Order Flow Segmentation, Liquidity and Price Discovery: The Role of Latency Delays[*]

Michael Brolley[†] David Cimon[‡]
Wilfrid Laurier University Bank of Canada

May 11, 2017

— preliminary draft —

Abstract

Latency delays—known as "speed bumps"—are an intentional slowing of order flow by exchanges. Supporters contend that delays protect market makers from high-frequency arbitrage, while opponents warn that delays promote "quote fading" by market makers. We construct a model of informed trading in a fragmented market, where one market operates a conventional order book, and the other imposes a latency delay on market orders. We show that informed investors migrate to the conventional exchange, widening the quoted spread; the quoted spread narrows at the delayed exchange. The overall market quality impact depends on the nature of the delay: "short" latency delays lead to improved trading costs for liquidity investors, but worsening price discovery; sufficiently "long" delays improve both.

[*]The authors would like to thank Corey Garriott, Andriy Shkilko, and Adrian Walton for valuable discussions. Michael Brolley gratefully acknowledges the financial support from the Social Sciences and Humanities Research Council. The views of the authors are not necessarily those of the Bank of Canada. All errors are our own.

[†]Email: mbrolley@wlu.ca; web: http://www.mikerostructure.com.

[‡]Email: dcimon@bank-banque-canada.ca; web: https://sites.google.com/site/dcimon.

"I am personally wary of prescriptive regulation that attempts to identify an optimal trading speed, but I am receptive to more flexible, competitive solutions that could be adopted by trading venues."

—SEC Chair, Mary Jo White, June 5, 2014

Liquidity suppliers prefer to transact against uninformed traders. These uninformed traders are valuable, as they are unlikely to move the market against market makers. Many exchanges, competing for scarce order flow, have specialized to attract these uninformed liquidity demanders. Inverse pricing, dark trading and retail order segmentation facilities have all been studied as ways in which exchanges try to draw these traders from other markets, in part by advertising their market design as a way to disincentivize informed traders from also participating. Recently, some exchanges have imposed latency delays—so-called "speed bumps"—as yet another way of segmenting away retail order flow. Measured on the order of milliseconds or microseconds, latency delays impose a time delay between an order's receipt at the exchange, and its execution.[1] Exchanges advertise latency delays as means of protecting market makers from adverse selection at the hands of high frequency traders (HFTs) acting on extremely short-horizon information; the savings are then passed on through a narrower spread.[2]

As with any market structure change, latency delays have not been without controversy. Beyond the comments from proponents, who tout improved market quality, other market participants have suggested that delays in order execution create an uneven playing field by allowing market makers to "fade" quotes ahead of large orders.[3] Quote "fading" refers to a market maker's ability to revise their quotes after an order is received, but not yet filled. By fading quotes, market makers execute incoming orders at less favourable prices than at the time of initial submission. Indeed, existing evidence from the academic community suggests

[1] A description of the mechanics behind latency delays is available in the appendix.

[2] For one example see "Regulators Protect High-Frequency Traders, Ignore Investors" in Forbes: `http://www.forbes.com/sites/jaredmeyer/2016/02/23/sec-should-stand-up-for-small-investors/\#1c96d49a1ec6`

[3] For one example see "Canada's New Market Model Conundrum" by Doug Clark at ITG: `http://www.itg.com/marketing/ITG_WP_Clark_Alpah_Conundrum_20150914.pdf`

that, not only do latency delays allow market makers to withdraw liquidity, but they may harm liquidity at other markets, by concentrating retail order flow at a single venue. (Chen, Foley, Goldstein, and Ruf 2016)

To resolve these competing explanations, we construct a static, three-period model of sequential trading. In our model, trading occurs in fragmented markets, where one exchange imposes a latency delay. We model traders who are aware of the possibility of an information event, which occurs with some probability in the second period. This information event can be interpreted in two ways, both as a scheduled event such as an earnings announcement or as a fleeting arbitrage opportunity. In the first case, traders are aware of an announcement at fixed point in the second period, while in the second case, traders are aware that arbitrage opportunities become public knowledge at a fixed point in the second period.[4]

In the first period, traders can choose to submit orders to one of two exchanges. One is a standard exchange, which executes orders immediately in the first period. The second is a latency-delayed exchange, which randomly executes orders either immediately in the first period, or after the information event becomes public in the second period.

To differentiate the effects of the delay on different traders, we model two types of traders, uninformed liquidity traders, and informed speculators. Liquidity traders arrive at the market with a need to trade, and have the choice between either submitting an order immediately, or waiting until after the information event. A liquidity trader who chooses to submit an order before the information event may send the order to either the open exchange, which executes instantly, or the speed bump, which delays the order with some probability. Liquidity traders who delay their order risk paying a form of delay cost, should the market move against them. This delay cost represents the need for these traders to seek additional capital, should prices become worse.

Alternatively, if a speculator arrives, they have the option for paying to become informed before the announcement. Similar to liquidity traders, speculators have the option to either

[4]The latter interpretation is similar in many respects to Budish, Cramton, and Shim (2015), who document the fleeting nature of arbitrage opportunities between New York and Chicago.

execute their order immediately at the non-delayed exchange, or submit their order to the delayed-exchange, and risk having the information event arrive before their order executes.

We relate the "length" of a latency delay to the probability that a known (or expected) information event occurs between the trader's order submission, and its execution. In this way, we assume that the delayed exchange imposes a delay that is randomly drawn within a fixed interval, such that private information becomes public within the latency delay with some (expected) probability. Our notion of a latency delay variation has two interpretations, a "longer" delay implies that either: i) the distribution of the random delay widens, or, ii) the time before the public announcement has been reduced.

We show that as the length of the latency delay increases, informed investors migrate away from the latency delayed-exchange. We use this migration to define our results in terms of a "segmentation point". The segmentation point is the delay length at which the speculator with the highest marginal utility for the delayed exchange migrates away from the delayed exchange. As the delay length increases towards the segmentation point, uninformed investor participation at the delayed exchange increases, reaching a maximum at the segmentation point. At the non-delayed exchange, there is a net migration of informed investors, and a net emigration of uninformed investors. The result is a wider quoted spread at the non-delayed exchange; moreover, some speculators who would acquire information in a setting with no delayed market, choose not to become informed.

Once the segmentation point is reached, further increases in the delay create very different results. Spreads at the latency-delayed exchange improve no further, as all informed traders have left the exchange, while uninformed traders continue to incur larger delays. As a result, uninformed traders begin to return to the non-delayed exchange improving bid-ask spreads at the non-delayed exchange and increasing informed trader participation. For a delay of sufficient length, the non-delayed exchange reverts to the conditions present before a latency delay was imposed, while the latency-delayed exchange contains only uninformed traders who previously did not participate in the market prior to the resolution of the information

event.

We make several empirical predictions regarding latency delays. First, we predict that initial prices should improve at the delayed-exchange while they should be worse at standard exchanges. Second, we predict market segmentation effects between exchanges. Liquidity trader participation should increase following the introduction of a delay, and their trading should be concentrated at this exchange. Informed participation should fall following the introduction of a delay, and their trading should be concentrated at standard exchange. Finally, we show that, latency delays have ambiguous effects on price discovery, depending on the length of the delay. Particularly, small latency delays decrease price discovery measures, while simultaneously increasing spreads at other exchanges.

Related Literature. While there is little existing literature on the topic of latency delays, the factors which have led to their creation have been well documented. The first group of relevant literature studies high frequency trading, and its effects on markets. Predatory high frequency trading is generally cited as the rationale for the use of speed bumps and, as such, is essential to understanding their purpose. The second group of literature covers both the drivers and effects of market fragmentation. As a means for exchanges to differentiate themselves, speed bumps can be discussed within this general trend of market fragmentation and competition between exchanges.

As latency delays are on the order of milliseconds or less, market makers who are able to make use of them in a strategic manner are inherently high frequency traders. Several studies of high frequency market makers have shown that they can improve liquidity (Brogaard and Garriott 2015, Brogaard, Hagströmer, Nordén, and Riordan 2015, Subrahmanyam and Zheng 2015). However, work on high frequency liquidity demanders finds that they may increase price efficiency (Carrion 2013) but also increase transaction costs (Chakrabarty, Jain, Shkilko, and Sokolov 2014). High frequency traders have also been shown to improve price discovery through both liquidity supply (Brogaard, Hendershott, and Riordan 2015, Conrad, Wahal, and Xiang 2015) and demand (Brogaard, Hendershott, and Riordan 2014).

Proponents argue that latency delays can curb "predatory" behaviours by high frequency traders, such as inter-market arbitrage. However, critics have suggested that latency delays may also lead to quote fading. Latza, Marsh, and Payne (2014) do not find evidence of predatory quote fading behaviour by HFTs, while Malinova and Park (2016) find that it does occur.[5] Our model confirms some forms of quote fading found in the empirical literature. While we do not allow market makers to fade quotes arbitrarily, we model market makers who may fade quotes in response to new information on the underlying asset value. We show that the ability to update quotes before an order arrives may allow market makers to quote at better initial prices.

Theoretically, the role of HFTs has been studied in a variety of contexts including their role in market-making (Jovanovic and Menkveld 2011), arbitrage (Wah and Wellman 2013), and the incorporation of new information (Biais, Foucault, and Moinas 2015).[6] Menkveld and Zoican (2016) model the effects of known latency within a single exchange, versus latency in reaching the exchange, a friction similar to an intentional latency delay. We complement the existing theoretical work on HFTs by modeling both intentional, randomized delays within exchanges as well as investor migration between exchanges, based on these delays. Further to previous literature, investors base their exchange choice not only with whether other market participants are delayed, but also on whether a delay at one exchange will remove their informational advantage.

The topic of market segmentation is not new within the academic literature. Existing empirical work has found that fragmented markets can have improved liquidity (Foucault and Menkveld 2008) and efficiency (Ye and O'Hara 2011). Additional work by Kwan, Masulis, and McInish (2015) and Gomber, Sagade, Theissen, Weber, and Westheide (2016) studies the use of both dark trading, and other mechanisms, in order to attract order flow.[7] As

[5]Related work by Ye, Yao, and Gai (2013) find evidence of a different behaviour known as quote "stuffing", which we do not address in this paper

[6]A further survey is topics surrounding HFT is present in both Angel, Harris, and Spatt (2011) and O'Hara (2015).

[7]Further theoretical work by Baldauf and Mollner (2016) shows that the net effects of increased fragmentation are ambiguous for liquidity suppliers.

latency delays are another means of attracting order flow, our work confirms the concept of segmentation and suggests additional avenues for empirical market segmentation work.

