L + \delta$ for any t_0 and δ . The expected utility from placing an order is given by

$$\begin{aligned} \mathbb{E}[u(\alpha, s)|\alpha, T_L] &= (\alpha\sigma - s) \overbrace{\mathbb{E}_{t_0} \left[\int_0^{t_0 - T_L} \varepsilon^{-1} e^{-\varepsilon^{-1}\delta} d\delta + \int_{t_0 - T_L}^{\infty} \varepsilon^{-1} e^{-\varepsilon^{-1}\delta} e^{-\gamma(T_L + \delta - t_0)} d\delta | t_0 > T_L \right]}^{\text{Pr(trade before MM learns } v)} \\ &\quad + (\alpha\sigma - \sigma) \overbrace{\mathbb{E}_{t_0} \left[\int_{t_0 - T_L}^{\infty} \varepsilon^{-1} e^{-\varepsilon^{-1}\delta} (1 - e^{-\gamma(T_L + \delta - t_0)}) d\delta | t_0 > T_L \right]}^{\text{Pr(trade after MM learns } v)} \\ &= \sigma(\alpha - \bar{\alpha}\chi(s)) \end{aligned} \quad (23)$$

with

$$\chi(s) = \frac{1}{\bar{\alpha}} \left[\frac{s}{\sigma} \left(1 - \frac{\varepsilon\gamma}{1 + \varepsilon\gamma} \frac{\varepsilon}{\varepsilon + \kappa} \right) + \frac{\varepsilon\gamma}{1 + \varepsilon\gamma} \frac{\varepsilon}{\varepsilon + \kappa} \right].$$

Note that the resulting equation is independent of T_L due to the memoryless property of $\{T_L, t_0\}$ and their independence.

2. If $T_L \geq t_0$, and the DLT still observes $\tilde{s} = s_0$ on the LOB, there are two possibilities: (i) her order is executed at $s_0 = s \in (0, \sigma)$ and (ii) at $s_1 = \sigma$. Similar arguments to the above case (but

incorporating the fact that the price is not updated at T_L) yield the following utility.

$$\begin{aligned}\mathbb{E}[u(\alpha, s) | \alpha, T_L > t_0, \tilde{s} = s_0] &= (\alpha\sigma - s) \overbrace{\int_0^\infty \varepsilon^{-1} e^{-\varepsilon^{-1}\delta} e^{-\gamma\delta} d\delta \frac{(\varepsilon^{-1} - \phi)e^{-\phi(T_L - t_0)}}{\varepsilon^{-1}e^{-\phi(T_L - t_0)} - \phi e^{-\varepsilon^{-1}(T_L - t_0)}}}^{\text{Pr(trade at } s_0 = s)} \\ &\quad + (\alpha\sigma - \sigma) \overbrace{\int_0^\infty \varepsilon^{-1} e^{-\varepsilon^{-1}\delta} \left(1 - e^{-\gamma\delta} \frac{(\varepsilon^{-1} - \phi)e^{-\phi(T_L - t_0)}}{\varepsilon^{-1}e^{-\phi(T_L - t_0)} - \phi e^{-\varepsilon^{-1}(T_L - t_0)}} \right) d\delta}^{\text{Pr(trade at } s_1 = \sigma)} \\ &= \sigma [\alpha - \bar{\alpha}\rho(s, T_L - t_0)],\end{aligned}\quad (24)$$

with

$$\rho(s, \tau) = \frac{1}{\bar{\alpha}} \left[\frac{s}{\sigma} \frac{1}{1 + \gamma\varepsilon} \frac{(1 - \varepsilon\phi)e^{-\phi\tau}}{e^{-\phi\tau} - \varepsilon\phi e^{-\varepsilon^{-1}\tau}} + 1 - \frac{(1 - \varepsilon\phi)e^{-\phi\tau}}{e^{-\phi\tau} - \varepsilon\phi e^{-\varepsilon^{-1}\tau}} \right].$$

The first and the second lines in (24) incorporate the probability that the DLT can trade prior to the HFT and the market maker, conditional on they have not arrived yet at T_L .

3. If $T_L > t_0$, and the price on the LOB is $\tilde{s} = s_1 = \sigma$, the DLT's expected utility is

$$\mathbb{E}[u(\alpha, s) | \alpha, T_L > t_0, \tilde{s} = s_1] = \sigma(\alpha - 1).$$

Therefore, the conditional expected utility is summarized as

$$\mathbb{E}[u(\alpha, s) | \alpha, T_L, \tilde{s}] = \begin{cases} \sigma [\alpha - \bar{\alpha}\chi(s)] & \text{if } T_L < t_0 \\ \sigma [\alpha - \bar{\alpha}\rho(s, T_L - t_0)] & \text{if } T_L \geq t_0 \text{ and } \tilde{s} = s_0 \\ \sigma(\alpha - 1) & \text{if } T_L \geq t_0 \text{ and } \tilde{s} = s_1. \end{cases}$$

In any cases, a DLT can fulfill her trading needs and obtain private value $\sigma\alpha$ if she places an order. However, the execution price varies depending on the timing of the liquidity shock, T_L , and the speed bump, δ . Hence, χ and ρ represent the expected execution price.

From the market maker's perspective, the aggregate measure of *active* DLTs matters to determine

$s_0 = s$, as it determines the size of liquidity trading.

Lemma 4. *If the DLTs arrive at stochastic time $T_L + \delta$ before the market maker updates her quote, they have the following measure*

$$w\left(\frac{s}{\sigma}, T_L\right) = \begin{cases} 1 - \chi(s) & \text{if } T_L < t_0, \\ 1 - \rho(s, T_L) & \text{if } T_L \geq t_0. \end{cases} \quad (25)$$

(25) is derived by $\mathbb{E}[u|\alpha, T_L] > 0$. It characterizes the price-elastic behavior of DLTs: the wider the spread, the less active the DLTs become.

4.2 Equilibrium spread

The market maker's expected profit from posting $s_0 = s$ is analogous to the benchmark model (1), but she takes into account the measure of liquidity trading in Lemma 4.

$$\begin{aligned} V_M(s) = & \mathbb{E}_{\delta, t_0} \left[s\beta(1 - \chi(s)) \left(\int_{\delta}^{t_0+\delta} e^{-\beta(t-\delta)} \mathbb{I}_{\{t < t_0\}} dt + \int_{\delta}^{t_0+\delta} e^{-\beta(t-\delta)} e^{-\gamma(t-t_0)} \mathbb{I}_{\{t \geq t_0\}} dt \right) \right. \\ & \left. + \int_{t_0+\delta}^{\infty} e^{-\beta(t-\delta) - \phi(t-t_0-\delta) - \gamma(t-t_0)} [s\beta(1 - \rho(s, t - \delta)) + (s - \sigma)\phi] dt \right]. \end{aligned}$$

The break-even condition leads to the following result.

Proposition 8. *The equilibrium spread, s^* , must satisfy the following equation:*

$$\frac{s^*}{\sigma} = \frac{\frac{\phi}{1+\lambda(\varepsilon)(\phi+\beta+\gamma)}}{\beta(1 - m(\frac{s^*}{\sigma}, \varepsilon)) + \frac{\phi}{1+\lambda(\varepsilon)(\phi+\beta+\gamma)}} \equiv G\left(\frac{s^*}{\sigma}, \phi\right) \quad (26)$$

with $\lambda(\varepsilon) \equiv \frac{\kappa(1+\varepsilon\gamma)+\varepsilon}{1+\kappa^{-1}\varepsilon}$ and

$$m\left(\frac{s}{\sigma}, \varepsilon\right) = \frac{\phi + \beta + \gamma}{1 + \lambda(\phi + \beta + \gamma)} \frac{1}{\bar{\alpha}} \left[\frac{s}{\sigma} \left(\frac{\lambda^2}{\kappa(1 + \varepsilon\gamma)} + \frac{p(\varepsilon)}{\psi} \right) + \lambda \left(1 - \frac{\lambda}{\kappa(1 + \varepsilon\gamma)} \right) + \frac{1}{\psi} [1 - p(\varepsilon)] \right]$$

where $p(\varepsilon) = \mathbb{E}_z \left[\frac{1}{1+\gamma\varepsilon} \frac{(1-\varepsilon\phi)e^{-\phi z}}{e^{-\phi z} - \varepsilon\phi e^{-\varepsilon^{-1}z}} \right]$ with $z \sim \text{Exp}(\phi + \beta + \gamma)$.

If and only if $\bar{\alpha} > 1$, the fixed point problem (26) has a unique stable solution s^* such that $\frac{s^*}{\sigma} \in (0, 1)$. The explicit formula for s^* is given by (39) in Appendix A.5.

Proof. See Appendix A.1 for the derivation. $\square$

As in the benchmark model, the arrival rate of the HFT in (26) is discounted by $\lambda(\varepsilon)(\phi + \beta + \gamma)$, where λ is increasing in ε . On top of that, the measure of liquidity trading is also reduced by $m(\frac{s}{\sigma})$, which represents the expected measure of *inactive* liquidity traders.

Due to the price-elastic behavior of the DLTs, s^* is determined by the fixed point problem in equation (26). Importantly, G is increasing in s , representing the “liquidity-begets-liquidity” argument. Namely, a narrow spread reduces the trading cost and attracts a large set of discretionary liquidity traders. In turn, more active liquidity trading mitigates adverse selection for the market maker and reduces the spread even more. In what follows, I refer to this externality as “liquidity feedback.”

4.3 Speed acquisition

Note that the arrival rate of the DLTs (β) and the HFT’s sniping probability are the same as the benchmark model (while the measure of liquidity trading is different). The optimal speed, ϕ^* , is chosen to maximize

$$W_{HFT} = \underbrace{\tilde{\pi}(\phi)h(\varepsilon)(\sigma - s^*)}_{\equiv W_{HFT}^{gross}} - C(\phi),$$

$$\text{s.t., } \frac{s^*}{\sigma} = G\left(\frac{s^*}{\sigma}\right),$$

where G is given by (26), $\tilde{\pi} = \frac{\phi}{\phi + \beta + \gamma} \frac{1}{1 + \kappa\beta}$, and $h = \frac{1}{1 + \varepsilon\gamma}$.

The following equation represents the impact of being faster on W_{HFT} , which captures the marginal

benefit versus the marginal cost of being faster.

$$\frac{dW_{HFT}}{d\phi} = \left[(\sigma - s^*) \frac{\partial \tilde{\pi}}{\partial \phi} - \tilde{\pi} \frac{ds^*}{d\phi} \right] h(\varepsilon) - C'(\phi) = \left[(\sigma - s^*) \frac{\partial \tilde{\pi}}{\partial \phi} - \tilde{\pi} \underbrace{\frac{\frac{\partial G(\frac{s^*}{\sigma})}{\partial \phi}}{1 - \underbrace{\frac{dm(\frac{s^*}{\sigma})}{ds} \frac{\partial G(\frac{s^*}{\sigma})}{\partial m}}_{\text{amplification} < 1}}}_{\text{price impact of speed}} \right] h(\varepsilon) - C'(\phi). \quad (27)$$

As shown by the second term in the brackets, the endogenous marginal cost of being faster (i.e., the price impact of speed) is amplified due to the liquidity feedback effect. This is because an increase in ϕ^* pushes up s^* by its direct effect and by reducing the measure of DLTs.

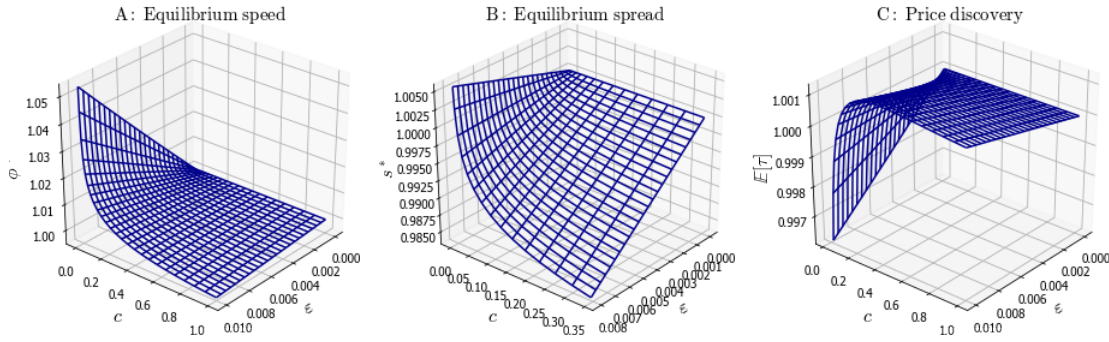
To see the impact of a speed bump on speed acquisition, consider the cross-derivative of (27).

$$\frac{\partial}{\partial \varepsilon} \left(\frac{dW_{HFT}}{d\phi} \right) = h(\varepsilon) \left[\underbrace{\frac{\frac{\partial s^*}{\partial \varepsilon} \frac{\partial \tilde{\pi}}{\partial \phi} - \tilde{\pi} \frac{\partial}{\partial \varepsilon} \frac{ds^*}{d\phi}}{1 - \underbrace{\frac{dm(\frac{s^*}{\sigma})}{ds} \frac{\partial s^*}{\partial m}}_{\text{amplification}}}}_{\text{original effects (i) and (ii) } > 0} - \tilde{\pi} \frac{\partial s^*}{\partial \phi} \frac{\partial}{\partial \varepsilon} \underbrace{\left(1 - \frac{dm(\frac{s^*}{\sigma})}{ds} \frac{\partial s^*}{\partial m} \right)^{-1}}_{=\Delta \text{amplification}} \right] + \frac{dW_{HFT}^{gross}}{d\phi} \frac{1}{h(\varepsilon)} \underbrace{\frac{dh(\varepsilon)}{d\varepsilon}}_{\text{effect (iii)}}. \quad (28)$$

The “original effects” in the first term captures the impact of a speed bump inherited from the benchmark model, i.e., positive effects (i) and (ii) in equation (9). Namely, a speed bump affects the level and the slope of the spread so that it facilitates speed acquisition. The denominator implies that liquidity feedback amplifies the benchmark effects because both effects (i) and (ii) cause a change in the spread. The last term (effect [iii]) is also inherited from the benchmark argument in equation (9), i.e., the crowding-out effect of a speed bump. Since the crowding-out effect stems from the exogenous cost, it is unaffected by liquidity feedback.

