Brolley, Michael; Cimon, David A.

Working Paper

Order flow segmentation, liquidity and price discovery: The role of latency delays

Bank of Canada Staff Working Paper, No. 2018-16

Provided in Cooperation with:
Bank of Canada, Ottawa

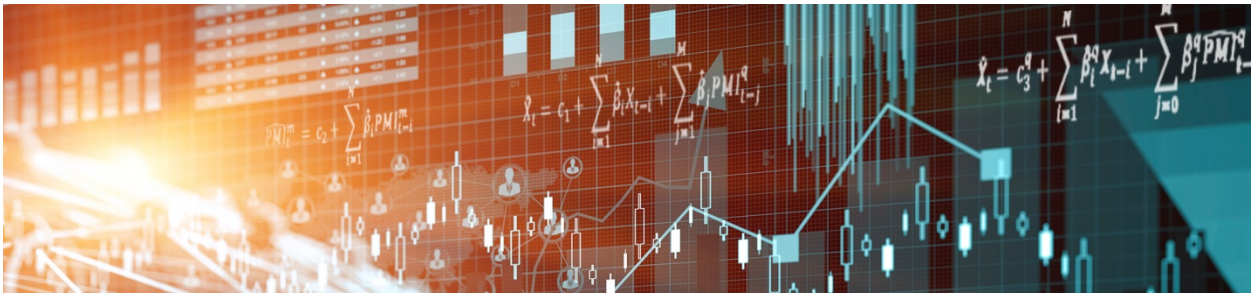
Suggested Citation: Brolley, Michael; Cimon, David A. (2018) : Order flow segmentation, liquidity and price discovery: The role of latency delays, Bank of Canada Staff Working Paper, No. 2018-16, Bank of Canada, Ottawa, https://doi.org/10.34989/swp-2018-16

This Version is available at:
https://hdl.handle.net/10419/197869

BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel 2018-16

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



by Michael Brolley and David A. Cimon

www.bank-banque-canada.ca

Bank of Canada Staff Working Paper 2018-16

April 2018

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

by

Michael Brolley[1] and David A. Cimon[2]

[1] Wilfrid Laurier University
mbrolley@wlu.ca

[2] Financial Markets Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
dcimon@bankofcanada.ca

ISSN 1701-9397

Acknowledgements

The authors would like to thank Sabrina Buti, Eric Budish, Sarah Draus, Corey Garriott, Terrence Hendershott, Peter Hoffmann, Katya Malinova, Albert Menkveld, Carol Osler, Andreas Park, Andriy Shkilko, Adrian Walton, Bart Yueshen, Marius Zoican, and participants at the 2017 Stern Microstructure Conference, the 2017 Erasmus Liquidity Conference, the 2017 SAFE Market Microstructure Conference, the 2017 EFA Annual Meeting, the 2017 NFA Annual Conference, the 2017 CEA Annual Meeting, and seminars at the Bank of Canada, University of Toronto, and Wilfrid Laurier University, for valuable discussions and comments. We also thank William Wootton for research assistance. Michael Brolley acknowledges financial support from the Social Sciences and Humanities Research Council of Canada. The views of the authors are not necessarily those of the Bank of Canada. All errors are our own.

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, while the quoted spread narrows at the delayed exchange. The overall market quality impact depends on the relative concentration of speculators who may become informed. If speculators are few relative to liquidity traders, total welfare falls; with relatively more speculators, total welfare rises.

Bank topics: Financial markets; Market structure and pricing; Financial system regulation and policies
JEL codes: G14, G18

Résumé

Les bourses ralentissent volontairement les flux d’exécution des ordres en y introduisant des délais de traitement ou « ralentisseurs ». Les partisans de ce mécanisme avancent que celui-ci protège les teneurs de marché contre l’arbitrage de place pratiqué par les opérateurs à haute fréquence; leurs détracteurs, eux, font valoir que ces délais favorisent l’effacement des cours (*quote fading*) auquel les teneurs de marché peuvent se livrer. Nous construisons un modèle faisant intervenir sur un marché fragmenté des opérateurs informés avec, d’un côté, une bourse où l’on gère un carnet d’ordres conventionnel et, de l’autre, une bourse qui introduit des délais dans le traitement des ordres de marché. Nous montrons que les investisseurs informés se tournent vers la bourse conventionnelle, ce qui a pour effet de creuser les écarts acheteur-vendeur, alors que ces écarts se rétrécissent sur la bourse où des délais de traitement sont introduits. L’impact de ce phénomène sur la qualité d’ensemble du marché dépend de la concentration relative de spéculateurs qui pourraient accéder à une meilleure information. Si les spéculateurs sont peu nombreux par rapport au nombre d’opérateurs en quête de liquidité, le bien-être global diminue; il augmente lorsque les spéculateurs sont relativement plus nombreux.

Sujets : Marchés financiers; Structure de marché et fixation des prix; Réglementation et politiques relatives au système financier
Codes JEL : G14, G18

Non-Technical Summary

Liquidity suppliers prefer to intermediate trades from uninformed investors. These uninformed trades are valuable to stock exchanges, as they are unlikely to move prices against their dedicated liquidity suppliers. These dedicated liquidity suppliers, often referred to as "market makers," play an important role in ensuring liquidity is available at the exchange. Many stock exchanges compete for order flow, and have specialized to attract trades from these uninformed liquidity demanders. Technological changes such as inverse pricing, dark trading and retail order segmentation facilities have all been studied by academics as ways in which exchanges try to draw these traders from other markets. These designs are advertised as a way to prevent informed traders from participating, allowing the market makers to provide more liquidity. Recently, some exchanges have imposed latency delays—popularized as "speed bumps"—as yet another way of attracting uninformed order flow. Measured on the order of milliseconds or microseconds, latency delays impose a time delay between when a trader sends an order and when it is processed by the exchange. In this paper, we model the impact of introducing such a delay.

Since their introduction, latency delays have been controversial. Exchanges advertise latency delays as means of protecting market makers from high-frequency traders, who act on extremely short-horizon information. They argue that competitive market makers then pass the savings forward by quoting a narrower bid-ask spread. Other market participants have suggested that delays create an uneven playing field by allowing market makers to "fade" their quotes: a behaviour where market makers quote one price, but alter it for a worse price before large orders reach the exchange. In our model, both these behaviours arise: the protection from the delay allows market makers to quote a better bid-ask spread, however, orders may be executed at a worse price than initially posted.