Existing theoretical work studies the choice of market based on fees (Colliard and Foucault 2012), dark liquidity (Zhu 2014), and the profitability of financial intermediaries (Cimon 2016). We extend existing work by modeling market segmentation based on differences in speed. Taken together with these earlier contributions, our work helps complete the set of factors which may influence market choice by financial system participants.

The closest work to ours is Chen, Foley, Goldstein, and Ruf (2016) who empirically study the introduction of a speed bump on TSX Alpha, a Canadian trading venue. They find that, following the introduction of a speed bump, total volume on the affected exchange decreases. High frequency traders provided a greater proportion of liquidity, compared to non-high frequency traders when the speed bump was in place. Adverse selection on the affected exchange also decreased. For all other exchanges, informed trading increased, leading to wider quoted and effective spreads.

The paper proceeds as follows. Section 1 outlines the model. Section 2 presents a benchmark model of two identical (fragmented) markets with no latency delay, and then extends it to consider the case where one exchange may impose a latency delay on incoming orders. In Section 3, we present empirical and policy predictions. Section 4 concludes.

1 The Model

Security. There is a single risky security with an unknown random payoff v that is equal to $v_0 - \sigma$ or $v_0 + \sigma$, with equal probability, where $\sigma \in (0, 1)$. The security is available for trading at $t = 1$ and $t = 2$. The security's value is publicly announced at $t = 2$ before trading begins. The asset is liquidated at $t = 3$.

Market Organization. There are two exchanges, Fast and Slow, that operate as displayed limit order books: posted limit orders display their quotes to all market participants. Market

orders sent to Exchange Fast fill immediately upon receipt, whereas exchange Slow fills market orders with a random delay. With probability $\delta \in (0,1)$, an order sent to exchange Slow is delayed until $t = 2$, and filled after the announcement of v.[8] Otherwise, the order is filled immediately.

There are two interpretations for this type of latency delay. First, a latency delay of this type can reflect a setting where investors expect an incoming information event (a scheduled announcement), though some investors may not be informed about its direction and magnitude. Alternatively, this type of latency delay can reflect the presence of fleeting arbitrage opportunities at other markets. Speculators who acquire information can be viewed as acquiring the necessary technology to exploit these opportunities. The random nature of the speed bump then represents the fact that, with a delay of any length, speculators may no longer be the first to trade.

Exchange Market Maker. A competitive market maker supplies buy and sell limit orders to both exchanges before investors submit their orders at $t = 1$ and $t = 2$. The market maker is risk-neutral, and receives only the public information, v_0, about the security's fundamental value. The market maker has a zero latency, permitting them to place (and update) limit orders to both exchanges at the beginning of periods $t = 1$ and $t = 2$, before investors place their orders. At $t = 2$, upon the announcement of v, the market maker updates their $t = 1$ limit orders to the public value, v.

The exogenous separation of market makers matches an important feature of latency-delayed venues. In general, orders are delayed, with the exception of orders used for market making purposes. On some venues, this consists of orders pegged at or near the midpoint, while on others it consists of large orders, above a certain size, providing liquidity. Thus, it is generally insufficient to merely submit a limit order to bypass the delay.

Investors. There is a unit mass of risk-neutral investors. At $t = 0$, an investor arrives at the market to trade a single unit of the security. The investor is either a speculator with

[8] A random delay is similar in nature to the latency delay imposed by TMX Alpha, a Canadian trading venue. TMX Alpha delays orders by a random time period of 1-3ms.

probability $\mu > 0$, or an uninformed investor endowed with liquidity needs. Upon arrival, a speculator receives an information acquisition cost γ_i that is distributed uniformly on $[0, 1]$. Speculators may pay γ_i at $t = 0$ to perfectly learn the random payoff v. We refer to those that acquire information as "informed investors", and their mass is denoted $\mu_I \in (0, \mu)$.

With probability $(1 - \mu)$, a liquidity investor arrives. Liquidity investors have no private information, but are endowed with a liquidity need that motivates them to trade. They also pay an additional cost to trade following an adverse price movement that is proportional to the innovation, $c_i = k\lambda_i\sigma$, where $k \in (0, \infty)$. λ_i is a private scaling parameter of the innovation that is distributed uniformly on $[0, 1]$. This cost represents the cost uninformed investors pay to acquire additional capital to trade when the price moves away from them. As this represents a re-capitalization cost, liquidity investors pay this cost only if the price moves against them, not if it moves in their favour. [9] The uninformed investor also pays a constant delay cost $K \in (\sigma, \infty)$ if they elect not to trade. Liquidity investors are buyers or sellers with equal probability.

An investor i may submit a single market order at $t = 1$ or $t = 2$, or not trade. Investors place orders to maximize (expected) profits. Finally, the structure of the model is known to all market participants. The model timeline is illustrated in Figure 1.

Investor Payoffs. The expected payoff to an investor who submits a buy order at $t = 1$ is given by their knowledge of the true value of v, minus the price paid, any information acquisition or delay costs incurred. We denote liquidity investors as L, and informed investors as I. The expected payoffs to investor $i \in \{I, L\}$ submitting an order to exchange $j \in \{\mathsf{Fast}, \mathsf{Slow}\}$ are given by:

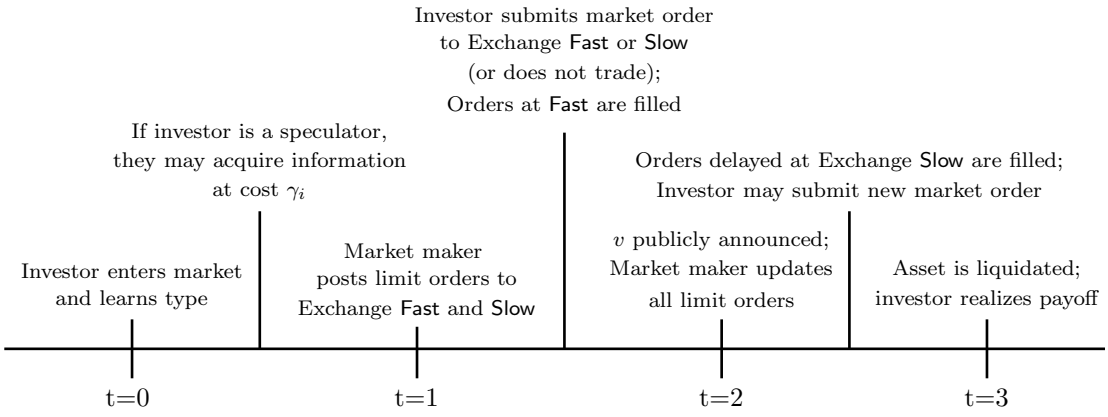
$$\pi_I(\gamma_i;\ \text{Buy at t=1}) = v - \mathsf{E}[\mathsf{ask}_1^j \mid \text{submit at exchange J}] - \gamma_i \tag{1}$$

$$\pi_L(c_i;\ \text{Buy at t=1}) = v_0 - \mathsf{E}[\mathsf{ask}_1^j \mid \text{submit to exchange J}] - \mathsf{Pr}(\text{order delay}) \times \frac{c_i}{2} \tag{2}$$

[9] We concede that a price movement can occur in a beneficial direction, and that the investor could earn a reinvestment return on the proceeds. We assume that the recapitalization cost exceeds the reinvestment return, and as such, normalize the reinvestment return to zero.

Figure 1: Model Timeline

This figure illustrates the timing of events upon the arrival of an investor at $t = 0$, until their payoff is realized at $t = 3$. Speculators face information acquisition costs γ_i, and liquidity investors face delay cost c_i.



The scaling factor of 1/2 in the delay cost of π_L reflects the fact that the asymmetric cost is only incurred if the price moves away from the liquidity investor, which occurs with probability 1/2. An investor i who submits a buy order at period $t = 2$ (or elects not to trade) has a payoff of $-\gamma_i$ if informed (speculators have payoff zero); uninformed investors earn a payoff to not trading of $-K < -\sigma$. Seller payoffs are similarly defined.

2 Equilibrium

In this section, we present two versions of our model: first, we outline a benchmark case where both exchanges are identical (no latency delay), and then subsequently compare our results to a model where Exchange Slow imposes a latency delay.

We search for a perfect Bayesian equilibrium in which the market maker chooses a quoting strategy such that they earn zero expected profits at each venue, and investors choose order submission strategies that maximize their profits. We also search for equilibria where investors use both exchanges. We study the impact of a processing delay at one exchange by comparing it to a case where both exchanges do not impose a processing delay; we refer to this as the *benchmark case*. In effect, a market with two identical exchanges is equivalent

to a single competitive exchange. Because the set-up of our model is symmetric for buyers and sellers, we focus our attention to the decisions of buyers, without loss of generality.

2.1 Identical Fragmented Markets (No Latency Delay)

In the exposition that follows, although both exchanges fill orders without delay, we continue to denote them as Exchange Fast and Slow, to maintain consistency in notation. If both exchanges impose no processing delay ($\delta = 0$), then investors' payoffs simplify considerably. Because any orders submitted to either exchange will be filled at the posted quote, investors who submit orders suffer no risk of the quote updating adversely. Speculator and liquidity investor payoffs to trading on an Exchange j are reduced to:

$$\pi_I(\gamma_i; \text{Buy at t=1}) = v - \mathsf{ask}_1^j - \gamma_i \tag{3}$$

$$\pi_L(c_i; \text{Buy at t=1}) = v_0 - \mathsf{ask}_1^j \tag{4}$$

Note that because a market buy order is filled immediately at the posted quote, the expected profit for a liquidity investor who submits a market buy order at $t = 1$ does not consider c_i directly; instead, the cost of c_i is considered when choosing whether to trade at $t = 1$, or wait until uncertainty is resolved at $t = 2$ (for which they pay c_i).

Given an expectation of investors' order submission strategies, the market maker populates the limit order books at exchanges Fast and Slow. The market maker quotes competitively, setting the ask (and bid) prices at $t = 1$ on exchange Fast and Slow—which we denote $\mathsf{ask}_1^{\mathsf{Fast}}$ and $\mathsf{ask}_1^{\mathsf{Slow}}$, respectively—to account for the expected adverse selection of an incoming buy order:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at Exchange } \mathsf{Fast}] \tag{5}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = \mathsf{E}[v \mid \text{Buy at Exchange } \mathsf{Slow}] \tag{6}$$

Prices $\mathsf{bid}_1^{\mathsf{Fast}}$ and $\mathsf{bid}_1^{\mathsf{Slow}}$ are analogously determined through symmetry of buyers and sellers.