The second term, denoted as “ Δ amplification”, represents how a speed bump affects the liquidity feedback loop and the amplification term in equation (27). Note that $m(\frac{s}{\sigma})$ represents the measure of inactive DLTs. Firstly, m becomes less responsive to s^* when a speed bump is long. With a long delay, each DLT is more likely to trade at $s_1 = \sigma$ rather than $s_0 = s^*$ (i.e., after the market maker learns v).

Figure 10: Equilibrium variables



Note: Figures in this section are illustrated by using the following parameter values. $\beta = \gamma = 1.00, \bar{\alpha} = 1.40, t_0 = 1.00$.

Thus, a change in s^* matters less to DLTs' participation decision.

Secondly, a longer speed bump changes the impact of m on the spread. A speed bump reduces the effective arrival rate of the HFT, $\frac{\phi}{1+\lambda(\phi+\beta+\gamma)}$, meaning that the market maker's profit relies more heavily on trading with DLTs. Thus, an increase in ε tends to make $\frac{\partial s^*}{\partial m}$ larger. Due to these competing effects, the change in the amplification term is ambiguous.

4.4 Equilibrium impact of a speed bump

As the fixed point problem in (26) is analytically intractable, further comparative statics rely on numerical methods with Python function `fsolve`. As in any numerical exercise, the following analyses cannot guarantee that the results below apply to all parameter values. However, it reveals some interesting and non-trivial properties of a speed bump and speed acquisition.

Figure 10 shows the behavior of the optimal speed (Panel A), the equilibrium spread (Panel B), and the price discovery process (Panel C). All panels plot *normalized* values of equilibrium variables to mute the level effect of c and visualize the impact of ε , e.g., for variable x , figures show $x(\varepsilon, c)/x(0, c)$.

Panel A checks the robustness of the benchmark results. The optimal speed of the HFT is increasing in $\varepsilon = \mathbb{E}[\delta]$ when c is small, while it tends to decline in high- c or high- ε regions because both c

and ε make the exogenous cost of speed more salient and strengthen the crowding-out effect. Thus, the result is consistent with the previous sections with exogenous liquidity traders.

Panel B shows that the previous sections' results regarding s^* are also robust, except that it is increasing in $\varepsilon = \mathbb{E}[\delta]$ even if $c = 0$ (i.e., the irrelevance result in Proposition 3 does not hold). This is because the DLTs add the new channel through which ϕ positively affects the spread: a faster HFT reduces DLTs' expected utility and hampers liquidity trading, aggravating adverse selection.

Panel C illustrates the measure of price discovery. In the single-HFT model, it takes $\tau \equiv \min\{T_M, T_H + \delta\}$ for the price to impound information about v . We can compute that

$$\mathbb{E}[\tau] = \frac{\varepsilon}{1 - \varepsilon\phi} \left(\frac{1}{\varepsilon(\phi + \gamma)} - \frac{\varepsilon\phi}{1 + \varepsilon\gamma} \right).$$

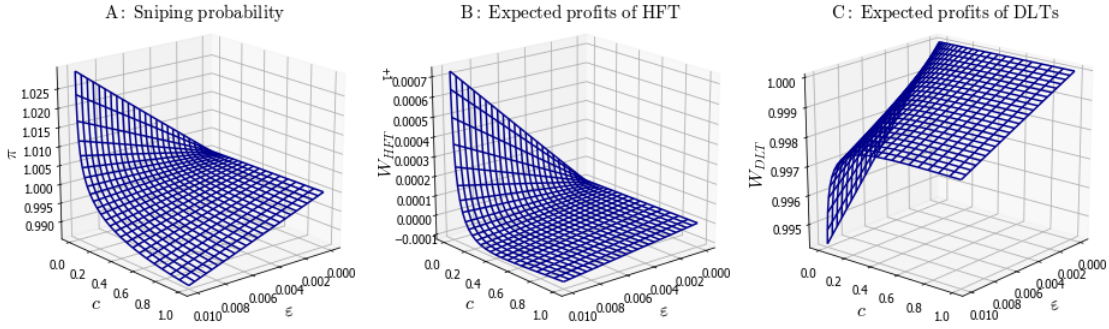
Although information revelation by the HFT's sniping is delayed by δ , the crowding-in effect of a speed bump expedites the price discovery process as long as c is small. The opposite result holds when c is large and the crowding-out effect becomes dominant.

Overall, the comparative statics with endogenous liquidity traders provide the same results as those with exogenous liquidity traders, such as Corollary 1, except for the special case with a single HFT and no exogenous costs.

4.5 Welfare: who benefits from a speed bump?

This subsection investigates the welfare implication of a speed bump. It is worth noting that the market maker earns zero profits regardless of a speed bump, i.e., her *equilibrium* profit is unaffected by δ or $\varepsilon = \mathbb{E}[\delta]$. Thus, the purpose of a speed bump is not to increase the market maker's *equilibrium* profit but to mitigate the asymmetric information problem, as it determines the spread and trading costs for other traders (Glosten and Milgrom, 1985).

Figure 11: Welfare impact of a speed bump



Note: The figure uses the same parameter values and normalization as in Figure 10.

HFT's expected trading profit The HFT's welfare is defined by her *ex-ante* expected profit (at $\phi = \phi^*$ and $s = s^*$):

$$W_{HFT} = \underbrace{\frac{1}{1 + \kappa\beta\phi^* + \beta + \gamma} \frac{\phi^*}{1 + \varepsilon\gamma}}_{=\pi(\phi, \varepsilon)} (\sigma - s^*) - C(\phi^*),$$

which is identical to the benchmark model, except that s^* is the solution of the fixed point problem in (26).

Figure 11 shows the comparative statics. Firstly, the reaction of the HFT's sniping probability π (Panel A) is positive when the exogenous cost (c) and the length of a delay (ε) are small. It comes from the crowding-in effect of a speed bump. In high- c or high- ε regions, however, the sniping probability becomes hump-shaped or declines, as the crowding-out effect weakens the reaction of ϕ^* to ε , and the negative direct effect of ε via $h(\varepsilon)$ tends to dominate. Panel B illustrates the expected trading profit of the HFT, W_{HFT} . The reaction of W_{HFT} to a speed bump strongly reflects the behavior of π but less responsive than π . This is because the sniping profit, $\sigma - s$, is negatively correlated with π via adverse selection for the market maker (e.g., when the HFT is more likely to snipe, the trading cost tends to be high).

DLTs' expected trading surplus For $\tau \geq t_0$, define $q(\tau, t_0)$ as the probability that $\tilde{s}(\tau) = s_0 (= s^*)$. By using the property of the exponential distribution,

$$q(\tau, t_0) = e^{-\gamma(\tau-t_0)} \frac{\varepsilon^{-1} e^{-\phi(\tau-t_0)} - \phi e^{-\varepsilon^{-1}(\tau-t_0)}}{\varepsilon^{-1} - \phi}.$$

Also, I denote $\bar{q} = 1 - q$.

A DLT's *ex-ante* expected utility (before a shock hits and α is realized) is formulated as follows.

$$W_{DLT} = \mathbb{E}_{\alpha, t_0} \left[\int_0^{t_0} \beta e^{-\beta\tau} \sigma [\alpha - \bar{\alpha} \chi(s^*)]^+ d\tau \right. \\ \left. + \int_{t_0}^{\infty} \beta e^{-\beta\tau} q(\tau, t_0) \sigma [\alpha - \bar{\alpha} \rho(s^*, \tau, t_0)]^+ d\tau + \int_{t_0}^{\infty} \beta e^{-\beta\tau} \bar{q}(\tau, t_0) \sigma (\alpha - 1)^+ d\tau \right] \quad (29)$$

with $x^+ \equiv \max\{0, x\}$. The first line represents the case that the shock hits before t_0 (i.e., $T_L < t_0$), while the second line shows the complimentary case. With probability q , the DLT can trade at expected price ρ , while it becomes $s = \sigma$ with probability $\bar{q}$. Also, the trading option is exercised only if a liquidity trader's type is sufficiently high. Since q depends on τ and t_0 only via their difference, $\tau - t_0$, equation (29) reduces to

$$W_{DLT} = \frac{\sigma \bar{\alpha}}{1 + \kappa \beta} \left[\beta [1 - \chi(s^*)]^2 + \kappa^{-1} \int_0^{\infty} \beta e^{-\beta u} q(u) [1 - \rho(s^*, u)]^2 du + \left(1 - \frac{1}{\bar{\alpha}}\right)^2 \kappa^{-1} \int_0^{\infty} \beta e^{-\beta u} \bar{q}(u) du \right],$$

where $q(u) \equiv q(u, 0)$.

Proposition 9. *The direct effect of a speed bump on DLTs' expected utility is negative, i.e., $\frac{\partial W_{DLT}}{\partial \varepsilon} < 0$.*

Proof. It is confirmed by taking the first order (partial) derivative of χ and $q(1 - \rho)^2$ with respect to ε . □

Besides the effect through the spread, a speed bump harms W_{DLT} . Due to a speed bump, DLTs are more likely to fall behind the market maker's learning. The DLTs' market order is more likely to experience "slippage" as a speed bump becomes longer. If there is no speed bump, DLTs hit by a

liquidity shock in $t < t_0$ can execute orders at price $s_0 = s^* < \sigma$. However, if a speed bump kicks in, DLTs placing orders during $t < t_0$ anticipate that their orders are executed at unfavorable prices, $s_1 = \sigma$, compared to what they currently have on the LOB (s^*). Although a speed bump may or may not reduce the spread, Panel C in Figure 11 demonstrates that the direct effect in Proposition 9 is dominant, weighing negatively on the expected utility of DLTs.

Overall, a speed bump renders taking liquidity more difficult and costlier both for the HFTs and the DLTs. The HFTs can compensate for the loss by increasing their speed, but liquidity traders do not have any tools to offset the negative impact of a speed bump. Therefore, by comparing the HFT and the DLTs in Panels B and C of Figure 11, the difference in the expected welfare of these two types of traders, $W_{HFT} - W_{DLT}$, widens as a speed bump becomes longer. The lack of speed-adjustment tools for DLTs leads to a pessimistic prediction that inequality between HFTs and ordinary passive traders widens due to a speed bump.

5 Discussion

5.1 Deterministic speed bump

In reality, asymmetric speed bumps in operation are all *random*, while several exchanges have proposed asymmetric and *deterministic* speed bumps. A deterministic delay can be analyzed as a special case of the benchmark model, although the results are somewhat trivial.

As in the benchmark model, a speed bump delays all market orders by δ but it is constant $\delta = \bar{\delta} \geq 0$. Following the discussion in the benchmark model, the market maker posting s_{det} at $t = 0$

anticipates to obtain²⁴

$$V_M(s_{det}) = \mathbb{E}_{t_0} \left[\left(\int_{\bar{\delta}}^{t_0+\bar{\delta}} \beta e^{-\beta(\tau-\bar{\delta})} \mathbb{I}_{\tau < t_0} d\tau + \int_{\bar{\delta}}^{t_0+\bar{\delta}} \beta e^{-\beta(\tau-\bar{\delta})} e^{-\gamma(\tau-t_0)} \mathbb{I}_{\tau \geq t_0} d\tau \right) s_{det} + e^{-\gamma\bar{\delta}-\beta t_0} \frac{1}{\phi + \beta + \gamma} [\phi(s_{det} - \sigma) + \beta s_{det}] \right].$$

The first term shows the case that a trade happens in the safe interval before the market maker can cancel her quote, while the second term represents a trade after the safe interval. The break-even spread is

$$s_{det} = \frac{\frac{\phi}{1+\lambda(\bar{\delta})(\phi+\beta+\gamma)}}{\frac{\phi}{1+\lambda(\bar{\delta})(\phi+\beta+\gamma)} + \beta} \sigma$$

with

$$\lambda(\bar{\delta}) = \frac{\kappa}{\kappa^{-1} - \gamma} \left(\kappa^{-1} - \gamma e^{-(\kappa^{-1}-\gamma)\bar{\delta}} \right). \quad (30)$$

Note that λ is monotonically increasing in $\bar{\delta}$. Moreover, the objective function of the HFT is

$$\phi^* = \arg \max_{\phi} e^{-\gamma\bar{\delta}} \frac{1}{1 + \kappa\beta} \frac{\phi}{\phi + \beta + \gamma} (\sigma - s_{det}) - C(\phi).$$

Thus, denoting $h(\bar{\delta}) = e^{-\gamma\bar{\delta}}$ and using λ in (30) make the model identical to the benchmark model.