Our model generates several testable implications related to latency delays. First, we predict that following the introduction of a delay, quoted prices should improve at the delayed exchange while they should be worse at other standard exchanges. Second, we predict trader

migration between exchanges. More uninformed investors should trade on exchanges following the introduction of an exchange with delay, and their trading should be concentrated at this exchange. Fewer informed investors should trade following the introduction of a delay, and their trading should be concentrated at other standard exchanges. The total effect is an increase in exchange-traded volume, with fewer traders choosing to use off-exchange internalizers and dark pools.

We generate policy-related predictions for welfare and price discovery. We define these results in terms of the ratio of speculators to liquidity investors. We show that when there are many speculators, introducing a delay improves welfare. Similarly, when there are few speculators, introducing a delay harms welfare. The results for price discovery are nearly the opposite. When there are many speculators, introducing a delay causes price discovery to fall. When there are fewer speculators, price discovery may increase, but only for longer delays. Combining these two results, we show that in some circumstances, both price discovery and welfare may actually increase for some longer delays.

"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-supplying market makers prefer to intermediate uninformed trades. These uninformed trades are valuable, as they are unlikely to move prices against market makers. Many exchanges, competing for scarce order flow, have specialized to attract trades from 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 participating. Recently, some exchanges have imposed latency delays—so-called "speed bumps"—as yet another way of segmenting away uninformed 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] In this paper, we model the market impact of introducing such a delay.

As with any market structure change, latency delays have been controversial. Exchanges advertise latency delays as a means of protecting market makers from adverse selection at the hands of high-frequency traders (HFTs), who act on extremely short-horizon information.[2] Exchanges argue that competitive market makers then pass the savings forward by quoting a narrower spread. Other market participants have suggested that delays create an uneven playing field by allowing market makers to "fade" quotes ahead of orders, executing them at worse prices than those that were posted at the time the order was sent.[3] In our model, these behaviours arise endogenously; the protection of a latency delay allows market makers

[1]Descriptions of the mechanics behind some implementations of latency delays are available in the appendix.

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

[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`

to quote a better spread, however, orders may be executed at a worse price if the arrival of information updates the spread before an order is filled.

Our model generates several testable implications related to 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. Uninformed investor participation should increase following the introduction of a delay, and their trading should be concentrated at this exchange. Information acquisition by speculators should fall following the introduction of a delay, and their trading should be concentrated at standard exchanges. The net effect is an increase in total exchange-traded volume, with fewer traders choosing to use off-exchange internalizers. In comparative static results, we show that as volatility increases, total trading volume falls; however, volume on the delayed exchange increases.

We show that the impact of a delay on price discovery and welfare depends on the relative concentration of speculators, who may acquire information. When most investors are speculators, price discovery falls; when few investors may acquire information, the impact on price discovery is dependent on the delay length. In this second case, price discovery falls for shorter delays, but may increase for longer delays. The impact on welfare is nearly the opposite. If the population of speculators is large, the introduction of a delay increases total welfare, but decreases total welfare if the measure of speculators is small. For more moderate concentrations of speculators, the welfare impact depends on the length of delay. In special cases, we show that both price discovery and welfare may increase. Given the set-up of our model, these predictions are testable using the correct data.

Finally, we analyze the incentives for exchanges to impose a latency delay. We consider two market organizations: i) an independently operated delayed exchange that competes with a standard exchange (e.g., IEX, Aequitas Neo), and ii) a delayed exchange that is a subsidiary of the standard exchange (e.g., TSX Alpha, NYSE MKT). In the first case, the exchange imposes a delay to maximize its own volume, while in the second, the exchange

seeks to maximize the combined volume across the two exchanges. Independent exchanges optimally impose shorter delays, while subsidiary exchanges optimally impose the longest delay possible. In either case, exchanges do not impose a delay that increases both welfare and price discovery, when this is possible, suggesting a role for regulators.

Our model is a three-period model of sequential trading. Trading occurs in fragmented markets, where one exchange imposes a latency delay. We model traders who are aware of an impending information event. This information event is interpreted as a fleeting arbitrage opportunity, which will eventually become public information for all market participants in the second period. Traders are able to learn of this arbitrage opportunity, at some cost, and are aware that it will become public knowledge. This interpretation is similar in many respects to Budish, Cramton, and Shim (2015), who document fleeting arbitrage opportunities between New York and Chicago.

We define the "length" of a latency delay as the probability that a trader who submits an order before the new information becomes public has the order executed after the information is impounded into prices. One interpretation is that the delayed exchange imposes a fixed delay, and that the information becomes public after a random period of time. In this case, there is a random probability that other agents may become informed and react before the order clears the delay. This interpretation is also consistent with agents who have some variation in their reaction time to new information. An alternative interpretation is that the delayed exchange imposes a random delay, drawn within a fixed interval, such that private information may become public within the latency delay. In this case, agents may have a fixed or random reaction time, and the random nature of the delay may come from the either the trader, the exchange, or both. Thus, the model is well suited to analyzing both the case of a fixed-length delay and that of a random-length delay.[4] In either case, the length of the delay is relative to the time in which information remains private. A long delay is one in which prices are likely to change before an order passes through the delay, while a short

[4]An example of a fixed-length delay is the Investors Exchange (IEX) in the United States. An example of a random-length delay is TMX Alpha in Canada.

delay is one in which prices are likely to remain the same.

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. They have the choice between either submitting an order immediately to one of the exchanges or waiting until after the information event and submitting their order to an off-exchange internalizer. A liquidity trader who chooses to submit an order before the information event may send the order to either the standard exchange, which executes instantly, or the delayed exchange, which delays the order with some probability. Liquidity traders who delay their order, either by submitting to the internalizer or by being delayed by the delayed exchange, risk paying a delay cost should the market move against them. This delay cost is similar in nature to the one introduced by Zhu (2014) and represents unmodelled risk aversion.