At period $t = 2$, v is announced, and the market maker updates their buy orders on both exchanges to $\mathsf{ask}_2^{\mathsf{Fast}} = \mathsf{ask}_2^{\mathsf{Slow}} = \mathsf{bid}_2^{\mathsf{Fast}} = \mathsf{bid}_2^{\mathsf{Slow}} = v$.

Each investor makes two decisions: whether to participate in the market at $t = 1$ (or at all), and if they participate, to which exchange should they submit an order. A speculator receives their information acquisition cost γ_i at $t = 0$, and weighs it against the expected profit of becoming informed. If they acquire information, they subsequently decide to which exchange they will submit an order. Similarly, liquidity investors receive their delay cost c_i at $t = 0$, and choose whether to delay trading to $t = 2$. If they decide to trade at $t = 1$, they choose to which exchange they submit an order.

We characterize these decisions via backward induction. At $t = 2$, speculators (informed and otherwise) have no information advantage, and thus their expected profit is zero. Liquidity investors who did not submit an order at $t = 1$ submit an order to either exchange at $t = 2$ and pay cost c_i. It is always optimal for a liquidity investor to submit an order at $t = 2$, as the cost to abstaining, $K > \max\{c_i\}$.

At $t = 1$, speculators who do not acquire information at $t = 0$ do not trade. If a speculator has chosen to acquire knowledge of v, the now-informed investor knows that delaying until period $t = 2$ is unprofitable, so they choose the optimal exchange to which they submit their order. We denote the probability with which an informed investor submits an order to Exchange Fast as $\beta \in (0, 1)$; they submit an order to Exchange Slow otherwise. Because γ_i only dictates the decision to acquire information, and doesn't factor directly into the venue choice, informed investors use a mixed strategy in β such that they earn a equal payoff at both exchanges. Similarly, a liquidity investor that chooses to trade in $t = 1$ finds that their venue choice is not directly impacted by c_i; they also submit orders to both venues with a mixed strategy, where we denote probability of submitting an order to Exchange Fast as

$\alpha \in (0,1)$, and Exchange Slow otherwise. Buyers' order choice indifference conditions are:

$$\text{Informed Buyer: } \left\{\beta \mid \pi_I^{\mathsf{Fast}}(\text{Buy t=1}) = \pi_I^{\mathsf{Slow}}(\text{Buy t=1}) \iff \mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{7}$$

$$\text{Liquidity Buyer: } \left\{\alpha \mid \pi_L^{\mathsf{Fast}}(\text{Buy t=1}) = \pi_L^{\mathsf{Slow}}(\text{Buy t=1}) \iff \mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{8}$$

We note here that, in the absence of direct impacts by γ_i and c_i, the sole determinant of venue choice for buyers are the ask prices (and similarly bid prices for sellers). If quotes are not equal across both exchanges, then (α, β) cannot be an equilibrium, as there would be migration from the high-priced exchange to the lower priced exchange until prices across both exchanges equate.

Given the venue choice strategies for informed and liquidity investors, the ask prices quoted by the market maker at $t = 1$ can now be characterized as:

$$\mathsf{ask}_1^{\mathsf{Fast}} = v_0 + \frac{\mathsf{Pr}(\text{informed trade at } \mathsf{Fast})}{\mathsf{Pr}(\text{trade at } \mathsf{Fast})} \cdot \sigma \tag{9}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = v_0 + \frac{\mathsf{Pr}(\text{informed trade at } \mathsf{Slow})}{\mathsf{Pr}(\text{trade at } \mathsf{Slow})} \cdot \sigma \tag{10}$$

Liquidity investors are buyers or sellers with equal probability, so only half of liquidity investors who choose to participate in the market at $t = 1$ will buy, independent of the realization of v. Sell prices $\mathsf{bid}_1^{\mathsf{Fast}}$ and $\mathsf{bid}_1^{\mathsf{Slow}}$ are similarly characterized.

Given α and β, investors make participation decisions at $t = 0$ that characterize the measure of speculators, μ_I and the measure of liquidity investors that participate before $t = 2$, which we denote $\Pr(c_i \geq \underline{c})$. That is, all investors with $c_i \geq \underline{c}$ face a large enough cost of delay c_i, such that they trade prior to period $t = 2$. Speculators receive γ_i in period $t = 0$, and decide whether paying their information acquisition cost is profitable. The mass of speculators that choose to acquire information determines μ_I. To find μ_I, we find the value of γ_i at which a speculator is indifferent to acquiring information and not trading. This is

equal to γ_i such that a speculator earns a zero expected profit from becoming informed:

$$\bar{\gamma} = \max\left\{v - \mathsf{ask}_1^{\mathsf{Fast}}, v - \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{11}$$

Hence, any speculator with $\gamma_i \leq \bar{\gamma}$ will acquire information, and the mass of informed investors at $t = 1$ is equal to: $\mu_I = \mu \times \mathsf{Pr}\left(\gamma_i \leq \bar{\gamma}\right)$. Similarly, we characterize the measure of liquidity investors that participate in the market at $t = 1$, $\Pr(c_i \geq \underline{c})$ by:

$$\underline{c} = \min\left\{v_0 - \mathsf{ask}_1^{\mathsf{Fast}}, v_0 - \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{12}$$

Therefore, any liquidity investors with a delay cost greater than $\underline{c}$ choose to trade at $t = 1$. The probability that such a liquidity investor arrives is given by $(1 - \mu) \times \mathsf{Pr}\left(c_i \geq \underline{c}\right)$.

An equilibrium in our model is characterized by: (i) investor participation measures, μ_I and $(1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right)$; (ii) investor venue strategies, α and β, and; (iii) market maker quotes at $t = 1$ for each exchange $j \in \{\mathsf{Fast}, \mathsf{Slow}\}$, ask_1^j and bid_1^j. These values solve the venue choice indifference equations (7)-(8), the market maker quoting strategy (9)-(10), and the investor participation conditions, (11)-(12).

Theorem 1 (Identical Fragmented Markets) *Let $\delta = 0$. Then for any $\beta \in (0, 1)$, there exists a unique equilibrium consisting of participation constraints $\mu_I \in (0, \mu)$, $\underline{c} \in [0, \frac{k\sigma}{2}]$ that solve (11)-(12), prices $\mathsf{ask}_1^{\mathsf{Fast}}, \mathsf{ask}_1^{\mathsf{Slow}}, \mathsf{bid}_1^{\mathsf{Fast}}$ and $\mathsf{bid}_1^{\mathsf{Slow}}$ that satisfy (9)-(10), and $\alpha \in (0, 1)$ that solves (7)-(8) such that $\beta = \alpha$.*

Theorem 1 illustrates that, in equilibrium, identical fragmented markets may co-exist, and moreover, they need not attract the same level of order flow, despite offering identical prices. For example, in an equilibrium where $(\alpha, \beta) = (3/4, 3/4)$, Exchange Fast receives three times the order flow of Exchange B, but because $\alpha = \beta$, these probabilities cancel out of the pricing equations (9)-(10), ensuring that the ask (and bid) prices of Exchange Fast and Slow are equal. We summarize this in the Corollary below.

Corollary 1 (Equilibrium Prices) *In equilibrium, ask and bid prices at* $t = 1$ *are equal to* $\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}} = v_0 + \frac{\mu_I}{\mu_I + (1-\mu)\mathsf{Pr}(c_i \geq \underline{c})} \cdot \sigma$ *and* $\mathsf{bid}_1^{\mathsf{Fast}} = \mathsf{bid}_1^{\mathsf{Slow}} = v_0 - \frac{\mu_I}{\mu_I + (1-\mu)\mathsf{Pr}(c_i \geq \underline{c})} \cdot \sigma$

In what follows, we define the identical fragmented market formulation of our model ($\delta = 0$) as the benchmark case. We denote the equilibrium solutions with the superscript BM (i.e., $\mathsf{ask}^{\mathsf{BM}}$, $\mathsf{bid}^{\mathsf{BM}}$).

2.2 Slow Exchange Imposes a Latency Delay

In this section, we examine the case where Exchange Slow fills investor orders with a random processing delay, such that orders sent to Exchange Slow are filled before $t = 2$ with probability $\delta \in (0, 1)$. The processing delta impacts payoffs to informed and liquidity investors differently. Informed investors face payoffs to Exchange Fast and Slow:

$$\pi_I^{\mathsf{Fast}}(\gamma_i; \text{Buy at t=1}) = v - \mathsf{ask}_1^{\mathsf{Fast}} - \gamma_i \tag{13}$$

$$\pi_I^{\mathsf{Slow}}(\gamma_i; \text{Buy at t=1}) = v - (1-\delta) \times \mathsf{ask}_1^{\mathsf{Slow}} - \delta \cdot v - \gamma_i \tag{14}$$

By submitting an order to Exchange Slow, informed investors face the possibility of losing their informational advantage. Liquidity do not know v, however, so their expectation of what the announcement of the true value will be is always v_0, and thus the processing delay does not impact their expectation of the future value when buying. Instead, the uncertainty about the outcome of the price manifests in an asymmetrical cost to trading, c_i, that they incur if the price moves in the direction of their desired trade ($v > \mathsf{ask}_1^{\mathsf{Slow}}$). The payoffs to liquidity investors then simplify to:

$$\pi_L(c_i; \text{Buy at t=1}) = v_0 - \mathsf{ask}_1^{\mathsf{Fast}} \tag{15}$$

$$\pi_L(c_i; \text{Buy at t=1}) = (1-\delta)(v_0 - \mathsf{ask}_1^{\mathsf{Slow}}) - \delta \cdot \frac{k\lambda_i}{2} \times \sigma \tag{16}$$

Taking this into account, the market maker sets its prices at $t = 1$ in the following way:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = \frac{\beta\mu_I}{\beta\mu_I + \mathsf{Pr}(\text{uninformed trade at } \mathsf{Fast})} \cdot \sigma \tag{17}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = \frac{(1-\beta)\mu_I}{(1-\beta)\mu_I + \mathsf{Pr}(\text{uninformed trade at } \mathsf{Slow})} \cdot \sigma \tag{18}$$

In period $t = 2$, the value v is publicly announced, so the market maker updates its prices to $\mathsf{ask}_2^{\mathsf{Fast}} = \mathsf{ask}_2^{\mathsf{Slow}} = v$.