The identity confirms the robustness of my main arguments in the previous sections.

As discussed in Appendix B and analyzed by Baldauf and Mollner (2017), it is believed that a random speed bump has an additional benefit in the recent segmented markets. This is because a random delay makes it harder to synchronize the arrival of “sprayed” orders across multiple exchanges. Although comprehensive analyses that impound all these elements are beyond the scope of this paper, the discussions in my model can be a building block for future research on the optimal design of speed bumps in multiple exchanges.

²⁴Note that they do not update continuously because the process for t_0 is memoryless.

5.2 Will the market introduce a speed bump?

“Will the market fix the market?” is the question put forth by Budish, Lee and Shim (2018) in the context of frequent batch auctions (FBA). They analyze whether competing exchanges have an incentive to introduce an FBA that aims to mitigate adverse selection caused by HFTs. Although adopting multiple exchanges and competition between them is beyond the scope of my model, incorporating the crowding-in effect of a speed bump into the above question can provide a new implication.

Specifically, the following claim is derived from my model:

Claim 1. With the appropriate design of a speed bump (value of $\varepsilon = \mathbb{E}[\delta]$) and the schedule of a speed fee (c), a speed bump promotes speed acquisition by HFTs and mitigates the adverse selection problem at the same time.

One of the examples comes from Corollary 1 in the single-HFT model that shows

$$\exists(\varepsilon, c) \text{ s.t., } \frac{d\phi^*}{d\varepsilon} > 0 \text{ and } \frac{ds^*}{d\varepsilon} < 0.$$

There is a region of ε and c such that the crowding-in effect dominates the crowding-out effect, leading to a faster HFT ($\frac{d\phi^*}{d\varepsilon} > 0$), whereas the direct effect of a speed bump on the spread dominates the speed-up by the HFT ($\frac{ds^*}{d\varepsilon} < 0$). Such parameter regions also exist in the general models with multiple HFTs or DLTs (e.g., see Figure 9 in Section 3).

Note that exchange platforms, in reality, earn a large portion of their profits from fees for fast data/exchange access, such as direct data feeds. As long as parameters (ε, c) satisfy the condition in Claim 1, an exchange can charge a higher price for speed by introducing a speed bump, as HFTs try to compensate for the loss from a delay and demand more speed services. Simultaneously, the exchange with a speed bump can achieve its intended goal, i.e., alleviating the adverse selection problem. Of course, we need to formalize the environment and the objective function of exchanges, but the above claim leads to the following prediction:

Conjecture 1. *Suppose that an exchange’s objective function consists of (i) maximizing fee revenues paid by*

HFTs for fast access to the exchange and (ii) alleviating the adverse selection problem. Then, there exists a speed bump with appropriate design such that the exchange has an incentive to adopt it.

The above conjecture goes counter to the result in Budish, Lee and Shim (2018), as their main argument is that exchanges do not have an incentive to introduce FBA.²⁵ Since their model does not have the crowding-in effect, introducing FBA always invalidates the speed advantage of HFTs. Then, HFTs stop purchasing fast data/connection services from exchanges, undermining exchanges' profits from speed fees. This pessimistic result holds because the previous studies focus on the positive side of a speed bump regarding speed acquisition and adverse selection.

In contrast, my model's crowding-in effect shows that a slow market structure does not always diminish HFTs' demand for speed. It can be seen as the negative side of a speed bump *per se* but proposes a more optimistic prediction once we consider exchanges' incentive: they will introduce a new market design that contributes to creating a level playing field.

5.3 Exogenous cost of speed and asset volatility

Discussions so far indicate that the *marginal* exogenous cost of speed $C'(\phi)$ is one of the important determinants of the effectiveness of a speed bump. In reality, typical high-frequency trading firms pay separate (monthly) subscription fees for physical and logical connectivity to an exchange, an additional connectivity charge to receive market data through acquired connectivity, and for the content of market data itself.

²⁵Budish, Lee and Shim (2018) show that, in equilibrium, exchanges charge positive fees on speed but no fees on transaction services. Such an equilibrium arises due to the trading rule in the US, such as RegNMS and UTP. On the one hand, trading services are homogeneous goods across exchanges, meaning that exchanges are fungible in terms of trading. On the other hand, traders need to purchase exchange specific speed technologies (ESST), making an exchange a monopolistic supplier of her ESST. In this situation, even if an exchange who deviates and adopts FBA earns profits by increasing trading fees, the profit margin easily disappears since other competing exchanges can imitate the FBA structure, resulting in lower profits compared to the status-quo structure.

My model provides some testable predictions, as the impact of a speed bump can be substantially different depending on the marginal cost of speed.

Conjecture 2. *All else equal, if the exogenous marginal cost of increasing speed is low, introducing a speed bump or its extension is accompanied by an increase in spread (and vice versa).*

However, high-frequency trading firms are notoriously secretive regarding their costs, and measuring the marginal cost of speed c is not easy. Thus, instead of measuring the actual cost of speed, considering the *effective* cost may provide an alternative prediction.

My model restricts HFTs' trading volume and assumes that the asset's volatility, σ , and the frequency of jumps, $\kappa = \mathbb{E}[t_0]$, are fixed. Varying these observable parameters shrinks or magnifies the effective cost of speed per sniping opportunity, leading to different consequences.

For example, the HFT's net profit per sniping is represented as

$$W_{HFT} = \frac{1}{\underbrace{1 + \beta\kappa}_{=\tilde{\pi}(\phi)} \frac{\phi}{\phi + \beta + \gamma}} h(\varepsilon) \times \sigma \left(1 - \frac{s}{\sigma}\right) - C(\phi),$$

where $\frac{s}{\sigma}$ is the effective spread and independent of σ . The above formula implies that a higher asset volatility, σ , reduces the effective marginal exogenous cost of speed *per sniping reward*, which is now defined by $\frac{C'}{\sigma\pi}$, weakening the crowding-out effect.

Also, the exogenous cost should be measured per sniping opportunity. For example, Virtu Financial, Inc. reported spending approximately \$209 million on communication and data processing and \$383 million on employee compensation and payroll taxes in 2019.²⁶ Although the aggregate cost looks large, it represents the *annual* amount. In contrast, according to the analysis by Laughlin (2014) and Bloomberg, Virtu conducts 2.5-3.5 million trades *per day* across all venues and all asset classes in the United States. The large number of transactions dwarfs the above operational cost, making

²⁶The expenditure in 2019 is the largest in these past 8 years. Virtu's SEC filings are available at <https://ir.virtu.com/financials-and-filings/sec-filings/default.aspx>.

the cost per transaction amount to \$0.67-\$0.94. Of course, the above back-of-envelope calculation cannot separate the cost for sniping from that for market making. Yet, what matters is the fact that a more frequent arrival of arbitrage opportunities reduces the cost per sniping and leads to a strong crowding-in effect (and vice versa). This discussion is translated into a change in κ (i.e., the expected time until the asset value jumps) that negatively affects π .

Conjecture 3. *All else equal, if the volatility of asset (measured by the frequency of a jump κ and its size σ) is high, the introduction of a speed bump or its extension is accompanied by an increase in spreads (and vice versa).*

5.4 Application to general information design

My result can fit into the discussion on more general information regulation in financial markets. For example, one of the closest market structures to a speed bump is “Last Look” in foreign exchange markets (FX): after a liquidity taker trades on a market maker’s posted price, the market maker obtains a small “latency buffer” in which she has an option to reject the trade. The design of the Last Look options differs across ECNs (Electronic Communication Networks) and OTC (Over-the-Counter) markets, but the primary goal is to mitigate adverse selection and other structural frictions for market makers that stem from the speed/informational advantage of liquidity takers in the highly fragmented FX markets. A couple of theoretical papers have studied the Last Look feature (e.g., Oomen, 2017; Cartea, Jaimungal and Walton, 2019), but the speed acquisition by liquidity takers is not taken into account.

In general, financial markets have been experienced many forms of informational regulation that try to mitigate information asymmetry between traders, such as Regulation Fair Disclosure (2000), Sarbanes–Oxley Act (2002), and Dodd-Franc Act (2010). The theoretical literature has argued that attempts to mitigate asymmetric information by disseminating public information can hammer private information acquisition by potential informed traders (i.e., the crowding-out effect; Verrecchia,

1982).²⁷ If the crowding-out effect is the only force at play, fair disclosure policies unambiguously mitigate asymmetric information and tighten the spread. However, “the empirical evidence [...] is still inconclusive” (Beyer et al., 2010). Even though my model focuses on speed acquisition, the mechanism for the crowding-in effect should survive in a general regulatory framework.

For example, in the benchmark model, parameter ϕ governs the learning process of the HFT. In the context of fair disclosure, we can re-interpret ϕ (or $\phi/(\phi + \beta)$) as the measure of traders with perfect private information about v . Also, parameter γ captures the arrival frequency of public news—fair disclosure—that tells all traders v . An intentional delay δ makes public disclosure more likely to happen before informed traders may take liquidity. Therefore, δ can be seen as the regulation that requests and promotes the swift dissemination of public news.

Following the discussion in Section 2, the regulation generates an additional opportunity for the informed trader to exploit her informational advantage. As a result, fair disclosure may promote traders’ information acquisition activity, i.e., the crowding-in effect. Although the formal analyses are left as a future research topic, my model predicts that the crowding-in effect should exist as long as regulations try to mitigate asymmetric information, whereas traders can strategically decide on the quality (speed and precision) of their private information.

6 Conclusion

Speed bumps have been widely adopted in many exchanges as one of the simplest and easiest ways to slow down fast informed trading. It seeks to slow down HFT and mitigate adverse selection for liquidity providers by imposing an intentional delay on HFTs’ order execution. In contrast to its intended purpose, however, my model shows that a speed bump can facilitate the equilibrium speed acquisition by the HFTs, i.e., it has a *crowding-in* effect. The crowding-in effect competes against a

²⁷The studies that propose the crowding-out effect include Diamond (1985), Gao and Liang (2013), Colombo, Femminis and Pavan (2014), and Banerjee, Davis and Gondhi (2018).

speed bump's intended effect, and the equilibrium speed acquisition and adverse selection exhibit rich reactions to a speed bump. Thus, this paper complements the literature by theoretically showing that a speed bump can have a positive, negative, and ambiguous impact on the market. It also proposes a couple of determinants, such as the cost of speed and the expected length of a delay.

This paper describes how fast HFTs obtain private information and act on it. Although this follows the literature and provides a tractable framework, it ignores another dimension of information acquisition, namely, the precision of a signal. Several studies, such as [Huang and Yueshen \(2018\)](#) and [Dugast and Foucault \(2018\)](#), argue that the precision of private information and the information processing speed may have a negative relationship. Hence, analyzing the impact of a speed bump on information acquisition, both in terms of speed and precision, would be a good extension of my model to capture a more holistic view on the impact of an exogenous speed restriction.

In this paper, I have not analyzed the market design problem, i.e., the optimal structure of speed bumps to improve market quality and traders' welfare. The existing analyses argue for a tradeoff generated by a speed bump: a decline in the spread (a lower trading cost) versus higher delay costs, such as quote fading. Thus, if the crowding-out effect is dominant, weighing these positive and negative effects should pin down the optimal delay. However, my model shows the crowding-in effect and complicates the discussion, as the speed bump may worsen liquidity, the trading cost, and trader welfare at the same time. In contrast, the model's prediction regarding exchanges' incentive to adopt a speed bump is optimistic, and modeling the exchange competition with a speed bump is the future research topic.

References

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

- Aït-Sahalia, Yacine and Mehmet Saglam**, “High frequency traders: Taking advantage of speed,” Technical Report, National Bureau of Economic Research 2013.
- Aldrich, Eric M and Daniel Friedman**, “Order protection through delayed messaging,” Technical Report, WZB Discussion Paper 2018.
- American, NYSE**, “NYSE American to Eliminate Speed Bump, Restore Floor-Based Trading,” January 2019.
- Anderson, Lisa, Emad Andrews, Baiju Devani, Michael Mueller, and Adrian Walton**, “Speed segmentation on exchanges: Competition for slow flow,” Technical Report, Bank of Canada Staff Working Paper 2018.
- Aquilina, Matteo, Eric B Budish, and Peter O’Neill**, “Quantifying the high-frequency trading “arms race”: A simple new methodology and estimates,” *Chicago Booth Research Paper*, 2020, (20-16).
- Baldauf, Markus and Joshua Mollner**, “High-frequency trading and market performance,” *SSRN Electronic Journal*, 2017.
- Banerjee, Snehal, Jesse Davis, and Naveen Gondhi**, “When transparency improves, must prices reflect fundamentals better?,” *The Review of Financial Studies*, 2018, 31 (6), 2377–2414.
- Baruch, Shmuel and Lawrence R Glosten**, “Flickering quotes,” *Columbia University*, 2013.
- and —, “Tail expectation and imperfect competition in limit order book markets,” *Journal of Economic Theory*, 2019, 183, 661–697.
- Beyer, Anne, Daniel A Cohen, Thomas Z Lys, and Beverly R Walther**, “The financial reporting environment: Review of the recent literature,” *Journal of Accounting and Economics*, 2010, 50 (2-3), 296–343.