Alternatively, speculators have the option of paying to become privately informed about the asset. Similar to liquidity traders, speculators can either execute their order immediately at the standard exchange or to submit their order to the delayed exchange, and risk having the information event arrive before their order executes. We assume that if a speculator's order executes, the information acquired by the speculator immediately becomes public knowledge before any other trades can occur. Thus, the speculator does not have the incentive to submit more than one order.

Related Literature. While there is little existing literature on the topic of latency delays, the factors that have led to their creation have been well documented. First, predatory high-frequency trading is generally cited as the rationale for the use of latency delays and, as such, is essential to understanding their purpose. Second, as a means for exchanges to differentiate themselves, latency delays 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, traders who make use of them in a strategic manner are inherently HFTs. Several studies of high-frequency liquidity suppliers

have shown that they can improve liquidity (Brogaard and Garriott 2015; Brogaard et al. 2015; Subrahmanyam and Zheng 2015). Work on high-frequency liquidity demanders finds that they may increase price efficiency (Carrion 2013) but also increase transaction costs (Chakrabarty et al. 2014). Further studies find that HFTs improve price discovery through both liquidity supply (Brogaard, Hendershott, and Riordan 2015; Conrad, Wahal, and Xiang 2015) and demand (Brogaard, Hendershott, and Riordan 2014). Our paper complements these existing works by generating new empirical predictions regarding the impact of HFTs on price efficiency and trading costs when a delay is present.

Proponents argue that latency delays can curb "predator" behaviours by HFTs, such as inter-market arbitrage. However, critics have suggested that latency delays may also lead to quote fading. Existing evidence is mixed, as 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] Evidence of these predatory behaviours, as they relate to latency delays, is also mixed. While Chen et al. (2016) find that liquidity demanders are able to access a lower proportion of posted liquidity following the introduction of a delay, Anderson et al. (2018) find that market-wide liquidity does not deteriorate following the same event. In our paper we do not allow market makers to fade quotes arbitrarily; instead, we model market makers who may rationally fade quotes in response to new information.

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] Closest to our paper, Menkveld and Zoican (2017) model the effects of processing latency within an exchange, versus latency in reaching the exchange, a friction similar to an intentional latency delay. We extend existing theoretical work on HFTs by modelling investor migration between multiple exchanges based on intentional delays.

[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 on topics surrounding HFT is present in both Angel, Harris, and Spatt (2011) and O'Hara (2015).

The topic of market segmentation is well studied within the academic literature. Existing empirical work has found that fragmented markets may have improved liquidity (Foucault and Menkveld 2008) and efficiency (Ye and O'Hara 2011). However, theoretical work by Baldauf and Mollner (2016) shows that the net effects of increased fragmentation are ambiguous for liquidity suppliers. Empirically, Kwan, Masulis, and McInish (2015) and Gomber et al. (2016) study the use of other market mechanisms, such as dark trading, in order to attract order flow. As latency delays are another means of attracting order flow, our work suggests additional avenues for empirical segmentation work.

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

1 The Model

Security. There is a single risky security with a random payoff v. v 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 unknown by the public in $t = 1$, but 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. Both operate as displayed limit order books, where posted limit orders are visible to all market participants. Market orders sent to Exchange Fast (also referred to as the standard or non-delayed exchange) in $t = 1$ fill immediately upon receipt. Market orders sent to Exchange Slow (also referred to as the delayed exchange) are subjected to a random delay. With probability $\delta \in (0, 1)$, an order sent to Exchange Slow in $t = 1$ is delayed until $t = 2$, and filled after the value v is

publicly announced. Otherwise, the order is filled immediately in $t = 1$.

The "length" of the latency delay is not specified in absolute terms but instead, relative to the fleeting nature of information. Consider a simple example of an outside arbitrage opportunity, which is exploitable by some market participants for an uncertain period of time, on the order of 5-10 milliseconds. In this example, a latency delay that lasts far less than 5 milliseconds will do little to stop speculation ($\delta = 0$), while one that lasts over 10 milliseconds will remove the opportunity entirely ($\delta = 1$). Alternatively, a delay between 5-10 milliseconds will remove some of these arbitrage opportunities, but leave others available ($\delta \in (0, 1)$). In that sense, a delay affects the ability of traders to act on information that remains exploitable for a short period of time, relative to the period of the delay.

The delay can be viewed as either deterministic or one with a random length. In the first interpretation, the random nature of the delay represents randomness in trader reaction time rather than in delay length. Figure 2 depicts this interpretation, where traders have a distribution of possible reaction times to new information. A deterministic delay slows this distribution by a fixed amount, increasing the probability that any non-delayed traders move first. This first case matches the original latency delay implemented by IEX. In the second interpretation, some orders are slowed for longer periods, while others are allowed to pass more quickly. This second case matches the delays implemented by Canadian venues TMX Alpha and Aequitas Neo.

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, receives only the public information, v_0, about the security's fundamental value and sets prices in a manner similar to Glosten and Milgrom (1985). The market maker has zero latency, and thus is able to place (and update) limit orders on both exchanges at the beginning of periods $t = 1$ and $t = 2$, before other investors place their orders. At $t = 2$, upon the announcement of v, the market maker updates its $t = 1$ limit orders to the public value, v. This update happens before orders that have been delayed at Exchange Slow are

able to reach the exchange.

The ability of market makers to bypass the delay matches an important feature of latency-delayed venues. Generally, these venues provide exceptions for some orders used for market making purposes. On some venues, such as IEX, the exception is that orders pegged at or near the midpoint update instantaneously in response to external factors. On others, such as TMX Alpha, the exception is that liquidity-supplying orders above a certain size bypass the delay. In general, it is insufficient to merely submit a limit order to bypass the delay.