When Exchange Slow imposes a processing delay, investors weigh the cost of trading on Exchange Fast immediately, against possibility of a) losing (all or part of) their information if they are informed, or b) paying a higher cost to acquire capital to complete their trade if they are a liquidity investor. A investor's order placement strategy has two equilibrium conditions: i) an indifference condition (IC) between orders to Exchange Fast and Slow, and ii) a participation constraint (PC). For a speculator, the participation constraint PC_I is the maximum information acquisition costs γ_i that lead a speculator to become an informed investor. Then, conditional on participation, the indifference condition IC_I represents the value of β such that an informed investor is indifferent to submitting an order to or B. These conditions are written as:

$$\text{IC}_I\text{: } \delta\sigma = \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] - (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] \tag{19}$$

$$\text{PC}_I\text{: } \mu_I = \mu \Pr(\gamma_i \leq \max\{\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}], (1-\delta)(\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}])\}) \tag{20}$$

Liquidity investors face two similar conditions. Their participation constraint PC_L describes the scaling of their delay costs $\underline{\lambda}$ at which they are indifferent to trading in $t = 1$ and waiting until $t = 2$. Then, conditional on participating, their indifference condition IC_L describes the value of $\bar{\lambda}$ such that a liquidity investor is indifferent to submitting an order to either

exchange. We write these conditions below.

$$\text{IC}_L\text{: } \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] = (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] + \delta \cdot \frac{k\bar{\lambda}}{2} \times \sigma \tag{21}$$

$$\text{PC}_L\text{: } \underline{\lambda} = \min\left\{\frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}]}{k\sigma}, \frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}]}{k\sigma}\right\} \tag{22}$$

Finally, an equilibrium is characterized by values k such that it is sufficiently costly to delay until $t = 2$ for at least some investors (i.e, $k > \underline{k} > 0$).

Lemma 1 (Costly Delay) *In any equilibrium that satisfies conditions (19)-(22), $k > 2$.*

We can now describe our equilibrium. An equilibrium in a model with a processing delay is characterized by: (i) Ask prices (17) and (18) (and similar bid prices) set by the market maker at exchanges A and B, respectively, such that they earn zero expected profit in expectation; (ii) a solution to the speculator's optimization problem, (19)-(20) and; (iii) a solution to the liquidity investor's optimization problem, (21)-(20). By solving this system, we arrive at the following theorem.

Theorem 2 (Existence and Uniqueness) *Let $k > 2$. For $\delta \in (0, 1)$, there exist unique values μ_I, $\underline{\lambda}, \bar{\lambda}$, β, and prices $\mathsf{ask}_1^{\mathsf{Fast}}, \mathsf{ask}_1^{\mathsf{Slow}}$ given by (17)-(18) that solve equations (19)-(22).*

The nature of the equilibrium depends on the parametrization of the latency delay and can take several forms. For a delay of sufficiently small size, market makers at the delayed exchange are not offered sufficient protection from informed trades. For these delays, both types of traders continue to trade at the delayed-exchange. However, there exists an inflection point, further discussed below, where the delay becomes sufficiently large that informed traders withdraw their flow from the delayed-exchange. For these larger delays, market makers are able to offer vastly improved prices on the latency-delayed exchange, drawing order flow only uninformed traders.

One interpretation of the latency delay is in the context of statistical arbitrage. Instead of interpreting the announcement event as an earnings announcement, it can instead be viewed

as the time with which market makers become aware of arbitrage opportunities. This is similar in many respects to Budish, Cramton, and Shim (2015), who document the fleeting nature of arbitrage opportunities between New York and Chicago. When viewed in this sense, a "short" speed bump is one which is similar in length to the lifespan of actionable arbitrage opportunities. Similarly, a "long" speed bump is one which delays orders sufficiently, such that statistical arbitrage is generally not possible.

3 Empirical Predictions and Policy Implications

We investigate the impact of a latency delay on measures of market quality and price discovery. When Exchange Slow imposes a latency delay, investors who submit an order to Exchange Slow at $t = 1$ face the possibility that private news may become public (i.e., the market maker will update their limit orders) before their order is filled. The latency delay impacts speculators and liquidity investors differently. Speculators do not benefit from a latency delay directly, as a latency delay increases the probability that they may lose their private information advantage, if they trade at Exchange Slow. Hence, ceteris paribus, they prefer an exchange that will execute their order immediately. A liquidity investor's preference, however, depends on their individual costs to delay. Those that have sufficiently low delay costs are impacted more by the price of the order than the possibility of delay, and hence, they may prefer an exchange with a latency delay, if the price offered is at a sufficient discount. Because speculators and liquidity investors' motives are not perfectly correlated, the introduction of a latency delay segments the order flow of the two investor groups, to varying degrees.

The degree of order flow segmentation depends on the parameters of the speed bump. A speed bump is not driven by the magnitude of the delay alone, but the likelihood that a delay of a given length would lead an investors' order to fill after private information becomes public, and hence face updated limit orders. In our model, the latency delay δ takes on this

interpretation. We identify a latency delay δ^*—which we refer to as the "segmentation point"—as the delay such that for all $\delta \geq \delta^*$, no informed traders submit orders to the delayed exchange ($\beta = 1$). Moreover, if no informed traders submit orders to Exchange Slow, then it must be that in equilibrium, $\mathsf{ask}_1^{\mathsf{Slow}} = 0$. Thus, because the cost of trading on exchange Slow is bounded above by the cost of delay, it must be that all uninformed investors participate in the market at $t = 1$ ($\underline{\lambda} = 0$). Given these solutions, we solve equations (19)-(22) for δ^*, yielding the equation:

$$\delta^*(k, \mu, \sigma) = \frac{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + (1-\mu)(1-\frac{2}{k})\mu\sigma} - (1-\mu)(1-\frac{2}{k})}{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + (1-\mu)(1-\frac{2}{k})\mu\sigma} + (1-\mu)(1-\frac{2}{k})} \tag{23}$$

We use δ^* to characterize our results on order flow segmentation in Proposition 1 below.

Proposition 1 (Order Flow Segmentation) *Relative to the benchmark value at $\delta = 0$, if Exchange* Slow *imposes a delay $\delta \in (0, 1)$, then:*

- *for $\delta \leq \delta^*$, informed trading on Exchange* Slow *falls ($\beta \downarrow$), and the measure of liquidity investors who submit orders only at $t = 2$ declines ($\underline{\lambda} \downarrow$).*
- *for $\delta > \delta^*$, informed trading concentrates on Exchange* Fast *($\beta = 1$), and all liquidity investors submit orders at $t = 1$ ($\underline{\lambda} = 0$). Moreover, liquidity trading on Exchange* Fast *increases ($\bar{\lambda} \downarrow$).*

While we find that $\beta = 1$ for all $\delta > \delta^*$, we do not predict full order flow segmentation of informed and uninformed investors, as uninformed investors whose delay costs are large enough ($\lambda_i \geq \bar{\lambda}$) still use Exchange Fast. The relationship between the value of δ and the participation of both investor types is shown in Figure 2.

Order flow segmentation represents one of the reasons why latency delays are often advertised by exchanges. Proponents argue that delays are a means of protecting liquidity suppliers from informed investors. Empirically speaking, existing work supports this fact and finds that that exchanges with latency delays have lower informed trading and higher

participation by uninformed orders (Chen, Foley, Goldstein, and Ruf 2016). We show that, for a sufficiently long delay, informed traders do optimally avoid these exchanges altogether, allowing liquidity suppliers to quote a near-zero spread for uninformed investors.

The existence of the latency delay implies that with some probability an order submitted to Exchange Slow will be delayed until after a public information announcement about the security being traded. Thus, the market maker is afforded the opportunity to update their limit orders before the delayed order arrives, allowing them to potentially avoid being adversely selected. Because the potential of updated quotes is equally costly to all informed investors, but not all liquidity investors, it is natural to hypothesize that quoted spreads would differ across exchanges Fast and Slow. Our model yields the following prediction on quoted spread behaviour between Exchanges Fast and Slow, given a latency delay, $\delta \in (0, 1)$.

Proposition 2 (Quoted Spreads) *For $\delta \in (0, 1)$ quoted spreads are narrower for Exchange* Slow ($\mathsf{ask}^{\mathsf{Slow}} \leq \mathsf{ask}^{\mathsf{BM}}$) *and wider at Exchange* Fast ($\mathsf{ask}^{\mathsf{Fast}} \geq \mathsf{ask}^{\mathsf{BM}}$)*, when compared to the benchmark case. For $\delta < \delta^*$, the spread widens at Exchange* Fast *as δ increases, while for $\delta > \delta^*$, the spread narrows at Exchange* Fast *as δ increases.*

While the market maker may have the opportunity to update their quotes before an informed trade clears the latency delay, they face additional costs at the non-delayed exchange. Informed traders concentrate at the non-delayed exchange, increasing adverse selection costs and forcing the market maker to quote worse prices than in the benchmark case. We illustrate the impact of δ on quoted spreads in Figure 4. Proposition 2 reflects the empirical results in Chen, Foley, Goldstein, and Ruf (2016), who find that spreads improve on the exchange with the latency delay, and become worse elsewhere.

An improvement in quotes at Exchange Slow is correlated with our result on order segmentation (Proposition 1): the migration of informed traders to Slow leads to an increase in market participation at $t = 1$ by liquidity investors. To analyze this order segmentation, we define total order submissions (OS) as the probability that an investor who enters, submits

an order at $t = 1$:

$$\mathsf{OS}_{t=1} = \mu\bar{\gamma} + (1-\mu) \times (1-\underline{\lambda}) \tag{24}$$

We then determine from Equation 25 how much of total order submissions at $t = 1$ are expected to result in trades before $t = 2$, our measure for trading volume before $t = 2$.

$$\mathsf{OS}_{t=1} = \mu\bar{\gamma} \times (\beta + (1-\beta)(1-\delta)) + (1-\mu) \times ((1-\bar{\lambda}) + (1-\delta)(\bar{\lambda} - \underline{\lambda})) \tag{25}$$

The right panel of Figure 3 shows that, as liquidity investors increase their participation, the migration of informed traders to Exchange Fast and the resulting increase in quoted spreads at Exchange Fast lead to a decline in informed trader participation, net of which our model predicts an increase in aggregate order submissions. This increase does not lead to an increase in total trading volume, however, as the increase in liquidity investor participation occurs primarily at Exchange Slow, orders at which, fill before $t = 2$ with probability $1 - \delta$. We summarize this result below.