- 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.
- and **Marius Zoican**, “Liquid speed: On-demand fast trading at distributed exchanges,” *arXiv preprint arXiv:1907.10720*, 2019.
- Brunnermeier, Markus K and Lasse Heje Pedersen**, “Market liquidity and funding liquidity,” *The review of financial studies*, 2009, 22 (6), 2201–2238.
- 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.
- Cartea, Alvaro, Sebastian Jaimungal, and Jamie Walton**, “Foreign exchange markets with last look,” *Mathematics and Financial Economics*, 2019, 13 (1), 1–30.
- Chakrabarty, Bidisha, Jianning Huang, and Pankaj K Jain**, “Effects of a Speed Bump on Market Quality and Exchange Competition,” *Available at SSRN 3280645*, 2019.

- Chen, Haoming, Sean Foley, Michael Goldstein, and Thomas Ruf**, "The Value of a Millisecond: Harnessing Information in Fast, Fragmented Markets," *SSRN Electronic Journal*, 2017.
- Colombo, Luca, Gianluca Femminis, and Alessandro Pavan**, "Information acquisition and welfare," *The Review of Economic Studies*, 2014, 81 (4), 1438–1483.
- 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.
- Dennert, Jürgen**, "Price competition between market makers," *The Review of Economic Studies*, 1993, 60 (3), 735–751.
- Diamond, Douglas W**, "Optimal release of information by firms," *The Journal of Finance*, 1985, 40 (4), 1071–1094.
- Du, Songzi and Haoxiang Zhu**, "What is the optimal trading frequency in financial markets?," *The Review of Economic Studies*, 2017, 84 (4), 1606–1651.
- Dugast, Jérôme and Thierry Foucault**, "Data abundance and asset price informativeness," *Journal of Financial Economics*, 2018, 130 (2), 367–391.
- Foucault, Thierry, Ailsa Roell, and Patrik Sandas**, "Market making with costly monitoring: An analysis of the SOES controversy," *The Review of Financial Studies*, 2003, 16 (2), 345–384.
- , **Ohad Kadan, and Eugene Kandel**, "Liquidity cycles and make/take fees in electronic markets," *The Journal of Finance*, 2013, 68 (1), 299–341.
- , **Roman Kozhan, and Wing Wah Tham**, "Toxic arbitrage," *The Review of Financial Studies*, 2016, 30 (4), 1053–1094.
- Frino, Alex, Vito Mollica, and Robert I Webb**, "The impact of co-location of securities exchanges' and traders' computer servers on market liquidity," *Journal of Futures Markets*, 2014, 34 (1), 20–33.

- Gao, Pingyang and Pierre Jinghong Liang**, “Informational feedback, adverse selection, and optimal disclosure policy,” *Journal of Accounting Research*, 2013, 51 (5), 1133–1158.
- 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.
- Han, Bing, Ya Tang, and Liyan Yang**, “Public information and uninformed trading: Implications for market liquidity and price efficiency,” *Journal of Economic Theory*, 2016, 163, 604–643.
- Han, Jung Suk, Mariana Khapko, and Albert S Kyle**, “Liquidity with high-frequency market making,” 2014.
- Hasbrouck, Joel and Gideon Saar**, “Low-latency trading,” *Journal of Financial Markets*, 2013, 16 (4), 646 – 679.
- Hendershott, Terrence and Pamela C Moulton**, “Automation, speed, and stock market quality: The NYSE’s hybrid,” *Journal of Financial Markets*, 2011, 14 (4), 568–604.
- Hoffmann, Peter**, “A dynamic limit order market with fast and slow traders,” *Journal of Financial Economics*, 2014, 113 (1), 156–169.
- Hu, Edwin**, “Intentional access delays, market quality, and price discovery: Evidence from IEX becoming an exchange,” *SEC Working Paper*, 2018.
- Huang, Shiyang and Bart Z Yueshen**, “Speed Acquisition,” *Available at SSRN 2845864*, 2018.
- Jones, Charles M**, “What do we know about high-frequency trading?,” *Columbia University Working Paper*, 2013.

- Khapko, Mariana and Marius Zoican**, “Do Speed Bumps Curb Speed Investment? Evidence from a Pilot Experiment,” *Evidence from a Pilot Experiment* (March 12, 2019), 2019.
- Kyle, Albert S and Jeongmin Lee**, “Toward a fully continuous exchange,” *Oxford Review of Economic Policy*, 2017, 33 (4), 650–675.
- Kyle, S Albert**, “Continuous Auctions and Insider Trading,” *Econometrica*, 1985, 53 (6), 1315–1335.
- Laughlin, Greg**, “Insights Into High Frequency Trading from the Virtu Initial Public Offering,” 2014.
- 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.
- Oomen, Roel**, “Last look,” *Quantitative Finance*, 2017, 17 (7), 1057–1070.
- 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.

Verrecchia, Robert E, “The use of mathematical models in financial accounting,” *Journal of Accounting Research*, 1982, pp. 1–42.

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 Appendix: Proofs

In what follows, I denote $\psi = \phi + \beta + \gamma$ in the single-HFT model and $\psi_i = \phi_S + \beta + \phi_i$ in the multiple-HFTs model to save the space.

A.1 Derivation of the equilibrium spread

Suppose that t_0 is fixed. The first and second coefficients of s in the first line of (1) are computed as follows.

$$\begin{aligned} \text{First coef.} &= \mathbb{E}_\delta \left[\int_\delta^{\max\{t_0, \delta\}} \beta e^{-\beta(\tau-\delta)} d\tau \right] \\ &= \int_0^{t_0} \varepsilon^{-1} e^{-\varepsilon^{-1}\delta} (1 - e^{-\beta(\max\{t_0, \delta\}-\delta)}) d\delta \\ &= 1 - e^{-\varepsilon^{-1}t_0} - \frac{\varepsilon^{-1}}{\varepsilon^{-1} - \beta} (e^{-\beta t_0} - e^{-\varepsilon^{-1}t_0}). \end{aligned}$$

$$\begin{aligned} \text{Second coef.} &= \mathbb{E}_\delta \left[\int_{\max\{t_0, \delta\}}^{t_0+\delta} \beta e^{-\beta(\tau-\delta)} e^{-\gamma(\tau-t_0)} d\tau \right] \\ &= \frac{\beta e^{-\beta t_0}}{\beta + \gamma} \left[\int_0^{t_0} \varepsilon^{-1} e^{-(\varepsilon^{-1}-\beta)\delta} (1 - e^{-(\beta+\gamma)\delta}) d\delta + (e^{(\beta+\gamma)t_0} - 1) \int_{t_0}^\infty \varepsilon^{-1} e^{-(\varepsilon^{-1}+\gamma)\delta} d\delta \right] \\ &= \beta e^{-\beta t_0} \frac{1}{1 + \varepsilon\gamma} \frac{1 - e^{-(\varepsilon^{-1}-\beta)t_0}}{\varepsilon^{-1} - \beta}. \end{aligned}$$

Also, the coefficient of the profit in the second line of (1) is

$$\begin{aligned}\text{Third coef.} &= \int_0^\infty \varepsilon^{-1} e^{-\varepsilon^{-1}\delta} \left(\int_{t_0+\delta}^\infty e^{-\phi(\tau-t_0-\delta)} e^{-\beta(\tau-\delta)} e^{-\gamma(\tau-t_0)} d\tau \right) d\delta \\ &= \frac{e^{-\beta t_0}}{\psi} \frac{1}{1+\varepsilon\gamma}.\end{aligned}$$

Then, the expected profit is reduced to

$$V_M \sim s\beta \frac{1}{\kappa^{-1} + \varepsilon^{-1}} \left(\varepsilon^{-1} + \frac{\kappa^{-1}}{1+\varepsilon\gamma} \right) + \frac{1}{1+\varepsilon\gamma} (\beta s + \phi(s-\sigma)) \frac{\kappa^{-1}}{\psi}, \quad (31)$$

leading to the equilibrium spread in the main text.

A.2 Proof of Proposition 4 (uniqueness)

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

Secondly, suppose that HFT i puts positive weight w on $p \in (s_{j,0}, \sigma)$. For this to be an equilibrium, there must exist positive ϵ such that HFT j does not charge prices in $[s_{j,0}, p + \epsilon]$. If not, HFT j can exploit the profit discontinuity at p , and she undercuts HFT i to obtain positive profits. Also, if HFT j charges prices below p , it is not optimal for HFT i to put a positive weight on p . Thus, HFT j must charge prices above $p + \epsilon$. In this case, however, it is optimal for HFT i to raise p .

Finally, suppose that $p = s_{j,0}$ has a positive weight. For $s_j \geq p = s_{j,0}$, the profits for HFM j satisfy

$$\begin{aligned}V_{M,j}(s_j) &\sim \frac{\phi_S + \beta(1 + \lambda\psi_j)}{\psi_j} (s_j - s_{j,0}) \Pr(s_j < s_i) + \left[\frac{\phi_S + \beta(1 + \lambda\psi_j)}{\psi_j} (s_j - s_{j,0}) - (1-m) \frac{\beta(1 + \lambda\psi_j)}{\psi_j} s_j \right] \Pr(s_j = s_i) \\ &\quad + \left[\frac{\phi_S + \beta(1 + \lambda\psi_j)}{\psi_j} (s_j - s_{j,0}) - \frac{\beta(1 + \lambda\psi_j)}{\psi_j} s_j \right] \Pr(s_j > s_i) \\ &= \frac{\phi_S + \beta(1 + \lambda\psi_j)}{\psi_j} (s_j - s_{j,0}) - \frac{\beta(1 + \lambda\psi_j)}{\psi_j} s_j [\Pr(s_j > s_i) + (1-m) \Pr(s_j = s_i)] \\ &< \frac{\phi_S + \beta(1 + \lambda\psi_j)}{\psi_j} (s_j - s_{j,0}) - \frac{\beta(1 + \lambda\psi_j)}{\psi_j} s_j \Pr(s_j > s_i).\end{aligned}$$

Then, we can define

$$\epsilon \equiv \frac{\beta(1 + \lambda\psi_j)s_{j,0} \Pr(s_j > s_i)}{\phi_S + \beta(1 + \lambda\psi_j)(1 - \Pr(s_j > s_i))} > 0$$

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

A.3 Proof of Proposition 5

In what follows, the comparative statics are analyzed by varying λ instead of ϵ because it affects the equilibrium spread only via $\lambda(\epsilon)$ and $\lambda'(\epsilon) > 0$.

Firstly, note that $\log x$ has the following properties for general $x \in (0, 1)$.

$$\log x + 1 - x < 0, \quad (32)$$

$$2(\log x + 1 - x) + (1 - x)^2 < 0. \quad (33)$$

Also, I denote

$$X_j = \log s_{j,0} + 1 - s_{j,0}.$$

Next, for later use, compute the following derivatives and cross-derivatives: with $\eta = \phi_j + \beta$,

$$\frac{d\mathbb{E}[s_j]}{d\phi_S} = -\frac{1 + \lambda\eta}{\beta} \frac{X_j}{(1 + \lambda\psi_j)^2} > 0, \quad (34)$$

$$\frac{d^2\mathbb{E}[s_j]}{d\phi_S^2} = -\frac{1 + \lambda\eta}{\beta\phi_S} \left[\frac{(1 - s_{j,0})^2 (1 + \lambda\eta) - 2\lambda\phi_S X_j}{(1 + \lambda\psi_j)^3} \right] < 0,$$

$$\frac{\partial \mathbb{E}[s_j]}{\partial \phi_j} = \lambda \frac{\phi_S}{\beta} \frac{X_j}{(1 + \lambda \psi_j)^2} < 0,$$

$$\frac{\partial}{\partial \phi_j} \left(\frac{d\mathbb{E}[s_j]}{d\phi_S} \right) = \frac{\lambda}{\beta} \frac{1}{(1 + \lambda \psi_j)^3} [X_j(1 + \lambda \beta) + (1 + \lambda \eta)(1 - s_{j,0})^2],$$

$$\frac{d\mathbb{E}[s_j]}{d\lambda} = \psi \frac{\phi_S}{\beta} \frac{X_j}{(1 + \lambda \psi_j)^2} < 0.$$

$$\frac{\partial}{\partial \lambda} \left(\frac{d\mathbb{E}[s_j]}{d\phi_S} \right) = \frac{1}{\beta(1 + \lambda \psi_j)^3} [[\phi_S + \psi_j(1 + \lambda \eta)]X_j + (1 + \lambda \eta)\psi_j(1 - s_{j,0})^2].$$

Also, $\phi_S^* = BR(\phi_j)$ satisfies the FOC:

$$0 = H(\phi_S, \phi_j, \lambda) = \frac{d\tilde{\pi}_S}{d\phi_S}(1 - \mathbb{E}[s_j]) - \tilde{\pi}_S \frac{d\mathbb{E}[s_j]}{d\phi_S}. \quad (35)$$

The SOC is computed as

$$\begin{aligned} \partial H / \partial \phi_S &= (1 - \mathbb{E}[s_j]) \frac{d^2 \tilde{\pi}}{d\phi_S^2} - 2 \frac{d\tilde{\pi}_i}{d\phi_S} \frac{d\mathbb{E}[s_j]}{d\phi_S} - \tilde{\pi} \frac{d^2 \mathbb{E}[s_j]}{d\phi_S^2} \\ &= -2 \frac{1}{\psi_j} \frac{d\mathbb{E}[s_j]}{d\phi_S} - \frac{\phi_S}{\psi_j} \frac{d^2 \mathbb{E}[s_j]}{d\phi_S^2} \\ &= \frac{1}{\psi_j} \frac{(1 + \lambda \eta)^2}{\beta(1 + \lambda \psi_j)^3} [2X_j + (1 - s_{j,0})^2] < 0, \end{aligned} \quad (36)$$

where the last inequality uses (33).