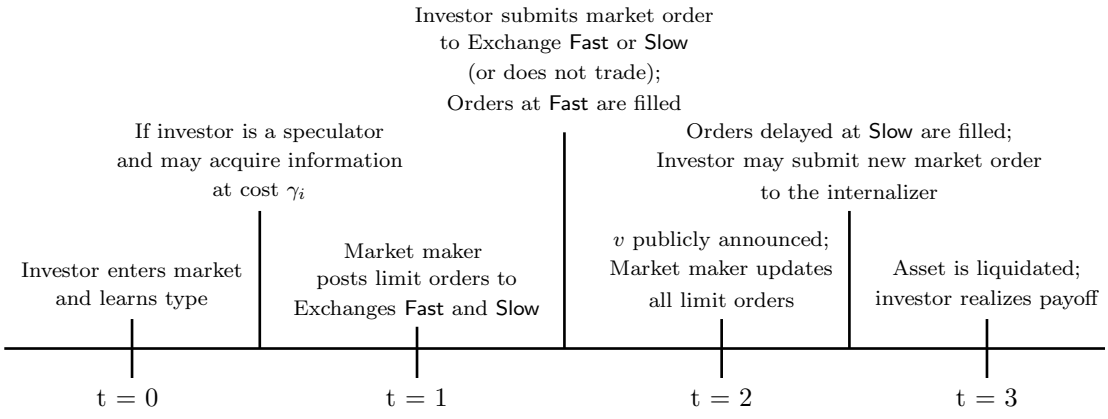
Investors. There is a unit mass of risk-neutral investors. At $t = 0$, an investor arrives at the market to trade a single unit of the security. The investor is either a speculator with probability $\mu > 0$, or an uninformed investor endowed with liquidity needs. Upon arrival, a speculator receives an information acquisition cost, $\gamma_i \sim \mathrm{U}[0, 1]$. Speculators may pay γ_i at $t = 0$ to perfectly learn the random payoff v. We refer to those who acquire information as "informed investors," and their mass is denoted $\mu_I \in (0, \mu]$. When information events are interpreted as fleeting arbitrage opportunities, speculators who acquire information can be viewed as acquiring the necessary technology to exploit these opportunities.

With probability $(1 - \mu)$, a liquidity investor arrives and is a buyer or seller with equal probability. Liquidity investors have no private information, but are endowed with a liquidity need that motivates them to trade. They pay an additional cost to trade following an adverse price movement. This cost, c_i, is proportional to the innovation such that $c_i = k\lambda_i\sigma$. $k \in (0, \infty)$ is a universal scaling parameter of the innovation, while λ_i is a private scaling parameter of the innovation, distributed uniformly on $[0, 1]$. The delay cost is similar to the delay cost imposed in Zhu (2014), and can represent a number of factors such as unmodelled risk aversion or a recapitalization cost. In both cases, this represents the cost uninformed investors pay when the price moves away from them and not if it moves in their favour.[7] Alternatively, the liquidity investor can elect not to trade and pay a constant delay cost

[7]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 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 the investor's payoff is realized at $t = 3$. Speculators face information acquisition cost γ_i, and liquidity investors face delay cost c_i.



$K \in (\sigma, \infty)$. We assume that the cost of not trading is large, such that $K > \max\{c_i\}$.

Investors place orders to maximize expected profits. An investor i may submit a single market order at $t = 1$ or $t = 2$, or not trade. An investor who submits an order at $t = 1$ must select one of the exchanges, while an investor who submits an order at $t = 2$ has the order sent to an off-exchange internalizer, which executes the order at the realized value v. Once an investor's order executes, any information acquired by the investor becomes public immediately, before any other trades occur. Thus, the investor has the incentive to submit only a single order.

Finally, the structure of the model is known to all market participants. We illustrate the timing of the model in Figure 1.

Investor Payoffs. The expected payoff to an investor who submits a buy order at $t = 1$ is given by his or her knowledge of the true value of v, minus the price paid and 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:

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

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

The scaling factor of $1/2$ in the delay cost of π_L reflects the fact that the asymmetric cost is incurred only if the price moves away from the liquidity investor, which occurs with probability $1/2$. An informed investor i who submits a buy order at period $t = 2$ (or elects not to trade) recovers no value from the information, and has a payoff of $-\gamma_i$ (speculators have payoff zero). 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 with a model where Exchange Slow imposes a delay. In our model, a market with two identical exchanges is functionally equivalent to a single competitive exchange.

We search for a perfect Bayesian equilibrium in which the market maker chooses a quoting strategy such that the market maker earns zero expected profits at each venue, and investors choose order submission strategies that maximize their profits. We also focus on equilibria where investors use both exchanges. Because the set-up of our model is symmetric for buyers and sellers, we focus our exposition on 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 Exchanges Fast and Slow to maintain consistency in notation. If both exchanges impose no processing delay ($\delta = 0$), 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:

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

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

Given that market orders are filled immediately at the posted quote, the expected profit for a liquidity investor who submits an order at $t = 1$ does not consider c_i directly; instead, the cost of c_i enters through the investor decision to trade at $t = 1$, or wait until uncertainty is resolved at $t = 2$ (for which they pay c_i if the price has moved against them).

The market maker populates the limit order books at Exchanges Fast and Slow, taking into account the expected order placement strategies by investors. The market maker quotes competitively, setting ask (bid) prices at $t = 1$ on each exchange to account for the expected adverse selection of an incoming buy (sell) order to that venue. We denote ask prices at Exchanges Fast and Slow at $t = 1$ as $\mathsf{ask}_1^{\mathsf{Fast}}$ and $\mathsf{ask}_1^{\mathsf{Slow}}$, respectively, and write them below:

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

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

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 its 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 so, to which exchange he or she submits an order. Speculators receive their information acquisition cost γ_i at $t = 0$, and weigh 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 the internalizer and pay cost c_i if the price has moved against them. 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 the informed investor chooses their order submission strategy over Exchanges Fast and Slow. We denote the probability with which an informed investor submits an order to Exchange Fast as $\beta \in (0, 1)$; otherwise, the investor submits an order to Exchange Slow. Similarly, a liquidity investor who chooses to trade in $t = 1$ submits an order to Exchange Fast with probability $\alpha \in (0, 1)$, and Exchange Slow otherwise. A buyer's order placement strategy over the two venues at $t = 1$ is characterized by:

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

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

We note here that, in the absence of direct impacts by γ_i and c_i, the sole determinant of venue choice for buyers is 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:

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

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

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, denoted μ_I, and the measure of liquidity investors who participate before $t = 2$, denoted $\Pr(c_i \geq \underline{c})$. Speculators receive γ_i in period $t = 0$, and decide whether paying their information acquisition cost is profitable. The mass of speculators who choose to acquire information determines μ_I. Similarly, all investors with $c_i \geq \underline{c}$ face a large enough cost of delay c_i, such that they trade prior to period $t = 2$.