Proposition 3 (Total Volume and Participation) *Relative to the benchmark value at $\delta = 0$, if Exchange* Slow *imposes a delay $\delta \in (0, 1)$, then liquidity investor participation improves ($\underline{\lambda} \downarrow$), and information acquisition falls ($\bar{\gamma} \downarrow$). Moreover, total market order submission at $t = 1$ increases, but expected trading volume prior to $t = 2$ declines.*

The latency delay affects incentives for both liquidity investors and informed investors. For liquidity investors, the improved prices offered by the market maker increases participation. As more liquidity investors enter the market and select the latency-delayed exchange, the market maker probability of adverse selection falls, further improving prices. For informed investors, the latency delay creates a disincentive for information acquisition. As δ increases towards δ^*, the proportion of liquidity investors to informed traders on the non-delayed exchange decreases, increasing spreads and decreasing total participation by informed investors. Moreover, a rise in δ improves the likelihood that an informed trader loses their information advantage if they trade on exchange Slow.

If the delay is sufficiently long, however, ($\delta = \delta^*$), all informed traders segregate to the non-delayed exchange, and all liquidity investors participate before $t = 1$. At this point, that any longer delay cannot improve the adverse selection costs on Exchange Slow, as these costs are already zero. Then, it must be that an increase in the delay probability beyond δ^* can only increase the probability that a liquidity investor pays their delay cost, which must be greater than $\mathsf{ask}_1^{\mathsf{Slow}} = 0$. Thus, for any $\delta > \delta^*$, liquidity investors must migrate from Exchange Slow to Exchange Fast (see Figure 2). For a sufficiently long delay, both informed traders and liquidity traders at the non-delayed exchange revert to the case where no delayed-exchange exists.

In comparison to the benchmark case, we find that the presence of a delayed exchange unequivocally reduces information acquisition by informed investors (and their subsequent market participation). We examine whether this fall in information acquisition arising from the presence of a delayed exchange contributes positively or negatively the price discovery process. In our framework, we define a measure of price discovery as the fraction of trades prior to the announcement of v that can be attributed to informed trades (that is, the permanent price impact of a trade).

$$\mathsf{Price\ Discovery} = \mu\bar{\gamma} \times (\beta \cdot \mathsf{ask}_1^{\mathsf{Fast}} + (1-\beta)(1-\delta) \cdot \mathsf{ask}_1^{\mathsf{Slow}}) \tag{26}$$

An informed investor's contribution to permanent price impact has three components: i) the probability of information acquisition, ii) the likelihood of a trade by an informed investor, and iii) the quote they hit (i.e. their price impact). From Proposition 3, we know that μ_I is lower for any $\delta \in (0, 1)$ when compared to the benchmark case, so the presence of a delayed exchange reduces permanent price impact under component (i). The impact of (ii) and (iii) are more nuanced, however. For $\delta < \delta^*$, the probability of trading at $t = 1$ for informed investors falls for those participating on Exchange Slow, and the quoted spread narrows. The countervailing force to this is that informed investors migrate their participation toward Exchange Fast, where trading before $t = 1$ is guaranteed *and* the quoted

spread is widening. For small δ, the reduction in informed investor volume and tightening of the quoted spread dominates, but for sufficiently large delays $\delta > \hat{\delta} >> \delta^*$ where informed trading is concentrated entirely on Exchange Fast, the latter dominates, and price discovery improves above that of the benchmark case.

Numerical Observation 1 (Price Discovery) *Relative to $\delta = 0$, there exists a unique $\hat{\delta} > \delta^*$, such that for all $\delta < \hat{\delta}$, average price movement attributed to informed trades (permanent price impact) at $t = 1$ worsens. For any $\delta > \hat{\delta}$, price discovery improves.*

An additional consequence of the latency delay is a change in pre-announcement price discovery, as shown in Figure 4. While price-discovery decreases for shorter delays, sufficiently long delays concentrate traders at the exchange with no delay and may improve price discovery measures. Unlike the previous results in this paper, which represent a transfer between liquidity traders and informed investors, the change price discovery information represents a cost imposed on the market by the delayed exchange. This prediction is somewhat at odds with the empirical results of Chen, Foley, Goldstein, and Ruf (2016), as we predict that price discovery may improve following the introduction of some forms of latency delay.

Curiously, we find that with sufficiently short delays, price discovery falls, but spread widen for informed traders. Here, markets lose benefits from price discovery, while informed traders continue to pay higher trading costs. Combined, these two changes represent a cost imposed on other exchanges from the introduction of a latency delay. This form of equilibrium is counter to conventional results, where increased price discovery results in wider spreads, and decreased price discovery allows market makers to quote narrower spreads.

Because of the ambiguous effects on price discovery, the effects on liquidity investors are also not definitive. We examine whether this effect has a positive transfer to liquidity investors via a reduction in trading costs paid on *average* (across liquidity investors of all delay cost types). We write this measure in the following way:

$$\mathsf{ATC} = \int_{\bar{\lambda}}^{1} \mathsf{ask}_1^{\mathsf{Fast}} \mathrm{d}\lambda + \int_{\underline{\lambda}}^{\bar{\lambda}} \mathsf{ask}_1^{\mathsf{Slow}} + \frac{k\sigma}{2}\lambda \mathrm{d}\lambda + \int_{0}^{\underline{\lambda}} \frac{k\sigma}{2}\lambda \mathrm{d}\lambda \tag{27}$$

We now examine how ATC is impacted by the introduction of an exchange with a latency delay, δ. Our result is presented graphically in Figure 4.

Numerical Observation 2 (Liquidity Investor Trading Costs) *There exist unique* $\underline{\delta}$ *and* $\overline{\delta}$ *such that* $0 < \underline{\delta} < \overline{\delta} < 1$ *where liquidity investors:*

- *pay lower average costs if* $\delta < \underline{\delta}$ *or* $\delta > \overline{\delta}$ *relative to* $\delta = 0$.
- *pay higher average costs if* $\delta \in [\underline{\delta}, \overline{\delta}]$ *relative to* $\delta = 0$.

Despite the fact that more liquidity investors participate in the market pre-announcement, the average delay costs borne by those traders increases. This, seemingly contradictory behaviour is a result of new liquidity traders submitting orders in $t = 1$, rather than delaying until $t = 2$. Without the latency delay, liquidity traders with the lowest delay cost are those who choose not to enter the market, and delay trading until the final period. With the latency delay, these low delay cost traders enter the market and trade on the delayed-exchange. For liquidity traders already in the market, an increase in the delay time increases the optimality of trading on the delayed exchange. While these traders are offered better prices, they incur higher delay costs, which increase their total cost of trading. Broadly speaking, traders who begin to enter the market at $t = 1$ as a result of the latency delay are made better off, while many of those who were already in the market are made worse off.

4 Conclusion

Latency delays have been a topic of controversy since their introduction. Proponents contend that they improve liquidity for uninformed investors via narrower spreads, while opponents claim that the liquidity improvement is illusory: the "improved" quotes may fade before they are ever hit. We construct a model of latency delays in order to disentangle potential effects from their introduction.

We find that many of the effects from latency delays depend on the length of the delay. Specifically, we define a "segmentation point", which is the shortest length of a latency

delay such that all informed traders cluster on the non-delayed exchange. As the length of a latency delay increases towards this point, the crowding of informed traders at the non-delayed exchange widens its bid-ask spread. Concurrently, more liquidity traders migrate to the delayed exchange, narrowing the its quoted spread, and increasing its total order flow.

Once the delay increases past the segmentation point, results change drastically. The spread at the latency-delayed exchange holds constant, and liquidity traders begin migrating to the non-delayed exchanges. This migration improves bid-ask spreads at non-delayed exchanges, and encourages more informed traders to (re-)enter the market. Finally, for sufficiently long latency delays, non-delayed markets are identical to the case with no delays, while the delayed markets contain only liquidity traders who did not trade in market with no delayed exchange.

Our model makes several empirical predictions. We predict that, following the introduction of a delay, quoted spreads should improve at the delayed exchange, while worsening at the standard exchanges. We also predict that the presence of a delayed exchange improves liquidity investor participation, and that informed trading should cluster on the non-delayed exchange. Our model also offers several predictions for policy makers. First, we find that the introduction of a delayed exchange can impact other exchanges. Other exchanges are likely to see an increased concentration of informed order flow and a withdrawal of retail order flow. Market makers on these exchanges may require additional protection, or they may withdraw from markets or quote at much worse prices. Alternatively, the delayed exchanges are particularly attractive to uninformed traders. This may create the need for special attention by regulators who may be concerned about protecting retail investors and non-professional market participants. Finally, sufficiently-short latency delays may create a loss in price discovery, combined with an increase in spreads at non-delayed exchanges. This combination represents a cost imposed on other markets from a delayed-exchange. Our model shows that, as with many market structure phenomena, policy makers must take a nuanced view to changes involving latency delays.

References

Angel, James J, Lawrence E Harris, and Chester S Spatt, 2011, Equity trading in the 21st century, *Quarterly Journal of Finance* 1, 1–53.

Baldauf, Markus, and Joshua Mollner, 2016, Trading in fragmented markets, *Available at SSRN*.

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

Brogaard, Jonathan, and Corey Garriott, 2015, High-frequency trading competition, .

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

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2014, High-frequency trading and price discovery, *Review of Financial Studies* 27, 2267–2306.

——— , 2015, Price discovery without trading: Evidence from limit orders, *Available at SSRN 2655927*.

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

Carrion, Allen, 2013, Very fast money: High-frequency trading on the nasdaq, *Journal of Financial Markets* 16, 680–711.

Chakrabarty, Bidisha, Pankaj K Jain, Andriy Shkilko, and Konstantin Sokolov, 2014, Speed of market access and market quality: Evidence from the sec naked access ban, *Available at SSRN 2328231*.

Chen, Haoming, Sean Foley, Michael A Goldstein, and Thomas Ruf, 2016, The value of a millisecond: Harnessing information in fast, fragmented markets, .

Cimon, David A, 2016, Broker routing decisions in limit order markets, *Available at SSRN*.

Colliard, Jean-Edouard, and Thierry Foucault, 2012, Trading fees and efficiency in limit order markets, *Review of Financial Studies* 25, 3389–3421.

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

Foucault, Thierry, and Albert J Menkveld, 2008, Competition for order flow and smart order routing systems, *Journal of Finance* 63, 119–158.

Gomber, Peter, Satchit Sagade, Erik Theissen, Moritz Christian Weber, and Christian Westheide, 2016, Spoilt for choice: Order routing decisions in fragmented equity markets, .