Impact of λ on $\phi_S^* = BR(\phi_j)$ Since (36) holds, it suffices to show that $\partial H / \partial \lambda > 0$.

$$\begin{aligned} \partial H / \partial \lambda &= - \left(\frac{d\tilde{\pi}_S}{d\phi_S} \frac{\partial \mathbb{E}[s_j]}{\partial \lambda} + \tilde{\pi}_S \frac{\partial}{\partial \lambda} \frac{d\mathbb{E}[s_j]}{d\phi_S} \right) \\ &= - \frac{\phi}{\beta} \frac{1 + \lambda \eta}{(1 + \lambda \psi_j)^3} [2X_j + (1 - s_{j,0})^2] \\ &> 0, \end{aligned} \quad (37)$$

where the last inequality comes from (33).

Impact of ϕ_j on $\phi_S^* = BR(\phi_j)$ Since (36) holds, it suffices to show that $\partial H/\partial \phi_j > 0$. The cross derivative of FOC is

$$\partial H/\partial \phi_j = -\frac{d\tilde{\pi}_S}{d\phi_S} \frac{\partial \mathbb{E}[s_j]}{\partial \phi_j} - \tilde{\pi}_S \frac{\partial}{\partial \phi_j} \frac{d\mathbb{E}[s_j]}{d\phi_S} + \frac{\partial}{\partial \phi_j} \frac{d\tilde{\pi}_S}{d\phi_S} (1 - \mathbb{E}[s_j]) - \frac{\partial \tilde{\pi}_S}{\partial \phi_j} \frac{d\mathbb{E}[s_j]}{d\phi_S}.$$

By using the FOC, the first and the last terms sum up to

$$A_0 = -\frac{d\tilde{\pi}_S}{d\phi_S} \frac{\partial \mathbb{E}[s_j]}{\partial \phi_j} - \frac{\partial \tilde{\pi}_S}{\partial \phi_j} \frac{d\mathbb{E}[s_j]}{d\phi_S} = -\frac{\phi_S}{\beta \psi_j} \frac{X_j}{\psi_j (1 + \lambda \psi_j)^2} (1 + 2\lambda \eta).$$

Also, the middle two terms amount to

$$\begin{aligned} A_1 &= -\tilde{\pi}_S \frac{\partial}{\partial \phi_j} \frac{d\mathbb{E}[s_j]}{d\phi_S} + (1 - \mathbb{E}[s_j]) \frac{\partial}{\partial \phi_j} \frac{d\tilde{\pi}_S}{d\phi_S} \\ &= \tilde{\pi}_S \left[-\frac{\partial}{\partial \phi_j} \frac{d\mathbb{E}[s_j]}{d\phi_S} + \left(\frac{d\tilde{\pi}_S}{d\phi_S} \right)^{-1} \frac{d\mathbb{E}[s_j]}{d\phi_S} \frac{\partial}{\partial \phi_j} \frac{d\tilde{\pi}_S}{d\phi_S} \right] \\ &= \frac{\tilde{\pi}_S}{(1 + \lambda \psi_j)^3} \left[-\frac{\lambda}{\beta} X_j (1 + \lambda \beta) - \frac{1}{\beta} \lambda (1 + \lambda \eta) (1 - s_{j,0})^2 + \frac{(1 + \lambda \eta)(1 + \lambda \psi_j)}{\psi_j \eta} X_j \right]. \end{aligned}$$

Therefore, in aggregate, $\partial H/\partial \phi_j = \frac{\tilde{\pi}_S}{(1 + \lambda \psi_j)^3} A$ with

$$\begin{aligned} A &= -\frac{1}{\beta \psi_j \eta} \{ X [\lambda \psi_j \eta (1 + \lambda \beta) + (\phi(1 + \lambda \eta) + \lambda \eta^2) (1 + \lambda \psi_j)] + \psi_j \eta \lambda (1 + \lambda \eta) (1 - s_{j,0})^2 \} \\ &> \frac{1}{2} \frac{1}{\beta \psi_j \eta} \{ [\lambda \psi_j \eta (1 + \lambda \beta) + (\phi(1 + \lambda \eta) + \lambda \eta^2) (1 + \lambda \psi_j)] - 2\psi_j \eta \lambda (1 + \lambda \eta) \} (1 - s_{j,0})^2 \\ &> 0, \end{aligned}$$

where the second inequality comes from (33).

A.4 Proof of Propositions 6 and 7

Denote the speed level in the symmetric equilibrium as ϕ . Also, I omit indices of other variables, such as ψ and s . Firstly, the FOC $H(\phi, \phi, \lambda) = 0$, defined by (35), is reduced to

$$0 = \hat{H}(\phi, \lambda) = \beta\eta(1 + \lambda\psi)^2 + \phi\psi(1 + \lambda\eta)(1 - s_0) + \phi(\phi + 2\eta(1 + \lambda\psi)) \log s_0.$$

It is easy to check $\hat{H}(0, \lambda) = \beta^2(1 + \lambda\beta)^2 > 0$. Also, at $\phi \rightarrow \infty$, $1 - s_0$ and $\log s_0$ converge to some constant values, and $\hat{H}(\phi, \lambda)$ can be written as a cubic function of ϕ with a negative coefficient on ϕ^3 , meaning that $\lim_{\phi \rightarrow \infty} \hat{H}(\phi, \lambda) < 0$. As shown in the following, $\hat{H}(\phi, \lambda)$ is a monotonically decreasing function of ϕ . Thus, the above argument implies that $\hat{H} = 0$ and $H = 0$ have a unique solution.

Next, by using the derivatives in Appendix A.3,

$$\begin{aligned} \frac{\partial H}{\partial \phi} &= \frac{\psi - 4\eta}{\psi^3} \left(1 + \frac{\phi}{\beta(1 + \lambda\psi)} \log s_0\right) \\ &+ \frac{\beta + \eta + 2\lambda\eta\beta + \lambda\phi\psi}{\beta\psi^2} \frac{X}{(1 + \lambda\psi)^2} + \frac{1}{\psi} \frac{1 + \lambda\eta}{\beta} \frac{(1 + \lambda\beta)(1 - s_0)^2 - 4\phi\lambda X}{(1 + \lambda\psi)^3}. \end{aligned}$$

Inequalities (32) and (33) imply that $\frac{\partial H}{\partial \phi} < 0$. Since Appendix A.3 shows that $\frac{\partial H}{\partial \lambda} > 0$, the implicit function theorem implies $\frac{d\phi}{d\lambda} > 0$, where the explicit formula is

$$\frac{d\phi}{d\lambda} = \frac{\psi^3\phi(1 + \lambda\eta) [2X + (1 - s)^2]}{\beta(\psi - 4\eta)(1 + \lambda\psi)^3 \left(1 + \frac{\phi}{\beta(1 + \lambda\psi)} \log s_0\right) + a_X X + \psi^2(1 + \lambda\eta)(1 + \lambda\beta)(1 - s)^2}, \quad (38)$$

with

$$a_X = \psi(1 + \lambda\psi)[\beta + \eta + 2\lambda\eta\beta + \lambda\phi\psi] - 4\psi^2\phi\lambda(1 + \lambda\eta).$$

Finally, we want to show that the expected spread is increasing in λ . At the symmetric equilibrium,

$$\begin{aligned}\frac{d\mathbb{E}[s]}{d\lambda} &= \overbrace{\psi \frac{\phi}{\beta} \frac{1-s_0 + \log s_0}{(1+\lambda\psi)^2}}^{=\frac{\partial \mathbb{E}[s]}{\partial \lambda}} + \frac{d\phi}{d\lambda} \overbrace{\left[-\frac{1+\lambda\beta}{\beta} \frac{\log s_0 + 1-s_0}{(1+\lambda\psi)^2} \right]}^{=\frac{\partial \mathbb{E}[s]}{\partial \phi}} \\ &= \frac{X}{\beta(1+\lambda\psi)^2} \left[\psi\phi - (1+\lambda\beta) \frac{d\phi}{d\lambda} \right].\end{aligned}$$

Since $X < 0$, inequality $\frac{d\mathbb{E}[s]}{d\lambda} > 0$ is reduced to

$$\frac{d\phi}{d\lambda} > \frac{\psi\phi}{1+\lambda\beta}.$$

By using the explicit formula (38), the above condition is reduced to

$$\begin{aligned}[2\psi^2(1+\lambda\eta)(1+\lambda\beta) - a_X] X &< \beta(\phi - 3\eta)(1+\lambda\psi)^3(1 - \mathbb{E}[s]) \\ &= -\beta(\phi - 3\eta)(1+\lambda\psi)^3 \frac{\phi\psi}{\beta\eta} (1+\lambda\eta) \frac{X}{(1+\lambda\psi)^2}\end{aligned}$$

where the last equality comes from the FOC. As a result, the condition is

$$\begin{aligned}[2\psi^2(1+\lambda\eta)(1+\lambda\beta) - a_X] X &< (3\eta - \phi)(1+\lambda\psi) \frac{\phi\psi}{\eta} (1+\lambda\eta) X \\ \therefore 2\psi^2(1+\lambda\eta)(1+\lambda\beta) - a_X &> (3\eta - \phi)(1+\lambda\psi) \frac{\phi\psi}{\eta} (1+\lambda\eta).\end{aligned}$$

We can compute that

$$\begin{aligned}\text{LHS} - \text{RHS} &= (1+\lambda\psi) \frac{\phi^2}{\eta} + 4\phi\lambda\psi(1+\lambda\eta) \\ &> 0.\end{aligned}$$

This inequality completes the proof.

A.5 DLTs in multiple HFTs model

Consider the model with multiple HFTs in Section 3. Due to the DLTs, the profits from market making is modified as

$$V_{M,i}(s_i) \propto [\phi_S(s_i - \sigma) + \theta(s_i, s_j)\beta(1 + \lambda\psi_i)s_i(1 - m(s_i))],$$

where the notation is the same as the main text. Following the mixed strategy equilibrium in Section 3, define the break-even spread with $\theta = 1$ as the solution of

$$s_{i,0} = \frac{\phi_S}{\phi_S + \beta(1 + \lambda\psi_i)(1 - m(s_{i,0}))}\sigma,$$

which is unique in $s \in [0, \sigma]$ if and only if $\bar{\alpha} > 1$. Then, the analogous arguments to Section 3 lead to the following mixed strategy distribution F_i

$$F_i(s) = 1 - \frac{\phi_S}{\beta(1 + \lambda\psi_j)(1 - m(s))} \frac{\sigma - s}{s},$$

and the ex-ante expected sniping profit of HFT i is given by

$$W_i^{gross} = \max_{\phi_i} \pi_i \left(1 - \int_{s_{j,0}}^{\sigma} s dF_j(s) \right).$$

The explicit formula for the expected spread is given by

$$\mathbb{E}[s_j] = -\frac{\phi_S}{\beta(1 + \lambda\psi_j)} \left[(1 - m_0)Q_0 \log s_0 - \frac{a_0}{m_1} \left(\frac{1}{1 - m_0 - m_1} - \frac{1}{1 - m_0 - m_1 s_0} \right) + \frac{a_1}{m_1} \log \left(\frac{1 - m_0 - m_1}{1 - m_0 - m_1 s_0} \right) \right]$$

where

$$\begin{aligned}
m_0 &= \frac{\psi}{1 + \lambda\psi} \frac{1}{\bar{\alpha}} \left(\varepsilon\gamma\hat{\chi} + \frac{1}{\psi} - \frac{1}{1 + \varepsilon\gamma} \frac{1}{2\psi - \beta} \right), \\
m_1 &= \frac{\psi}{1 + \lambda\psi} \frac{1}{\bar{\alpha}} \left(\hat{\chi} + \frac{1}{1 + \varepsilon\gamma} \frac{1}{2\psi - \beta} \right), \\
Q_0 &= \frac{1}{(1 - m_0)^2}, a_0 = 1 - m_0 - m_1 \\
a_1 &= \frac{m_1}{1 - m_0} - 1,
\end{aligned}$$

and

$$s_0 = \frac{\phi_S + \beta(1 + \lambda\psi_i)(1 - m_0) - \sqrt{[\phi_S + \beta(1 + \lambda\psi_i)(1 - m_0)]^2 - 4\phi_S m_1 \beta(1 + \lambda\psi_i)}}{2m_1 \beta(1 + \lambda\psi_i)}. \quad (39)$$

Note that s_0 in (39) is the equilibrium spread in the single-HFT model with DLTs.