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 who 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 $(1-\mu)\times\mathsf{Pr}\left(c_i \geq \underline{c}\right)$. 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.

An equilibrium in our model is characterized by: (i) investor participation measures, μ_I

and $\underline{c}$; (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 Slow, but because $\alpha = \beta$, these probabilities cancel out of the pricing equations (9)-(10), ensuring that the ask (and bid) prices of Exchanges 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 B (i.e., $\mathsf{ask}^{\mathsf{B}}$, $\mathsf{bid}^{\mathsf{B}}$).

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 $\delta \in (0, 1)$. We search for an equilibrium where investors use both exchanges. The

processing delay impacts payoffs to informed and liquidity investors differently. Informed investors face payoffs to Exchanges 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. As liquidity investors do not know v, their expectation of the true value is always v_0, and thus the processing delay does not impact their expectation of the future value. Instead, the uncertainty about the outcome of the price manifests in an asymmetrical cost to trading, c_i, which 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:

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

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

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 the possibility of a) losing (all or part of) their information if they are informed, or b) paying a delay cost to complete their trade if they are a liquidity investor. An investor's order placement strategy has two equilibrium conditions: i) an indifference condition (IC) between orders to Exchanges 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 Fast or Slow. 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 delay parameter $\underline{\lambda}$ at which they are indifferent to trading in $t = 1$ and waiting until $t = 2$ to trade at the internalizer. 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:

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

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

We can now describe our equilibrium. An equilibrium in our 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 Fast and Slow, 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)-(22). By solving this system, we arrive at the following theorem.

Theorem 2 (Existence and Uniqueness) *Let $\delta \in (0,1]$ and $k > \underline{k} = \frac{4((1-\mu)+2\mu\sigma)}{(1-\mu)+4\mu\sigma}$. Given $\beta \in (0,1]$, there exists unique values μ_I, $\underline{\lambda}, \bar{\lambda}$, and prices $\mathsf{ask}_1^{\mathsf{Fast}}, \mathsf{ask}_1^{\mathsf{Slow}}$ given by (17)-(18) that solve equations (19)-(22). Moreover, there exists a unique $\delta^* \in (0,1)$ such that: i) $\delta < \delta^* \Rightarrow \beta \in (0,1)$, and ii) $\delta \geq \delta^* \Rightarrow \beta = 1$.*

The nature of the equilibrium depends on both the total mass of liquidity traders with

high latency sensitivity (i.e., the magnitude of k) and the parametrization of the latency delay. Firstly, an equilibrium requires that the measure of liquidity traders with high latency sensitivity (i.e., a high $\lambda_i k$) be large enough such that not all liquidity traders leave the fast exchange. $\underline{k}$ is decreasing in both fundamental volatility (σ) and the measure of speculators, (μ), which drives the speculator's information acquisition: a lower measure of speculators or degree of fundamental volatility reduces a speculator's potential returns to acquiring information. With the subsequent reduction in adverse selection generated by the reduced information acquisition, the benefit of a potentially narrower spread at the delayed exchange does not outweigh the delay costs for sufficiently many liquidity traders. With a high enough k, high latency sensitive investors who strictly prefer the conventional exchange, even for small delays, incentivize information acquisition, even in the presence of low μ and σ.

Second, the parametrization of the delay affects where informed traders submit their orders: for a delay of sufficiently small size, informed traders will send orders to the delayed exchange, implying that market makers at the delayed exchange may not be entirely protected from adverse selection by the delay. However, for a large enough delay, informed traders migrate the entirety of their flow from the delayed exchange.

3 Impact of Latency Delays

When Exchange Slow imposes a latency delay, an investor who submits an order to Exchange Slow at $t = 1$ faces the possibility that private information may become public (i.e., the market maker will update its limit orders) before his or her 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. Unlike speculators, liquidity investors' preferences depend on their individual cost to delay. Those who have sufficiently

low delay costs are impacted more by the price of the order than the possibility of delay, and hence may prefer an exchange with a latency delay, if the price is sufficiently discounted. Because speculators' and liquidity investors' motives are not identical, the introduction of a latency delay segments the order flow of the two investor groups, to varying degrees.

Using the setting with two non-delayed exchanges as a benchmark ($\delta = 0$), which we refer to as the "benchmark case," we investigate the impact of a latency delay on market quality, price discovery and welfare. In what follows, we assume delay is sufficiently costly ($k > \underline{k}$), as required by Theorem 2. The latency delay δ^* takes on a special interpretation here: for all $\delta \geq \delta^*$, no informed traders submit orders to the delayed exchange ($\beta = 1$). We refer to this as the "segmentation point." 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 liquidity investors participate in the market at $t = 1$ ($\underline{\lambda} = 0$). Given these solutions, we solve equations (19)-(22) for δ^*, yielding the expression:

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

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

Proposition 1 (Order Flow Segmentation) *Compared with the benchmark case ($\delta = 0$), if Exchange* Slow *imposes a delay $\delta \in (0, 1)$,*

- *informed trading at Exchange* Slow *falls ($\beta \uparrow$); for $\delta \geq \delta^*$, informed traders use only Exchange* Fast *($\beta = 1$).*
- *orders sent to the internalizer by liquidity investors decrease ($\underline{\lambda} \downarrow$); for $\delta \geq \delta^*$, no liquidity investors use the internalizer ($\underline{\lambda} = 0$), and liquidity investors send more orders to Exchange* Fast *($\bar{\lambda} \downarrow$).*

Empirical Prediction 1 (Order Flow Segmentation) *If an exchange imposes a latency delay, its concentration of informed trading decreases while the concentration of informed trading on all other exchanges increases.*

While we find that $\beta = 1$ for all $\delta \geq \delta^*$, we do not predict full order flow segmentation of informed and liquidity investors, as liquidity 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 3.

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. 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 liquidity investors. Empirically speaking, existing work supports this fact and finds that exchanges with latency delays have lower informed trading and higher participation by uninformed orders (Chen et al. 2016; Anderson et al. 2018). In the context of the model, we predict that the increase in uninformed investor participation will come from investors who migrate from venues such as off-exchange internalizers, rather than from non-delayed exchanges.