Jovanovic, Boyan, and Albert J Menkveld, 2011, Middlemen in limit order markets, *Western finance association (WFA)*.

Kwan, Amy, Ronald Masulis, and Thomas H McInish, 2015, Trading rules, competition for order flow and market fragmentation, *Journal of Financial Economics* 115, 330–348.

Latza, Torben, Ian W Marsh, and Richard Payne, 2014, Fast aggressive trading, *Available at SSRN 2542184*.

Malinova, Katya, and Andreas Park, 2016, modern market makers, *Retrieved from* ***http://firn.org.au/wp-content/uploads/2016/05/Modern-Market-Makers-Park-Malinova.pdf*** *, October 26, 2016*.

Menkveld, Albert J, and Marius A Zoican, 2016, Need for speed? exchange latency and liquidity, *Review of Financial Studies (Forthcoming)*.

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

Subrahmanyam, Avanidhar, and Hui Zheng, 2015, Limit order placement by high-frequency traders, *Available at SSRN 2688418*.

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

Ye, Mao, and Maureen O'Hara, 2011, Is market fragmentation harming market quality?, *Journal of Financial Economics* 100, 459–474.

Ye, Mao, Chen Yao, and Jiading Gai, 2013, The externalities of high frequency trading, *Available at SSRN 2066839*.

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

A Appendix

In the appendix, we include a description of the mechanics underlying latency delays, all proofs and figures not presented in-text.

A.1 Latency Delays.

Broadly speaking, latency delays are means by which an exchange imposes a delay on some or all of their incoming orders. Despite being a relatively new feature offered by exchanges, many varieties of latency delay exist.

The most well known type of latency delay is that of IEX in the United States. This delay, sometimes referred to as the magic shoebox, indiscriminately slows down all orders entering the exchange by 350 microseconds. This alone would not prevent multi-market strategies, as traders could simply send their orders to the delayed exchange in advance. However, markets such as IEX generally allow traders to post pegged orders, which move instantaneously in response to external factors Since these pegged orders move instantaneously if trading occurs on other exchanges, market makers using these orders are offered some protection from multi-market trading strategies.

The pegged orders at IEX are available in multiple forms, but the one most relevant to this paper is what is called the "discretionary peg". This order type uses a known algorithm to determine if a price movement is likely, a behaviour IEX refers to as a "crumbling quote".[10] If IEX determines that the quote in a particular security is likely to move, it automatically reprices orders placed at "discretionary pegs", without the 350 microsecond delay.

A second type of delay allows some forms of liquidity supplying orders to simply bypass the delay. These limit orders often have a minimum size, or price improvement requirement, which differentiates them from a conventional limit order. By allowing some orders to bypass the latency delays, market makers who use these orders are able to update their quotes in

[10]Complete documentation is available in the IEX Rule Book, Section 11.190 (g), available here: `https://www.iextrading.com/docs/Investors\%20Exchange\%20Rule\%20Book.pdf`

response to trading on other venues. If the delay is calibrated correctly, this updating can occur before the same liquidity demanding orders bypass the latency delay. Critics contend that these delays also potentially allows market makers to fade their quotes, removing liquidity before any large order reaches the exchange.

This form of latency delay is used on the Canadian exchange TSX Alpha. In the case of TSX Alpha, orders entering the exchange are delayed by a period of 1 to 3 milliseconds before reaching the order book. A special order type, a limit order referred to as a "post only" order, is able to bypass this delay. Unlike a conventional limit order, the "post only" order also contains a minimum size requirement based on the price of the security. These sizes range from 100 shares for high priced to 20,000 shares for lower priced securities.[11]

Finally, a third type of latency delay explicitly classifies traders into two groups. Some traders are affected by the delay, and have their orders held up for a fixed period of time. Other traders are simply not affected and trade as normal. Unlike the other two types of delays which rely on order types, this form requires the explicit division of traders by the exchange into two types. This is used on the Canadian exchange Aequitas NEO, which divides traders into Latency Sensitive Traders, who are affected by the speed bump, and non-Latency Sensitive Traders, who are not.[12] Those are are deemed to be "latency sensitive" are subjected to a randomized delay of between 3 to 9 milliseconds.

[11]Complete documentation is available on the TMX Group website, here: `https://www.tsx.com/trading/tsx-alpha-exchange/order-types-and-features/order-types`

[12]The factors underlying this determination are outlined in Section 1.01 of the Aequitas Neo rule book, available here: `https://aequitasneoexchange.com/media/176022/aequitas-neo-trading-policies-march-13-2017.pdf`

A.2 Proofs

Proof Sketch (Theorem 1).

Investors who choose to buy at $t = 1$ at Exchange j have profit functions given by:

$$\pi_I(\gamma_i; \text{Buy at t=1}) = v - \mathsf{ask}_1^j - \gamma_i \tag{28}$$

$$\pi_L(c_i; \text{Buy at t=1}) = v_0 - \mathsf{ask}_1^j \tag{29}$$

Because exchanges are identical in their operation, it must be that in any equilibrium, their ask and bid prices are identical. These prices are given by the following:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = \frac{\beta\mu_I}{\beta\mu_I + (1-\mu)\alpha\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma \tag{30}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = \frac{(1-\beta)\mu_I}{(1-\beta)\mu_I + (1-\mu)(1-\alpha)\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma \tag{31}$$

We then solve $\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}}$ for $(\alpha, \beta) \in (0,1)^2$, for all μ_I and $\underline{c}$:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = \mathsf{ask}_1^{\mathsf{Slow}} \tag{32}$$

$$\iff \frac{\beta\mu_I}{\beta\mu_I + (1-\mu)\alpha\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma = \frac{(1-\beta)\mu_I}{(1-\beta)\mu_I + (1-\mu)(1-\alpha)\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma \tag{33}$$

$$\iff \beta(1-\alpha) = (1-\beta)\alpha \Rightarrow \beta = \alpha \tag{34}$$

Given equilibrium prices in (30) and (31), we then solve for μ_I and $\underline{c}$. To solve for μ_I, we solve the equation:

$$\mu_I = \mu \times \mathsf{Pr}(\gamma_i \leq \min\left\{v - \mathsf{ask}_1^{\mathsf{Fast}}, v - \mathsf{ask}_1^{\mathsf{Slow}}\right\}) \tag{35}$$

$$\Rightarrow \bar{\gamma} - (v - \mathsf{ask}_1^{\mathsf{Fast}}) = 0 \tag{36}$$

where the simplification in (36) arises from the fact that the ask prices at Exchanges Fast and Slow are identical in equilibrium. We then show that there exists a unique $\bar{\gamma} \in [0,1]$

that solves (36). Given this $\bar{\gamma}$, $\mu_i = \mu \times \bar{\gamma}$ exists, and is unique.

$$\bar{\gamma} = 0 : 0 - (v - 0) < 0 \tag{37}$$

$$\bar{\gamma} = 1 : 1 - \sigma \left(1 - \frac{\mu}{\mu + (1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \right) > 0 \tag{38}$$

where (38) is positive because $\sigma < 1$. Then differentiate equation (36) by $\bar{\gamma}$:

$$\frac{\partial}{\partial \bar{\gamma}} (\bar{\gamma} - (v - \mathsf{ask}_1^{\mathsf{Fast}})) = 1 + \sigma \left(\frac{(1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right)}{(\mu + (1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right))^2} \right) > 0 \tag{39}$$

for all $\underline{c}$. Then, to show there exists a unique $\underline{c}$, consider the participation constraint for liquidity investors, $\underline{c} - \mathsf{ask}_1^{\mathsf{Fast}} \geq 0$:

$$\underline{c} = 0 : 0 - \frac{\mu_I}{\mu_I + (1-\mu)\mathsf{Pr}\left(c_i \geq 0\right)} \cdot \sigma < 0 \tag{40}$$

$$\underline{c} = 1 : 1 - \sigma > 0 \tag{41}$$

where (41) is positive because $\sigma < 1$. Then differentiate $\underline{c} - \mathsf{ask}_1^{\mathsf{Fast}} \geq 0$ by $\underline{c}$:

$$\frac{\partial}{\partial \underline{c}} (\underline{c} - \mathsf{ask}_1^{\mathsf{Fast}}) = 1 + \sigma \left(\frac{(1-\mu)\mu_i}{(\mu + (1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right))^2} \right) > 0 \tag{42}$$

Thus, a unique equilibrium exists for all $\beta = \alpha \in (0,1)^2$. ■

Proof (Lemma 1). For an equilibrium to exist, we require that liquidity investors will trade before $t = 1$ for a non-zero measure of λ_i on *both exchanges*. To ensure this, a sufficient condition is that the scaling of the cost of delaying trade, k, must be large enough, to entice investors with the largest valuations ($\lambda_i \geq 1 - \epsilon$, for ϵ arbitrarily close to zero) to trade at an exchange that posts the widest possible spread, equal to 2σ. Then, k must satisfy:

$$\frac{k(1-\epsilon)\sigma}{2} > \sigma \iff k > \frac{2}{1-\epsilon} > 2$$

Hence, in any equilibrium where investors use both exchanges, $k > 2$. ■

Proof (Theorem 2). The proof of Theorem 2 proceeds similarly to Theorem 1, except

that we solve the liquidity investor constraints for $\bar{\lambda}$ and $\underline{\lambda}$, instead of $\underline{c}$ and α.

There are three equilibrium cases, defined through the (mixed) strategies of speculators: $\beta = 0$, $\beta(0,1)$, and $\beta = 1$.

Speculators use only Exchange Slow ($\beta = 0$): In this part, we show that no equilibrium exists for $\beta = 0$. To do so, we consider the informed investor's incentive compatibility constraint, evaluated at $\beta = 0$.

$$\text{IC}_I\text{: } \sigma - 0 - (1-\delta)(\sigma - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})}) = \delta\sigma + \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})} > 0 \tag{43}$$

Moreover, because $\mathsf{ask}^{\mathsf{Fast}} = 0$, then $\bar{\gamma} < 1$, implying that informed investors would always have an incentive to deviate to the fast exchange.