B Speed bump implementation and institutional details

This section briefly describes the institutional details of speed bumps. Although the model in the main text is built to analyze asymmetric speed bumps rather than a symmetric speed bump in IEX, I start this section by explaining IEX speed bump below, as it is a precursor of all other speed bumps.

Symmetric speed bump with deterministic delay

Who is protected? A symmetric and deterministic speed bump is first adopted by IEX in 2013 and followed by NYSE American (while NYSE American has decided to remove it based on NYSE, 2019). It delays all incoming and outgoing orders by 350 microseconds. The type of orders protected by the speed bump is “Pegged Order.” Pegged order is the type of non-displayed limit orders whose price is dynamically adjusted by reference to the national best bid and offer (NBBO).²⁸ Although IEX imposes

²⁸There are three types of pegged-order: Primary Peg (P-Peg), Discretionary Peg (D-Peg), and Midpoint Peg (M-Peg). P-Peg and D-Peg orders are resting at one tick below or above the NBBO. P-Peg orders have discretion to trade at the NBBO, while D-Peg orders have discretion to trade up to the midpoint. M-peg orders stay and are traded at the midpoint of the

a delay on incoming orders and outgoing information, the messaging of SIP-related information is not delayed. Thus, the price of pegged orders is dynamically adjusted by IEX with no delays. If the NBBO changes, a speed bump allows IEX to adjust the pegged orders, and HFTs cannot snipe them at the stale price.

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

Asymmetric speed bump

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

Empirically identifying the impact of speed bumps is not straightforward, as a speed bump typically comes with changes in other market structures and trading rules. For example, along with a speed bump, TSX Alpha sets a minimum size requirement for liquidity providing orders and adopts an inverted maker-taker fee structure. Due to these structural changes, market makers must pay NBBO and has a higher priority than D-Peg orders at the mid-point. Whether the discretion of each order is exercised is determined by the "IEX signal" that determines if the NBBO is volatile (i.e., "scrambling") by using a specific measure. The discretion is not exercised if the signal is "on," meaning that the bid-ask in the market is volatile.

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

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

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

For example, consider an informed trader who wants to fill a large order (say 1,000 shares). Exchanges A has 300 shares available, and Exchange B has 500 shares. Thus, the informed trader may spray orders to both exchange to fill 1,000 shares. Suppose that it takes t_A and t_B to send and execute orders on Exchanges A and B, respectively. Now, a speed bump is applied in Exchange B. If the length of delay δ is deterministic, the informed trader can stagger the timing of order entry to make $t_A = t_B + \delta$.²⁹ By the synchronized execution, there is no information leakage, and the trader can fulfill all orders.

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

²⁹In reality, there must involve some unexpected delays due to random factors, such as precipitation and temperature, and even a SOR cannot perfectly synchronize the order arrival timing.

liquidity providers bear less severe adverse selection than in the case with a deterministic speed bump.

Internet Appendix

A Sequential trading model

In this section, I relax the assumption in the benchmark model that a trade happens only once. Accommodating multiple trading rounds needs some modifications but the main implications and behavior of the model are the same as the baseline model. The basic idea follows the canonical model by [Glosten and Milgrom \(1985\)](#).

Trading takes place over the course of multiple trading rounds for $t > 0$, and t_k represents the time at which the k -th trade takes place. The sequence of time $\{t_k\}_{k=1}^{\infty}$ is notional in this model, in the sense that it is determined by the stochastic arrival of traders. Thus each t_k is random, as specified below.

The competitive market maker (MM) learns v from order flows and updates her belief. Upon observing the execution of a previous limit order at t_k , the market maker posts a new limit order with bid-ask prices $(-s_{k+1}, s_{k+1})$ anticipating that the next trade happens at t_{k+1} .

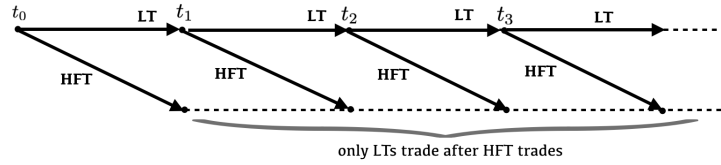
As in the benchmark model, a single HFT tries to take liquidity. She becomes informed at $T_H \sim \exp(\phi)$ and places a market order (or an informed HFT places an order at $t = 0$ and it arrives at T_H). The order arrives at the market with a delay $\delta \sim \exp(\varepsilon^{-1})$. The HFT trades only once and takes at most one unit of position.

Finally, there are n liquidity traders (LT). Each LT trades one unit of asset. They sequentially arrive at the market following the Poisson process with intensity β . For simplicity, I assume that LTs are not subject to a speed bump, and the arrival process of LTs are defined over $[0, \infty)$ rather than $[\delta, \infty)$. This assumption is reasonable if we think about a long-run model. Namely, $t = 0$ is thought of as the timing of the asset value jump rather than the timing that the market opens. An LT can

arrive at the market and take liquidity at $t \geq 0$ (rather than $t \geq \delta$) because it is possible that an LT is hit by the shock in $t < 0$.

Flow of events and the behavior of agents in each trading round are specified as follows and summarized in Figure 12.

Figure 12: Flow of events



Preliminary Let $b = \varepsilon^{-1}$ and $\tau_H \equiv T_H + \delta$. From the exponential distribution, τ_H has the following density function:

$$f_H(\tau_H = \tau) = \frac{b\phi}{b - \phi} \left(e^{-\phi\tau} - e^{-b\tau} \right).$$

Conditional on $\tau_H > m$, it becomes

$$f_H(\tau|m) \equiv f_H(\tau_H = \tau | \tau_H > m) = b\phi \frac{e^{-\phi\tau} - e^{-b\tau}}{be^{-\phi m} - \phi e^{-bm}}.$$

Thus, the process for τ_H is not memoryless. Also, denote the probability that τ_H happens after τ conditional on $\tau_H > m$ is

$$\bar{F}_H(\tau|m) = \Pr(\tau_H > \tau | \tau_H > m) = \frac{be^{-\phi\tau} - \phi e^{-b\tau}}{be^{-\phi m} - \phi e^{-bm}}.$$

In contrast, the arrival process of LTs is memoryless:

$$f_L(\tau|m) = \Pr(T_L = \tau | T_L > m) = \beta e^{-\beta(\tau-m)}$$

and

$$\bar{F}_L(\tau|m) = \Pr(T_L > \tau | T_L > m) = e^{-\beta(\tau-m)}.$$

Market maker's behavior

Consider a competitive market maker. From the market maker's perspective, the trading game goes as in Figure 12.

At $t = t_0 = 0$, the MM arrives and posts the spread s_1 anticipating that the next trade happens at stochastic time t_1 , where it is triggered either by the HFT or an LT with density $f_H(t_1|t_0)\bar{F}_L(t_1|t_0)$ and $\bar{F}_H(t_1|t_0)f_L(t_1|t_0)$, respectively.

Next, given the trade happens at t_1 , the MM posts a new price s_2 for the second trade that may happen at t_2 . If the first trade is done by the HFT, all remaining trades are with LTs (the state in Figure 12 transits from the upper horizontal line to the lower horizontal line). If the first trade is done by an LT, in contrast, the MM repeats the same game, i.e., both the HFT and an LT can trade with density $f_H(\tau|t_1)\bar{F}_L(\tau|t_1)$ and $\bar{F}_H(\tau|t_1)f_L(\tau|t_1)$, respectively (the state in Figure 12 stays on the upper horizontal line). This process repeats until $n + 1$ trades take place (n LTs and one HFT) and all traders fulfill their trading needs.

Equilibrium spread At $t = t_0 = 0$, the value function of a MM who posts spread s is

$$\begin{aligned} V_0(s, t_0) &= (s - \sigma) \overbrace{\int_{t_0}^{\infty} f_H(\tau|t_0)\bar{F}_L(\tau|t_0)d\tau}^{=g_H(t_0)} + s \overbrace{\int_{t_0}^{\infty} \bar{F}_H(\tau|t_0)f_L(\tau|t_0)d\tau}^{=g_L(t_0)} \\ &\equiv (s - \sigma)g_H(t_0) + sg_L(t_0). \end{aligned} \tag{40}$$

Thus, the competitive spread for the first trade is

$$s_1(t_0) = \frac{g_H(t_0)}{g_H(t_0) + g_L(t_0)}\sigma.$$

g_H and g_L denote the probability of HFT's and LT's arrival. Explicitly,

$$g_H(t) = \int_t^\infty f_H(\tau|t) \bar{F}_L(\tau|t) d\tau = \frac{\frac{\phi}{\phi+\beta} b e^{-\phi t} - \frac{b}{b+\beta} \phi e^{-bt}}{b e^{-\phi t} - \phi e^{-bt}}$$

and

$$g_L(t) = \int_t^\infty f_L(\tau|t) \bar{F}_H(\tau|t) d\tau = \beta \frac{\frac{b}{\beta+\phi} e^{-\phi t} - \frac{\phi}{b+\beta} e^{-bt}}{b e^{-\phi t} - \phi e^{-bt}}.$$

Consider the MM posting the new price $s = s_2$ at $t = t_1$. Her value function from posting s now depends on (i) the timing of the first trade t_1 and (ii) the identity of the first trader. Point (i) is due to the absence of the memoryless property.

$$\begin{aligned} V(s, t_1) &= \overbrace{g_L(t_0)}^{\text{LT trades in the 1st round}} \times \left[(s - \sigma) \overbrace{\int_{t_1}^\infty f_H(\tau|t_1) \bar{F}_L(\tau|t_1) d\tau}^{=g_H(t_1)} + s \overbrace{\int_{t_1}^\infty \bar{F}_H(\tau|t_1) f_L(\tau|t_1) d\tau}^{=g_L(t_1)} \right] \\ &+ \overbrace{g_H(t_0)}^{\text{HFT trades in the 1st round}} \times (s - \sigma) \int_{t_1}^\infty f_L(\tau|t_1) d\tau \\ &= g_L(t_0) [g_H(t_1)(s - \sigma) + g_L(t_1)s] + g_H(t_0)(s - \sigma). \end{aligned} \quad (41)$$

The first line shows the case that the first trade is done by an LT. In this case, the second trade is by the HFT, and the MM earns $s - \sigma$ with probability $g_H(t_1)$. Also, the second trade can be with an LT with probability $g_L(t_1)$, and the MM earns s . The second line shows the case that the first trade is done by the HFT. In this case, the market maker knows that the second trade is with an LT and $v = \sigma$, as the HFT bought the asset at t_1 . From the break-even condition, the spread posted at t_1 for the second trade is

$$s_2(t_1) = \frac{g_L(t_0)g_H(t_1) + g_H(t_0)}{g_L(t_0) [g_H(t_1) + g_L(t_1)] + g_H(t_0)} \sigma.$$

Note that we can write (41) by using (40) as follows.

$$V(s, t_1) = g_L(t_0)V_0(s, t_1) + g_H(t_0)(s - \sigma).$$

Due to the same argument, the MM's expected profit from posting s_3 at time t_2 is written as

$$\begin{aligned} V(s, t_2) &= g_L(t_0) [g_L(t_1) (g_H(t_2)(s - \sigma) + g_L(t_2)s) + g_H(t_1)(s - \sigma)] + g_H(t_0)(s - \sigma) \\ &= g_L(t_0)g_L(t_1)V_0(s, t_2) + [g_L(t_0)g_H(t_1) + g_H(t_0)](s - \sigma). \end{aligned}$$

By induction, for general $k < n$,

$$V(s, t_k) = V_0(s, t_k)\Pi_{j=0}^{k-1}g_L(t_j) + (s - \sigma)\sum_{l=0}^{k-1}g_H(t_l)\Pi_{j=0}^{l-1}g_L(t_j).$$

By substituting V_0 in (40), the equilibrium spread that the MM post at time t_k for the $(k + 1)$ -th trade is

$$s_{k+1} = s(t^k) = \frac{\sum_{l=0}^k g_H(t_l)\Pi_{j=0}^{l-1}g_L(t_j)}{\Pi_{j=0}^k g_L(t_j) + \sum_{l=0}^k g_H(t_l)\Pi_{j=0}^{l-1}g_L(t_j)}\sigma \quad (42)$$

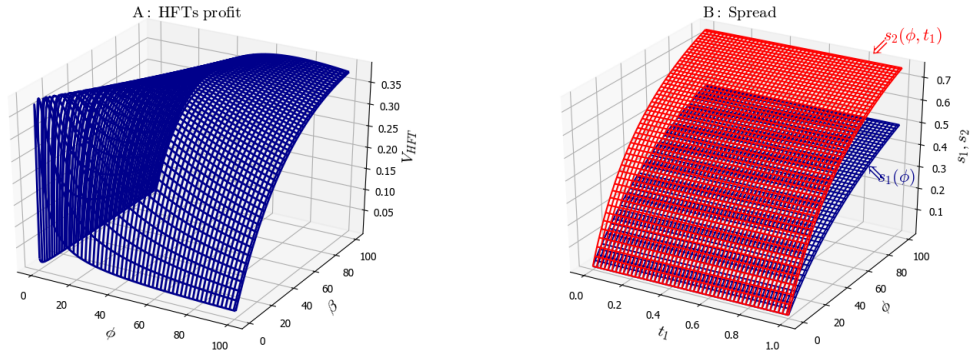
where $t^k = \{t_0, t_1, \dots, t_k\}$ denotes the history of transaction time. It depends on the history because τ_H is not memoryless.