In our model, the mechanism underlying the latency delay is that an order submitted to Exchange Slow may be delayed until after a public information announcement about the security. Thus, the market maker is afforded the opportunity to update its limit orders before the delayed order arrives, potentially avoiding adverse selection from informed investors. 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 *($\mathsf{ask}^{\mathsf{Slow}} \leq \mathsf{ask}^{\mathsf{B}}$) and wider at Exchange* Fast *($\mathsf{ask}^{\mathsf{Fast}} \geq \mathsf{ask}^{\mathsf{B}}$), when compared*

with the benchmark case ($\delta = 0$). For $\delta < \delta^*$, the spread widens at Exchange Fast *as δ increases, while for $\delta \geq \delta^*$, the spread narrows at Exchange* Fast *as δ increases.*

Empirical Prediction 2 (Quoted Spreads) *If an exchange imposes a latency delay, its quoted spreads will tighten, while the quoted spreads on all other exchanges widen.*

While the market maker may have the opportunity to update its quotes at the delayed exchange before an informed trade clears, it faces 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 5. Proposition 2 reflects some of the empirical results in Chen et al. (2016), who find that spreads worsen on non-delayed exchanges, driven by the redistribution of adverse selection.

The imposition of a delay at Exchange Slow and the subsequent improvement in quotes at Exchange Slow informs our trading volume result: liquidity investors with moderate delay costs who would submit an order to the off-exchange internalizer at $t = 2$ now choose to submit orders to Exchange Slow at $t = 1$. To show this, we define total exchange trading volume (Volume) as the probability that an investor who enters, submits an order at $t = 1$:

$$\text{Volume} = \mu\bar{\gamma} + (1 - \mu) \times (1 - \underline{\lambda}). \tag{24}$$

The right panel of Figure 4 illustrates that, as liquidity investors increase their participation, the migration of informed traders to Exchange Fast and the resulting worsening of quoted spreads at Exchange Fast lead to a decline in information acquisition by speculators. The net effect, however, is an increase in total exchange volume. We summarize this result in the proposition below.

Proposition 3 (Total Exchange Volume) *Compared with the benchmark case ($\delta = 0$), if Exchange* Slow *imposes a delay $\delta \in (0, 1)$, then liquidity investor participation improves ($\underline{\lambda} \downarrow$), and information acquisition falls ($\bar{\gamma} \downarrow$). The net effect is an increase in total exchange trading volume.*

Empirical Prediction 3 (Total Exchange Volume)

- *If one exchange imposes a latency delay, total exchange-traded volume increases.*
- *If one exchange imposes a latency delay, total exchange-traded volume from informed traders decreases, while total exchange-traded volume from liquidity traders increases.*

The latency delay affects incentives for both liquidity investors and speculators. For liquidity investors, the improved prices offered at Exchange Slow incentivize those with moderate delay costs to select the latency-delayed exchange over the internalizer, reducing adverse selection and improving quotes by the market maker. For speculators, 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 information acquisition by speculators.

If the delay is sufficiently long ($\delta \geq \delta^*$), all informed traders segregate to the non-delayed exchange, and all liquidity investors trade on-exchange. At this point, a 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 his or her delay cost. Thus, for any $\delta \geq \delta^*$, liquidity investors migrate from Exchange Slow to Exchange Fast (see Figure 3). For long delays, such that orders at Exchange Slow trade at $t = 2$ with probability one ($\delta = 1$), the measure of informed traders and liquidity traders at the non-delayed exchange reflects the case where no delayed-exchange exists ($\delta = 0$).

Proposition 3 predicts that imposing a delay at Exchange Slow siphons volume away from off-exchange internalizers. The subsequent increase in exchange volume and relative changes in price impacts through spreads at both exchanges raise a question as to the impact of the delay on the permanent price impact of trades (price discovery) at $t = 1$. We examine whether the use of delays to protect market makers from adverse selection impacts the speed of price discovery: do speculators reduce the information they bring to market ahead of market makers?

To study price discovery in our setting, we use the proportional pricing error measure of price discovery in Zhu (2014), defined as the root-mean-squared error of the fundamental value and the expected permanent price impact at $t = 1$, conditional on a positive innovation, σ, scaled by σ.

$$\mathsf{Price\ Discovery}_{t=1} = \frac{\sqrt{\mathsf{E}[(v - \text{price impact})^2 \mid v = v_0 + \sigma]}}{\sigma}. \tag{25}$$

Equation (25) reduces to a function of quotes at $t = 1$, as the absence of non-informational transaction costs in our model implies that quotes are equal to their price impact. We can then write price discovery explicitly:

$$\mathsf{Price\ Discovery}_{t=1} = \sqrt{1 - \frac{\mu\bar{\gamma}^*(\beta^* \cdot \mathsf{ask}^{\mathsf{Fast}*} - (1-\delta)(1-\beta^*) \cdot \mathsf{ask}^{\mathsf{Slow}*})}{\sigma}}. \tag{26}$$

Compared with the benchmark case, Proposition 3 predicts that the availability of a delayed exchange unequivocally reduces information acquisition by speculators (and their subsequent market participation). This in itself does not necessarily predict how efficiently the acquired information is impounded into prices. We find that the impact of a delay on price discovery depends on both the relative concentration of speculators and the length of a delay. For high concentrations of speculators, the result is an unambiguous fall in price discovery. However, when liquidity traders arrive more often, the length of the delay plays a key role. Indeed, in some cases, the introduction of a delay may increase price discovery. We illustrate these results in the right panel of Figure 5.

While an increase in price discovery may appear counter-intuitive, breaking price discovery into its constituent components reveals the cause. An informed investor's contribution to permanent price impact has three components: i) the probability of information acquisition; ii) the likelihood of a non-delayed trade by an informed investor; and iii) their price impact, measured by the bid-ask spread. If the concentration of speculators is relatively low, the probability of information acquisition decreases when a delay is introduced. This decrease

is less pronounced than when the concentration of speculators is higher. When $\delta > \delta^*$, these informed investors are concentrated at Exchange Fast, where they face no possibility of delay, and thus contribute to price discovery with certainty. The net result is that for low values of μ, there exists a sufficient long delay ($\delta > \hat{\delta}$), such that price discovery increases.