Speculators use both exchanges ($\beta \in (0,1)$): We now solve the following system of equations for $\bar{\lambda}, \underline{\lambda}, \bar{\gamma}$ and β, using the method as in the proof of Theorem 1.

$$\text{IC}_I\text{: } \delta\sigma = \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] - (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] \tag{44}$$

$$\text{PC}_I\text{: } \mu_I = \mu \Pr(\gamma_i \leq \max\{\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}], (1-\delta)(\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}])\}) \tag{45}$$

$$\text{IC}_L\text{: } \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] = (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] + \delta \cdot \frac{k\bar{\lambda}}{2} \times \sigma \tag{46}$$

$$\text{PC}_L\text{: } \underline{\lambda} = \min\left\{\frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}]}{k\sigma}, \frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}]}{k\sigma}\right\} \tag{47}$$

We write (44)-(47) explicitly as:

$$\text{IC}_I\text{: } 1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\left(1 - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})}\right) = 0 \tag{48}$$

$$\text{PC}_I\text{: } \bar{\gamma} - \sigma\left(1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})}\right) = 0 \tag{49}$$

$$\text{IC}_L\text{: } \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} - \frac{\delta k\bar{\lambda}}{2} = 0 \tag{50}$$

$$\text{PC}_L\text{: } \frac{\delta k\underline{\lambda}}{2} - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} = 0 \tag{51}$$

We first show that, for all $(\bar{\lambda}, \underline{\lambda}, \bar{\gamma}) \in (0,1)^3$, there is a unique $\beta^* \in (0,1)$ that solves (44).

$$\text{IC}_I \mid_{\beta=0} : \delta\sigma + (1-\delta)\frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})} > 0 \tag{52}$$

$$\text{IC}_I \mid_{\beta=1} : \delta\sigma - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} < 0, \ \forall \delta < \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} = \delta^* \tag{53}$$

Thus, for all $\delta < \delta^*$, there exists a $\beta \in (0,1)$ by the intermediate value theorem that satisfies (48). To show that β^* is unique, we differentiate (48) with respect to β.

$$\frac{\partial}{\partial\beta}(\text{IC}_I) = -\frac{\mu\bar{\gamma}(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} - \frac{\mu\bar{\gamma}(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda}))^2} < 0 \tag{54}$$

Thus, β^* is unique for all $(\bar{\lambda}, \underline{\lambda}, \bar{\gamma}) \in (0,1)^3$.

We then rearrange (48) to:

$$\delta = \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} \tag{55}$$

Equation (55) can then be substituted into (50) and simplified to yield an expression for $\bar{\lambda}$:

$$\bar{\lambda} = \frac{2}{k} \tag{56}$$

Next, we use equation (48) and (49) to solve for $\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}]$, in terms of δ, σ and $\bar{\gamma}$, which we substitute into equation (51):

$$\underline{\lambda} = \frac{2}{k}\left(1 - \frac{\bar{\gamma}}{\sigma(1-\delta)}\right) \tag{57}$$

Then, because the right-hand side equals $\frac{2}{k}\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] \in (0,1)$, and $\bar{\lambda} = \frac{2}{k}$, $\underline{\lambda}^*$ exists and is unique for all $\bar{\gamma} \in (0,1)$. Lastly, we show that there exists a unique $\bar{\gamma}^*$ that solves (20), given $\bar{\lambda}^*(\bar{\gamma}), \underline{\lambda}^*(\bar{\gamma})$, and $\beta^*(\bar{\gamma})$.

First, we show that $\bar{\gamma}^* \in [0,1]$ exists, by appealing to the intermediate value theorem:

$$\text{PC}_I \mid_{\bar{\gamma}=0} : 0 - \sigma < 0 \tag{58}$$

$$\text{PC}_I \mid_{\bar{\gamma}=1} : 1 - \sigma(1 - \frac{\mu\beta}{\mu\beta + (1-\mu)(1-\bar{\lambda})}) > 0 \tag{59}$$

where (59) holds by the fact that $\sigma < 1$. Thus, $\bar{\gamma}^* \in (0,1)$ exists. To show that $\bar{\gamma}^*$ is unique, we differentiate (20) by $\bar{\gamma}$:

$$\frac{\partial}{\partial \bar{\gamma}}(\mathrm{PC}_I) = \sigma \frac{\mu(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial \beta}{\partial \bar{\gamma}} \cdot \frac{\mu\bar{\gamma}(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial \bar{\lambda}}{\partial \bar{\gamma}} \cdot \frac{\mu\bar{\gamma}(1-\mu)\beta}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} < 0 \tag{60}$$

Where the third term is zero by the fact that $\frac{\partial \bar{\lambda}}{\partial \bar{\gamma}} = 0$. Now all we need to show is that $\frac{\partial \beta}{\partial \bar{\gamma}} \geq 0$. If we differentiate (19) by $\bar{\gamma}$, and solve for $\frac{\partial \beta}{\partial \bar{\gamma}}$, we find:

$$\frac{\partial IC_I}{\partial \bar{\gamma}} = -\frac{\mu(1-\mu)(1-\bar{\lambda}) + \frac{\partial \beta}{\partial \bar{\gamma}}\mu\beta(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} - (1-\delta)\frac{k}{2} \cdot \frac{\partial \underline{\lambda}}{\partial \bar{\gamma}} = 0 \tag{61}$$

$$\iff \frac{\partial \beta}{\partial \bar{\gamma}} = -\frac{\frac{\mu(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta+(1-\mu)(1-\bar{\lambda}))^2} + (1-\delta)\frac{k}{2} \cdot \frac{\partial \underline{\lambda}}{\partial \bar{\gamma}}}{\frac{\mu\beta(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta+(1-\mu)(1-\bar{\lambda}))^2}} > 0 \tag{62}$$

where (62) is positive by the fact that the partial derivative of $\underline{\lambda}$ with respect to $\bar{\gamma}$ is:

$$\frac{\partial \underline{\lambda}}{\partial \bar{\gamma}} = -\frac{2}{\sigma k(1-\delta)} < 0$$

which implies that:

$$\frac{\mu(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} - \frac{1}{\sigma} < 0$$

Hence, $\bar{\gamma}^*$ is unique.

Speculators use only Exchange Fast ($\beta = 1$): Lastly, we solve equations (44)-(47) for the case where $\beta =$. Inputting $\beta = 1$, we have:

$$\mathrm{IC}_I\text{: } \delta - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} \geq 0 \tag{63}$$

$$\mathrm{PC}_I\text{: } \bar{\gamma} - \sigma\left(1 - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})}\right) = 0 \tag{64}$$

$$\mathrm{IC}_L\text{: } \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} - \frac{\delta k\bar{\lambda}}{2} = 0 \tag{65}$$

$$\mathrm{PC}_L\text{: } \frac{\delta k\underline{\lambda}}{2} = 0 \tag{66}$$

Equation (63) pins down the relation between β and δ: for all $\delta \geq \frac{\mu\bar{\gamma}}{\mu\bar{\gamma}+(1-\mu)(1-\lambda)}$, $\beta^* = 1$.

Moreover, by inspection, we see that $\underline{\lambda}^* = 0$. To prove the existence of a unique $\bar{\gamma}$, we solve equation (64) for $\bar{\gamma}$:

$$\bar{\gamma}^* = \frac{\sqrt{(1-\mu)^2(1-\bar{\lambda})^2 + (1-\mu)(1-\bar{\lambda})\mu\sigma} - (1-\mu)(1-\bar{\lambda})}{2\mu} \tag{67}$$

By inspection, $\bar{\gamma}^*$ exists and is unique as long as the limit $\mu \to 0$ exists, and is in the interval [0,1]. To calculate this limit, we need to apply L'Hôpital's Rule.

$$\lim_{\mu \to 0} \left(\frac{\frac{\partial}{\partial \mu}\left(\sqrt{(1-\mu)^2(1-\bar{\lambda})^2 + (1-\mu)(1-\bar{\lambda})\mu\sigma} - (1-\mu)(1-\bar{\lambda})\right)}{\frac{\partial}{\partial \mu}(2\mu)} \right) = \frac{\bar{\lambda} + \sigma}{4} \in [0,1] \tag{68}$$

Lastly, we show that there exists a unique $\bar{\lambda} \in [0,1]$ that solves (65).

$$\text{IC}_L \mid_{\bar{\lambda}=0} : \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)} - 0 > 0 \tag{69}$$

$$\text{IC}_L \mid_{\bar{\lambda}=1} : 1 - \frac{k}{2} < 0 \tag{70}$$

Thus, $\bar{\lambda}^*$ exists. To show that it is unique, we differentiate (65) with respect to $\bar{\lambda}$:

$$\frac{\partial}{\partial \bar{\lambda}}(IC_L) = \frac{\mu\bar{\gamma}(1-\mu)}{(\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial \bar{\gamma}}{\partial \bar{\lambda}} \cdot \frac{\mu(1-\bar{\lambda})(1-\mu)}{(\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda}))^2} - \frac{\delta k}{2} < 0 \tag{71}$$

Since $\bar{\lambda} \leq 2/k$, the following holds.

Thus, a unique equilibrium exists for $\left\{\beta, \bar{\lambda}, \underline{\lambda}, \bar{\gamma}\right\} = \left\{1, \bar{\lambda}^*, 0, \bar{\gamma}^*\right\}$ ■

Figure 2: Market Participation by Investor Type

The left panel below depicts the unconditional probabilities of a speculator's action prior to $t = 2$ (β), as a function of the latency delay δ. The right panel illustrates the market participation choices of liquidity investors, as a function of the latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.