Finally, at $t = t_n$, the market maker posts a LO for the $n + 1$ st trade. However, the fact that the trade happens $n + 1$ times implies that the HFT must have traded or trades in the $n + 1$ st round, telling the market maker the true value of asset $v = \sigma$. Therefore, the equilibrium spread posed at t_n is

$$s_{n+1} = \sigma.$$

Thus, the price converges to the true value of asset at the end of the game. Note that letting $n \rightarrow \infty$ makes the model correspond to the original Glosten-Milgrom model with the HFT arrives only once

Figure 13: HFT's profits and spread



(but it is not tractable because $n \rightarrow \infty$ involves the infinite number of states, $\{t_i\}_{i=0}^{\infty}$).

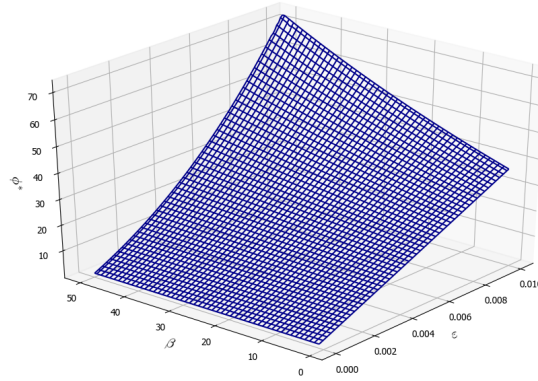
Speed acquisition by the HFT

In what follows, I assume that $C = 0$ because the purpose of this appendix is to show the existence of the crowding-in effect. The HFT knows the structure of the game (i.e., Figure 12). Thus, her expected sniping profit is

$$\begin{aligned}
 W_{HFT}^{gross} = & \overbrace{\int_{t_0}^{\infty} f_H(t_1|t_0) \bar{F}_L(t_1|t_0) (\sigma - s_1(t_0)) dt_1}^{\text{trade in the 1st round}} \\
 & + \underbrace{\int_{t_0}^{\infty} f_L(t_1|t_0) \bar{F}_H(t_1|t_0)}_{\text{forgo the 1st round}} \int_{t_1}^{\infty} \left[\underbrace{f_H(t_2|t_1) \bar{F}_L(t_2|t_1) (\sigma - s_2(t_1))}_{\text{trade in the 2nd round}} + \underbrace{f_L(t_2|t_1) \bar{F}_H(t_2|t_1)}_{\text{forgo the 2nd round}} \int_{t_2}^{\infty} (\dots) dt_3 \right] dt_2 dt_1.
 \end{aligned}$$

The first line represents the case that the HFT trades as the first trader at t_1 . The second line captures the case that an LT trades as the first trader, in which case the HFT repeats the game with the same structure as the first round.

Figure 14: HFT's optimal speed



Analyses

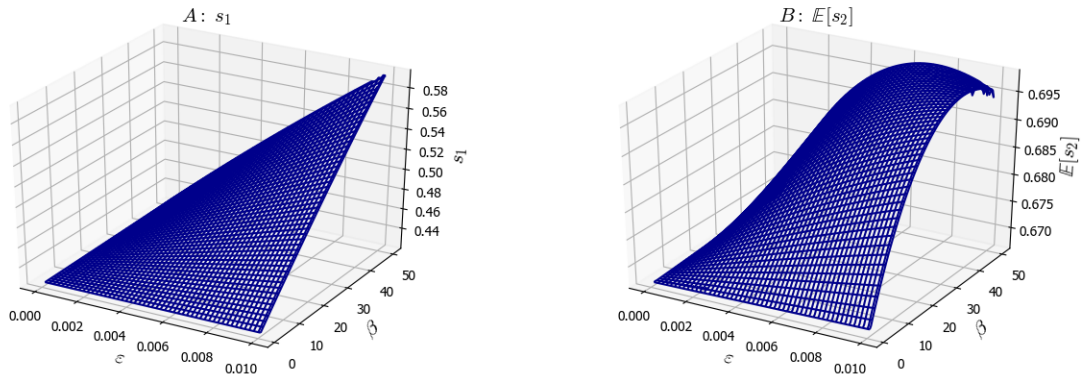
In what follows, I provide numerical examples with $n + 1 = 3$, i.e., trading takes place three times. Firstly, Panel A in Figure 13 plots the expected trading profits for the HFT (W_{HFT}^{gross}) against her speed ϕ and the arrival rate of LTs β . Besides $\mathbb{E}[\delta] = \varepsilon$, which we discuss below, the arrival rate of LTs is the only exogenous parameter of the model. With a wide range of β , Figure 13 shows that W_{HFT}^{gross} takes a hump-shaped curve against ϕ , meaning that the speed-acquisition problem is well defined.

The reason for the hump-shaped W_{HFT}^{gross} is provided by Panel B in Figure 13, which plots the spreads s_1 and s_2 . Note that the spread for the first round trading s_1 is the same as the benchmark model. The spread for the second trade s_2 depends on the timing of the first trade t_1 . As Panel B demonstrates, the second round's spread is an increasing function of the HFT's speed regardless of t_1 . Therefore, the spreads still work as an endogenous cost of being faster for the HFT.

Moreover, from (42), it can be analytically shown that s_{k+1} is higher than s_k , as the market maker updates belief and quote. This implies that the HFT has an incentive to increase her speed to trade in an earlier stage.

Impact of speed bump Figure 14 shows the impact of a speed bump on speed acquisition of the HFT with a wide range of β to check the robustness. Consistent with the benchmark model, a longer speed

Figure 15: Equilibrium spread



bump promotes the speed of the HFT, i.e., the crowding-in effect exists. Note that a large β makes it easier for the HFT to improve her speed, leading to a stronger impact of a speed bump in a high- β region.

Panels A and B in Figure 15 show the reaction of the (expected) spreads, s_1 and $\bar{s}_2 = \mathbb{E}[s_2]$, to a speed bump. Panel A shows that the first-round's spread s_1 is monotonically increasing in ϵ , reflecting the crowding-in effect. In contrast, Panel B demonstrates that the expected second-round's spread, $\bar{s}_2$, takes a hump-shaped reaction to a speed bump when β is sufficiently large. This is because of strong speed acquisition by the HFT: as the HFT becomes faster, she is more likely to arrive as the first trader, reducing the adverse selection cost for the second trade. Together with the direct protection by the speed bump, this negative impact generates the hump-shaped reaction of $\bar{s}_2$.

B Order-by-order speed bumps

The benchmark model analyzes a market-wide random delay, and all orders are delayed by the same amount of time δ . Modeling a speed bump as a market-wide delay follows [Brolley and Cimon \(2017\)](#) and helps derive analytical results and clear intuition. In reality, however, more complicated designs are possible, such as order-basis delays. In this section, I numerically check that the results in the main text are robust even if each order independently draws δ from $\text{Exp}(b)$.

Consider the same environment as the benchmark model. The only exception is that a market order placed by the HFT is delayed by δ_H , while an order from a liquidity trader is delayed by δ_L with $\delta_i \stackrel{\text{iid}}{\sim} \text{Exp}(b)$ and $\varepsilon \equiv b^{-1}$.

The process for the liquidity shock starts at $t = 0$, and liquidity traders arrive at $\tau_L = T_L + \delta_L$ which has density

$$f(\tau_L = \tau) = \frac{b\beta}{b - \beta}(e^{-\beta\tau} - e^{-b\tau}).$$

Also, conditional density is

$$f(\tau_L = \tau | \tau_L > m) = b\beta \frac{e^{-\beta\tau} - e^{-b\tau}}{e^{-\beta m} - e^{-bm}}.$$

Moreover, the HFT's order arrives at $\tau_H = T_H + \delta_H$. For $\tau > t_0$, the density function is

$$f_H(\tau_H = \tau) = \frac{b\phi}{b - \phi} \left(e^{-\phi(\tau - t_0)} - e^{-b(\tau - t_0)} \right)$$

Therefore, at $t = t_0$, the probability that the HFT trades faster than the liquidity traders and learning by the market maker is

$$\begin{aligned} q_H &\equiv \Pr(\tau_H = \tau < \min\{\tau_L, T_M\} | \tau \geq t_0) \\ &= \frac{b\phi}{e^{-\beta t_0} - e^{-bt_0}} \left(\frac{b}{\psi(b + \eta)} e^{-\beta t_0} - \frac{\beta}{(2b + \gamma)(\phi + \gamma + b)} e^{-bt_0} \right). \end{aligned}$$

Also, the probability that a trade with liquidity traders happens before the HFT's arrival and the market maker's learning is

$$\begin{aligned} q_L &\equiv \Pr(\tau_L = \tau < \min\{\tau_H, T_M\} | \tau \geq t_0) \\ &= \frac{b\beta}{e^{-\beta t_0} - e^{-bt_0}} \left(e^{-\beta t_0} \frac{b + \psi}{\psi(b + \beta + \gamma)} - \frac{2b + \gamma + \phi}{(\phi + b + \gamma)(2b + \gamma)} e^{-bt_0} \right). \end{aligned}$$

Therefore, the competitive spread is given by

$$s = \frac{q_H}{q_H + q_L} \sigma = \frac{\phi}{\beta L + \phi} \sigma$$

with

$$L = \frac{e^{-\beta t_0} \frac{b+\psi}{\psi(b+\eta)} - \frac{2b+\gamma+\phi}{(\phi+b+\gamma)(2b+\gamma)} e^{-bt_0}}{\frac{b}{\psi(b+\eta)} e^{-\beta t_0} - \frac{\beta}{(2b+\gamma)(\phi+\gamma+b)} e^{-bt_0}}.$$

Now, consider the HFT's profit. By using q_H , it amounts to

$$W_{HFT}^{gross} = \frac{b\phi}{e^{-\beta t_0} - e^{-bt_0}} \left(\frac{b}{\psi(b+\eta)} e^{-\beta t_0} - \frac{\beta}{(b+\gamma+\phi)(2b+\gamma)} e^{-bt_0} \right) \times (\sigma - s).$$

Figure 16 plots W_{HFT}^{gross} against ϕ and β in Panel A, while the optimal ϕ^* and s^* are shown in Panels B and C. Although the figure is drawn with a couple of parameter values, I conducted exhaustive numerical calculations to check robustness. Specifically, the parameter space for (β, γ) is set to cover $[1.0, 100.0] \times [1.0, 100.0]$ with grid size $\Delta = 0.0001$. For $\varepsilon \in [0, 0.001]$, the result shows that W_{HFT}^{gross} has a unique solution ϕ^* , and both ϕ^* and s^* are monotonically increasing in ε .

C Speed bumps with other distributions

Suppose that $\delta \in [0, \bar{\delta}]$ has cdf $Q(\delta)$. Note that we allow $\bar{\delta} \rightarrow \infty$. From the same argument as equation (31), the break-even spread is given by

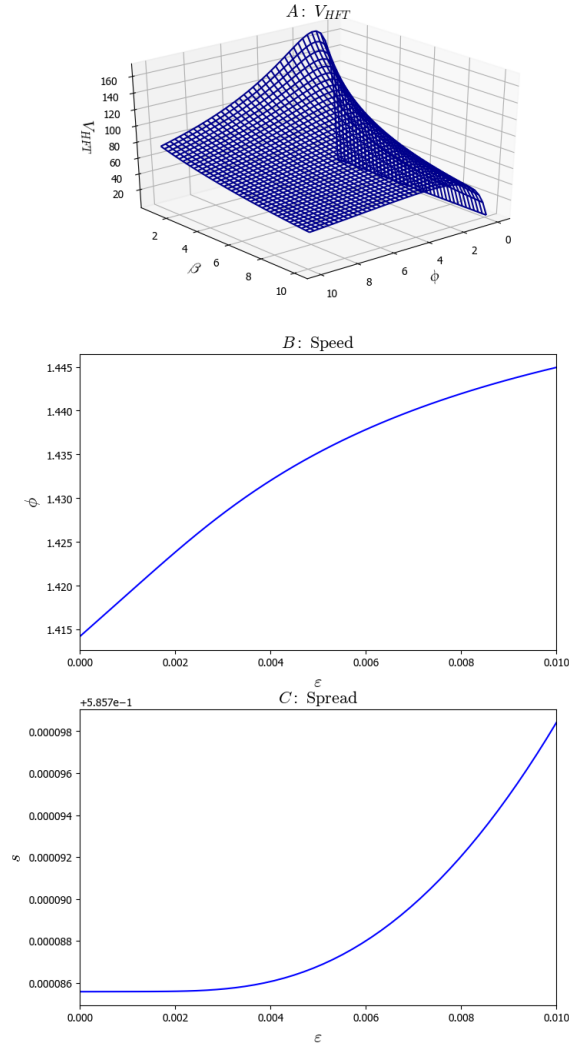
$$s = \frac{\phi}{\beta(1 + \lambda\psi) + \phi} \sigma$$

where

$$\lambda(Q) = \frac{1}{\kappa^{-1} - \gamma} \left(1 - \frac{\mathbb{E}_\delta[e^{-\kappa^{-1}\delta}]}{\mathbb{E}_\delta[e^{-\gamma\delta}]} \right).$$

I define “a longer speed bump” as follows:

Figure 16: Profits, optimal speed, and spread



Definition 1. Consider two speed bumps, δ and $\hat{\delta}$, with cdfs Q and $\hat{Q}$, respectively. Speed bump δ is defined as longer than $\hat{\delta}$ if and only if Q first-order stochastically dominates $\hat{Q}$.