Numerical Observation 1 (Price Discovery) *Compared with the benchmark case* ($\delta = 0$)*, the impact of a delay on price discovery is dictated by the measure of speculators,* μ*:*

- *for sufficiently low* μ*, there exists a* $\hat{\delta}$ *such that any* $\delta > \hat{\delta}$ *improves price discovery;*
- *for higher* μ*, price discovery worsens for any delay.*

Our prediction on price discovery illustrates that speculators are squeezed from both sides by a latency delay: while the delayed exchange offers an improved quote due to increased liquidity trading (and lower informed trading), the probability that informed traders realize the additional profit falls with the length of the delay. Moreover, by retreating to the non-delayed exchange en masse, adverse selection worsens, and the quoted spread widens.

Beyond price discovery, delays are touted by exchanges as a way to protect market makers from adverse selection, and thus improve trading costs for retail and institutional investors. We measure these trading costs using the liquidity investor's expected payoff. From a broader welfare perspective, this represents only half of the story, as speculators who become informed (and thus trade) spend resources on information acquisition. Thus, a welfare measure defined in the sense of allocative efficiency (e.g., Bessembinder, Hao, and Zheng 2015) does not merely net out to a sum of private values (or delay costs, in our case). Instead, the measure includes the additional information acquisition cost. In our framework, the expected profits to the average investor (total gains from trade) is a cost-minimization

measure, which we express as:

$$\mathsf{W} = \int_{\bar{\lambda}}^{1} \left(\mathsf{ask}_1^{\mathsf{Fast}} - v_0\right) \mathrm{d}\lambda + \int_{\underline{\lambda}}^{\bar{\lambda}} \left((1-\delta)(\mathsf{ask}_1^{\mathsf{Slow}} - v_0) + \delta\frac{k\sigma}{2}\lambda\right) \mathrm{d}\lambda + \int_{0}^{\underline{\lambda}} \frac{k\sigma}{2}\lambda \mathrm{d}\lambda$$
$$+ \mu\bar{\gamma}\left(\beta(v_0 + \sigma - \mathsf{ask}_1^{\mathsf{Fast}}) + (1-\beta)\left((1-\delta)(v_0 + \sigma - \mathsf{ask}_1^{\mathsf{Slow}})\right) - \int_{0}^{\bar{\gamma}} 1 \mathrm{d}\gamma\right) \tag{27}$$
$$= \mu\bar{\gamma}^2 + (1-\mu) \times \left(\delta(\bar{\lambda}^2 - \underline{\lambda}^2) + \underline{\lambda}^2\right) \times \frac{k\sigma}{4}. \tag{28}$$

We now examine how welfare is impacted by the introduction of an exchange with a latency delay, δ. Our result is numerical, which we present graphically in Figure 6.

Numerical Observation 2 (Expected Welfare) *In comparison with the benchmark case ($\delta = 0$), the impact of a delay on expected welfare is dictated by the measure of speculators, μ. There exists values $0 < \underline{\mu} < \overline{\mu} < 1$ such that:*

- *for $\mu < \underline{\mu}$, expected welfare worsens with any delay;*
- *for $\mu \in [\underline{\mu}, \overline{\mu}]$, the impact on expected welfare is ambiguous;*
- *for $\mu > \overline{\mu}$, expected welfare improves for any delay.*

The introduction of a delay (weakly) increases the delay costs paid by liquidity investors and (weakly) decreases the information costs paid by speculators who pay to become informed. The welfare impact depends on whether the increase in delay costs dominates the decrease in information acquisition costs, or vice versa. We illustrate these relative costs on the right panel of Figure 6. For extreme concentrations of speculators, the results are unambiguous. When there are very few speculators, the increase in delay costs dominates and welfare falls; when there are many speculators, the decrease in information acquisition costs dominates and welfare increases. For more moderate values of μ, the impact on welfare depends on the length of the delay.

The parameter-dependent impact on investor welfare is driven by the negative correlation between a liquidity investor's expected delay costs and the expected information acquisition costs. Despite the fact that the introduction of a delayed exchange incentivizes more liquidity

investors to submit orders to exchanges at $t = 1$, these investors have the lowest relative delay costs. They enter the market through the delayed exchange and, therefore, do not fully mitigate their delay costs. Moreover, average delay costs for investors already submitting orders to Exchange Slow are higher in the presence of a delay.

Conversely, speculators are less willing to pay high information acquisition costs when delays are imposed. Informed investors migrate to Exchange Fast, thereby increasing price impact (and thus lowering information rents). This increase in price impact peaks at δ^*, where informed investors are fully segmented to Exchange Fast. For any longer delay, high-delay cost liquidity investors migrate to Exchange Fast, improving liquidity through a reduction in $\mathsf{ask}_1^{\mathsf{Fast}}$, which incentivizes higher levels of spending on information acquisition.

4 Exchange Competition and Optimal Latency Delays

Our empirical predictions on market quality focus on a two-marketplace setting, where one exchange exogenously imposes a speed bump of length δ. However, it is not immediately apparent that an exchange would optimally choose to impose a speed bump, or if they did, what level of delay is optimal. In Subsection 2.1, we show that two identical exchanges can coexist for any split of total market share. This result motivates the incentive for imposing a delay: by providing a differentiated product, a small exchange may be able to siphon a significant share of order flow away from a dominant exchange, increasing its presence in the market. In this section, we analyze the choice of an exchange to impose a delay in a setting where another non-delayed market participates. We assume exchanges maximize profits by maximizing total trading volume. We contend that this assumption correlates with profit maximization (abstracting from data sales, co-location services, etc.), as profits from order flow arise from exchange fees, which are paid only upon execution of a trade.

When examining the profit maximization motive of an exchange, we acknowledge that while some delayed exchanges (e.g., IEX, Aequitas Neo) operate as stand-alone venues, other

delayed exchanges are subsidiaries of larger, parent exchanges (e.g., TSX Alpha, NYSE MKT). The industrial organization differences between these two settings necessitate different approaches to profit maximization. A stand-alone venue imposes a delay (if any) that maximizes its own volume, regardless of the overall market impact. A subsidiary exchange, however, would be reluctant to impose a delay that siphons order flow away from its parent exchange, especially if this order flow migration results in fewer trades. Hence, we assume that a subsidiary exchange imposes a delay (if any) that maximizes the joint volume across both the delayed exchange and the parent exchange.