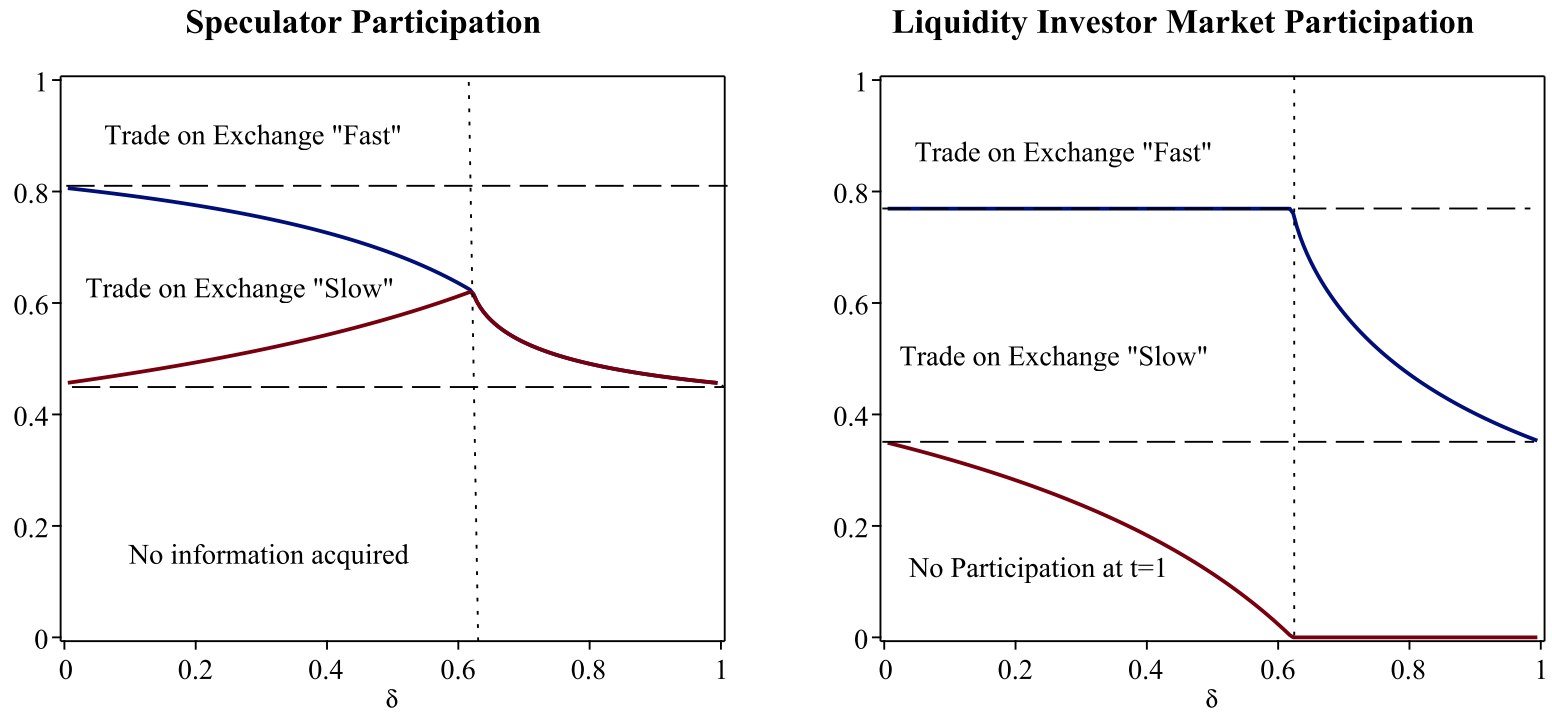


Figure 3: Order Submissions, Trades, and Market Participation

The left panel below depicts total orders submitted and trades executed pre-announcement (prior to $t = 2$), as a function of the Exchange Slow latency delay δ. The right panel illustrates market participation by speculators (μ_I) and liquidity investors (μ_L), as a function of the latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.

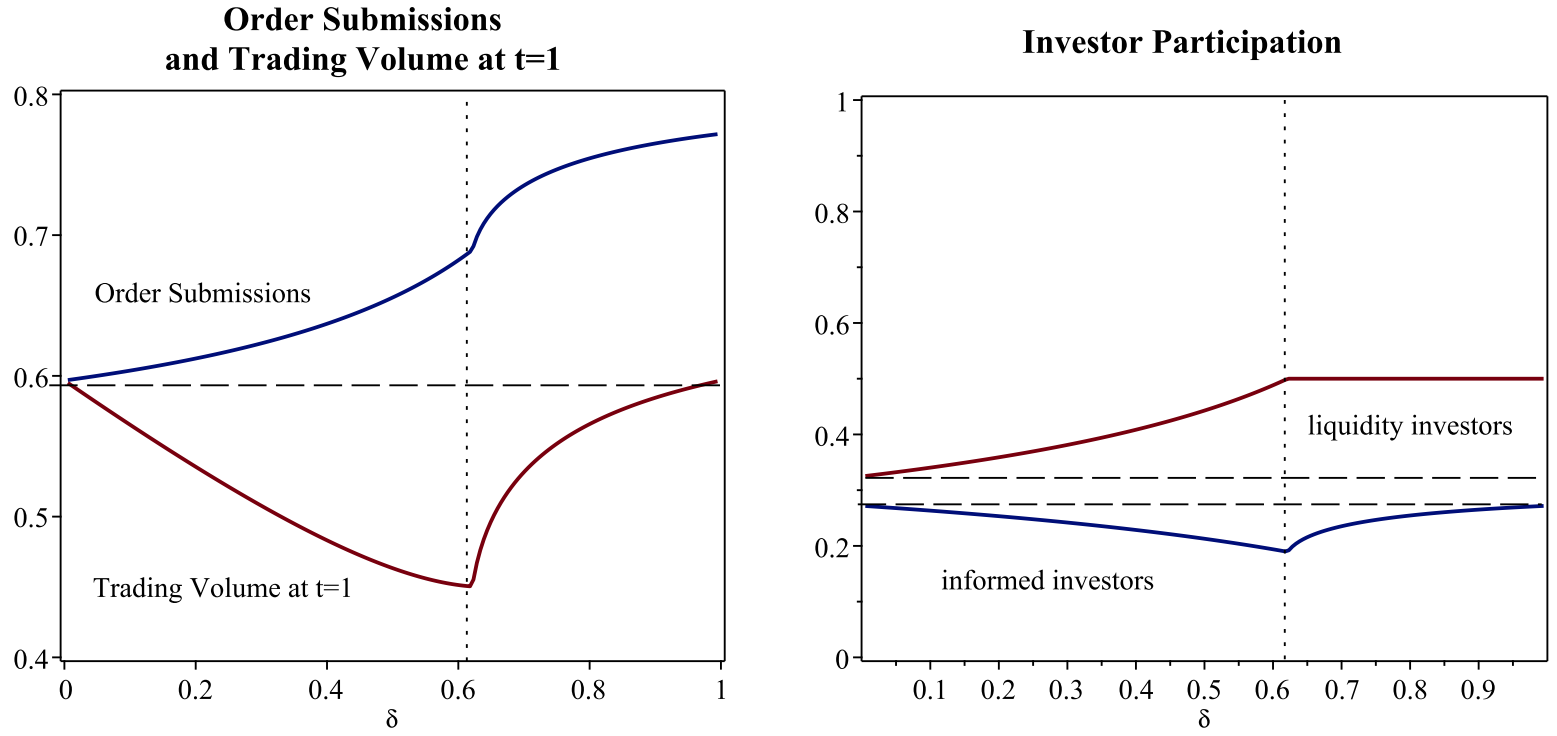


Figure 4: Quoted Spreads and Price Discovery

The left panel below presents the quoted half-spreads for exchanges Fast and Slow at $t = 1$, as a function of the latency delay δ. The right panel depicts price discovery pre-announcement, which we measure as average price movement attributed to informed trades prior to $t = 2$ (the announcement date of v), as a function of the Exchange Slow latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.

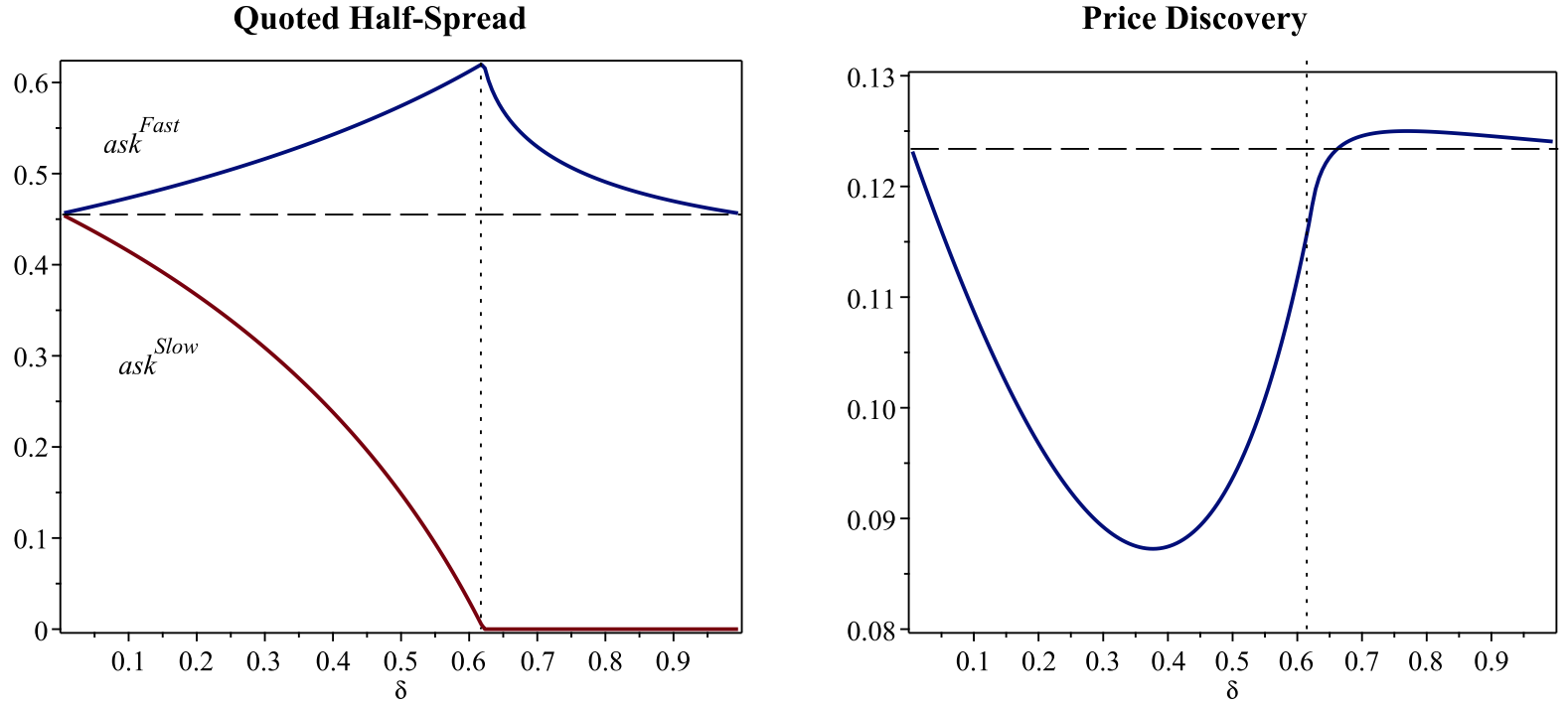


Figure 5: Liquidity Investor Trading Costs

The left panel below illustrates the average trading costs paid by a liquidity investor who enters the market at $t = 0$. In the right panel, we present the trading costs due to delay and the trading costs due to realized quotes separately, as well as the aggregation (from the left panel). We present these costs as a function of the latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.