Note that this definition is stronger than the definition of the length measured by $\mathbb{E}[\delta]$.

Suppose that Q first-order stochastically dominates (FOSD) $\hat{Q}$. Also, denote λ as $\lambda(Q)$. Then, the restriction for the distribution Q that makes the generalized model with Q identical to the baseline model is given by

$$\frac{1}{a - \gamma} \frac{\mathbb{E}_Q[e^{-\gamma\delta} - e^{-a\delta}]}{\mathbb{E}_Q[e^{-\gamma\delta}]} > \frac{1}{a - \gamma} \frac{\mathbb{E}_{\hat{Q}}[e^{-\gamma\delta} - e^{-a\delta}]}{\mathbb{E}_{\hat{Q}}[e^{-\gamma\delta}]}$$

where $\mathbb{E}_Q$ denotes the expectation operator with Q . The above inequality may arise from the defi-

inition of the FOSD: for any weakly increasing functions, $u(\delta)$, it holds that $\int u(\delta)dQ > \int u(\delta)d\hat{Q}$. Note that this applies to the denominator, $\mathbb{E}_Q[e^{-\gamma\delta}] < \mathbb{E}_{\hat{Q}}[e^{-\gamma\delta}]$, while the entire inequality cannot be analyzed unless further specifications for Q are provided.

For example, if Q is the uniform distribution, i.e., $Q = U[0, \bar{\delta}]$, then,

$$\lambda = \frac{1}{\kappa^{-1} - \gamma} \left(1 - \gamma\kappa \frac{1 - e^{-\kappa^{-1}\bar{\delta}}}{1 - e^{-\gamma\bar{\delta}}} \right).$$

Note that increasing $\bar{\delta}$ corresponds to a longer speed bump. It is straightforward to check that λ is monotonically increasing in $\bar{\delta}$, leading to the same results in the benchmark model.

D Model with N snipers

One of the natural extensions of the model in Section 3 is to have $N \geq 3$ high-frequency snipers.³⁰ To this end, suppose that there are $N + 2$ HFTs and N of them are randomly selected as high-frequency snipers (HFS). The remaining HFTs serve as HFMs. Note that, N can be thought of as the number of snipers, n_S , multiplied by the number of market orders placed by each sniper, n_M , so that $N = n_S n_M$. The following discussion assumes $n_M = 1$, while $n_M > 1$ can be analyzed with minor modifications.

I denote the set of snipers and market makers as S and M , respectively. Due to the exponential distribution, the snipers' arrival timing (without delays) is given by the exponential distribution with intensity $\phi_S \equiv \sum_{i \in S} \phi_i$. HFMs' behavior is the same as the model in Section 3 except that her limit order is taken by the HFTs with intensity ϕ_S .

³⁰Introducing $M > 2$ high-frequency market makers does not generate additional intuition, as it renders the model closer to the model with competitive market makers in Section 2.

Consider HFT i 's decision on ϕ_i before her role is determined. Her expected profits are given by

$$\begin{aligned} W_i &= \frac{N}{N+2} \sum_{j \in \mathbb{M}} \pi(\phi_i, \phi) (\sigma - \mathbb{E}[s_j]) - C(\phi_i) \\ &= \frac{2N}{N+2} \frac{1}{1+\beta\kappa} \underbrace{\frac{1}{1+\varepsilon\phi} \frac{\phi_i}{\beta + \phi_i + N\phi}}_{= \pi(\phi_i, \phi)} \left(\sigma - \int_{s_0}^{\sigma} s dF(s) \right) - C(\phi_i). \end{aligned} \quad (43)$$

where the second line comes from the symmetric speed by other HFTs (denoted as ϕ). Without a speed bump, the potential probability that HFT i can snipe HFM j 's quote is given by

$$\tilde{\pi}(\phi_i, \phi_j) = \frac{\phi_i}{\beta + \phi_i + \sum_{k \in \mathbb{S}, k \neq i} \phi_k + \phi_j}.$$

By applying the symmetric speed ($\phi_j = \phi$ for all $j \neq i$), we obtain π in (43). Also, by applying $\phi_S = \phi_i + (N-1)\phi$ to the model in Section 3, the expected spread is

$$\mathbb{E}[s] = -\frac{1}{\beta} \frac{\phi_i + (N-1)\phi}{1 + \lambda(\beta + \phi_i + N\phi)} \log s_0$$

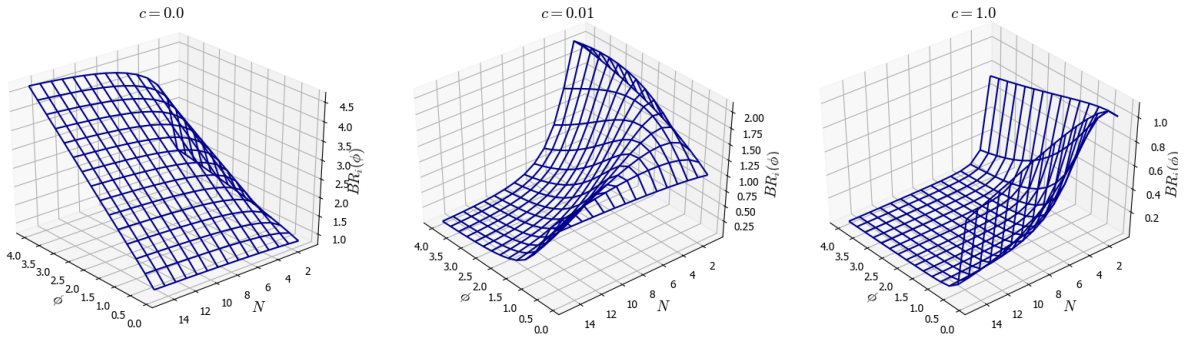
with $s_0 = \frac{\phi_i + (N-1)\phi}{\phi_i + (N-1)\phi + \beta(1 + \lambda(\phi_i + (N-1)\phi + \beta))}$. Moreover, equation (21) for the price discovery process is generalized as follows:

$$\mathbb{E}[\tau] = \frac{1}{(N+2)\phi} \left(1 + \frac{N\varepsilon\phi}{1 + 2\varepsilon\phi} \right).$$

Arms race Figure 17 plots the best response function of HFT i to her rivals' speed ϕ with different values of N and c . First, we observe that the number of HFS (N) has a non-monotonic impact on the arms race. When the cost parameter c is small (resp. large), accommodating more snipers increases (resp. decreases) the best response of HFT i . This result stems from the following property of the spread:

Corollary 3. *The price impact of speed is decreasing in the number of snipers, i.e., $\frac{\partial}{\partial N} \left(\frac{d\mathbb{E}[s]}{d\phi_i} \right) < 0$.*

Figure 17: Best response, speed bump, and N with different cost



Note: The figures are illustrated by using the following parameter values: $\beta = t_0 = 1.0, \varepsilon = 10^{-4}$. Each panel plots the normalized value of variable, e.g., $BR_i(\phi, N)/BR_i(0, 1)$.

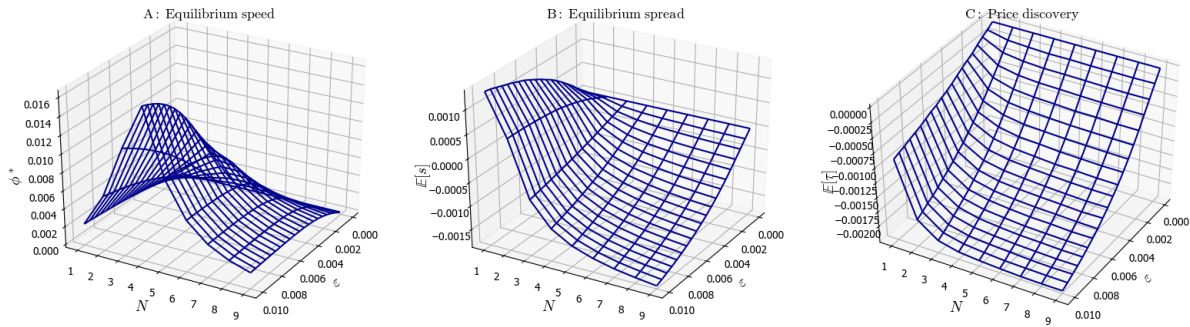
Proof. By letting $\phi_s \equiv \phi_i + (N - 1)\phi$, $\frac{dE[s]}{d\phi_i}$ is given by (34) in Appendix A.3. Taking the partial derivative with respect to N yields the result. $\square$

As a result, having more snipers (N) has two competing effects on HFT i 's speed acquisition. Firstly, it diminishes her incentive to be faster because the impact of ϕ_i on HFT i 's sniping probability π is dwarfed by a large N . Namely, stiffer competition between snipers makes each HFT reluctant to invest in speed technologies. Secondly, however, N alleviates the price impact of HFT i 's speed-up, thereby promoting her investment into speed.

Figure 17 demonstrates that there is a threshold N^* , and the first negative impact via stiffer competition becomes dominant when N exceeds N^* . This is intuitive because the first negative impact becomes more salient when N is large, π is small, and the effective exogenous cost of being faster (C/π) is large. The same logic explains why the best response function negatively reacts to N when the cost parameter c is large.

Equilibrium speed and market quality Figure 18 shows the equilibrium speed (Panel A), the expected spread (Panel B), and the measure of price discovery (Panel C). Firstly, the reaction of equilibrium variables to a speed bump (ε) that we establish with $N = 1$ is robust. When c is small, a longer delay promotes speed acquisition of HFTs, widens the expected spread, and generates the U-shaped

Figure 18: Equilibrium impact of speed bump and N



Note: The figures are illustrated by using the following parameter values: $\beta = t_0 = 1.0$. Each panel plots the normalized value of variable, e.g., $\phi^*(\epsilon, N)/\phi^*(0, 1)$.

Table 2: 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.

curve for the price discovery process, while these results revert when c becomes large. Somewhat more surprising is the impact of N on speed acquisition in Panel A. Due to the effect in Corollary 3, stiffer competition can promote the equilibrium speed acquisition when the number of sniper N is relatively small. Once again, the threshold N^* is the determinant of the above result. As it depends on the deep parameters of the model, such as β and t_0 , the value of N (or N^*) in Figure 18 does not directly translate into reality.

E Share of high-frequency trading firms

My model takes into account traders' strategic motives in their speed decision (i.e., they internalize the impact of their speed decision on a market). The strategic motive arises because major high-

speed financial institutions have significant shares in the trading volume in markets. For example, Table 2 shows the top five high-frequency financial institutions and their shares in the BrokerTec platform, through which more than half of the U.S. Treasury is traded. When an institution decides a speed technology, she becomes aware of the market impact of her choice because a sizable number of transactions (as in Table 2) are affected by the speed technology and may impact the equilibrium price.

Also, 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).

F Additional proof

F.1 Proof of Corollary 1

The cross-derivative of the FOC, evaluated at the optimal speed ϕ^* , is reduced to

$$\begin{aligned}\frac{\partial}{\partial \lambda} \left(\frac{dW_{HFT}}{d\phi} \right) |_{\phi=\phi^*} &= 2\beta^2 \phi \frac{1+\lambda\eta}{K^3} - \gamma \frac{d\varepsilon}{d\lambda} C'(\phi) \\ &= 2\phi \left[\beta^2 \frac{1+\lambda\eta}{(\phi + \beta(1+\lambda\psi))^3} - c\gamma \frac{d\varepsilon}{d\lambda} \right]\end{aligned}\tag{44}$$

where $K \equiv \phi + \beta(1 + \lambda\psi)$, and the second line uses the quadratic cost function.

The first term of (44) is bounded from above, and there exists $L < \infty$ such that

$$\forall c, L > \beta^2 \frac{1+\lambda\eta}{(\phi + \beta(1+\lambda\psi))^3}.$$

In contrast, $c\gamma \frac{d\varepsilon}{d\lambda}$ is monotonically increasing in c and not bounded from above. Therefore, there

exists $\bar{c}$, such that

$$c > \bar{c} \Rightarrow c\gamma \frac{d\varepsilon}{d\lambda} > L \Leftrightarrow \frac{d\phi^*}{d\lambda} < 0.$$

Next, we know from the FOC that $\phi^*|_{c=0} < \infty$, as long as $\lambda < \sqrt{\beta(\beta + \gamma)}$. Therefore, the first of (44) term is bounded from below, meaning that there exists $\underline{c}$ such that

$$c < \underline{c} \Rightarrow \frac{d\phi^*}{d\lambda} > 0.$$