Suppose that the exchange considering to impose a delay operates independently from the standard exchange. The goal is to siphon sufficient order flow away from the standard exchange and the order internalizer, such that their total volume exceeds the level for $\delta = 0$. For a delayed exchange, any delay $\delta \leq \delta^*$ has three effects on volume: i) informed investors migrate to the non-delayed exchange; ii) speculators reduce their information acquisition; and iii) liquidity traders migrate to the delayed exchange from the internalizer. At δ^*, the spread at the delayed exchange reaches the lower bound $\mathsf{ask}^{\mathsf{Slow}} = 0$, implying that informed order flow is zero. Then, for any $\delta > \delta^*$, liquidity traders that send orders to the delayed exchange do not become "cheaper", but do face higher delay costs. The result is that volume is lower at the delayed exchange for all $\delta > \delta^*$ when compared with $\delta \leq \delta^*$, where volume is maximized.

For a subsidiary of a standard exchange, the goal of imposing a delay is to maximize total volume across both exchanges. The effects are similar to the previous case, except the cross-exchange migration has a net-zero effect. Instead, the exchange is concerned with i) reduced information acquisition by speculators; and ii) emigration of liquidity traders from the internalizer. We find that, while information acquisition is lower for all $\delta \in [0, \delta^*]$, the order flow siphoned from the internalizer exceeds the loss of informed order flow. Moreover, for $\delta > \delta^*$, the migration of liquidity traders to the non-delayed exchange (from the delayed exchange) reduces adverse selection on the non-delayed exchange, incentivizing an increase

in speculator information acquisition (and informed order flow). Thus, the optimal delay length for this case is $\delta = 1$. We summarize these findings in the proposition below.

Proposition 4 (Optimal Latency Delay) *If the delayed exchange is independent of the non-delayed exchange, the delay that maximizes volume on Exchange* Slow *is given by* $\delta = (0, \delta^*]$. *If the delayed exchange is a subsidiary of the non-delayed exchange, then the delay that maximizes joint volume on Exchanges* Fast *and* Slow *is* $\delta = 1$.

The optimal delay differs substantially based on the ownership structure of the two exchanges. When a delayed exchange is operated separately from the other venues, it selects a shorter delay to profit from the net migration impact of the delay. The delayed exchange is motivated to provide an option for liquidity traders who care about price improvement but also want timely execution. The non-delayed exchange offers instantaneous execution, but this also attracts informed traders, and thus adverse selection and worse quotes. Here, the independent exchange has an opening in the exchange "product space" to offer a sufficiently long delay to disincentivize informed traders, while providing a reduction in delay costs. The result is a delay, $\delta \in (0, \delta^*]$.

We observe that total volume at the delayed exchange is constant for all δ in the range $\delta \in (0, \delta^*]$, as any outflow of informed traders to the non-delayed exchange is exactly offset by an inflow of liquidity traders from the internalizer. In light of this, an exchange seeking to choose an optimal delay may elect the longest delay in the range, δ^*. This is in keeping with the motivation of delayed exchanges to limit adverse selection, as δ^* fully segments informed traders to the non-delayed exchange.

If both exchanges are operated by a single firm, the optimal delay is the maximum length, $\delta = 1$, such that no trades occur at the delayed exchange before the market maker can update its quotes. The firm is motivated to effectively open an internalizer of its own, providing a venue to which low-delay-cost investors will send orders. The firm prefers the maximum delay over any shorter delay, as it not only incentivizes low-delay-cost liquidity traders to migrate to the delayed exchange, but also maintains the presence of high-delay-cost liquidity

traders at its non-delayed exchange, thus ensuring maximum speculator participation. Taken together, these effects maximize the total number of orders.

Proposition 4 is important in the context of our results on price discovery and welfare. Figures 5 and 6 illustrate that, for relatively low concentrations of speculators, there exists a delay where both price discovery and investor welfare may improve. However, this region exists only for $\delta \in (\delta^*, 1)$. Thus, Proposition 4 implies that introducing a delayed venue as either a competitor to a non-delayed venue or as a subsidiary of the non-delayed exchange to provide "product differentiation" will not yield a delay length that is both welfare and price-discovery improving.

The previous discussion suggests that delayed exchanges—whether independent or a subsidiary—will select some positive level of delay as part of a volume-maximization strategy. However, this optimal delay value is impacted by the fundamental volatility of the particular security. We proxy this volatility by the innovation to the security's value, σ.

Proposition 5 (Adverse Selection) *The segmentation point $\delta^*(\sigma)$ is increasing in σ.*

Empirical Prediction 4 (Adverse Selection and Exchange Volume) *As adverse selection increases, total exchange-traded volume falls, but volume at the delayed exchange increases.*

Proposition 5 predicts that fundamental volatility explicitly impacts only the decision of the speculator to acquire information: as σ increases, it is more profitable to acquire information, so $\bar{\gamma}$ also increases. The outcome is that δ^* increases in σ: for high-volatility stocks, more information acquisition occurs, increasing adverse selection on Exchange Fast, and slowing the migration of informed investors from Exchange Slow, when δ increases towards δ^*. Empirically, our model predicts that higher fundamental volatility should result in lower total exchange volume, driven by the reduction in volume at the non-delayed exchange. Conversely, volume on the delayed exchange weakly increases.

5 Conclusion

Latency delays are one of the latest means by which exchanges differentiate themselves. These delays are introduced to segment uninformed order flow from the broader market, by preventing informed traders acting on fleeting information. We find that latency delays have a mixed impact on market liquidity: the imposition of a delay improves liquidity on the delayed exchange, but worsens liquidity on standard exchanges. Moreover, the presence of a delayed exchange reduces overall information acquisition, but the subsequent impact on price discovery depends on the ratio of speculators to liquidity traders: with a greater presence of speculators, a delayed exchange worsens overall price discovery, whereas markets with fewer speculators see price discovery improvements when one market imposes a delay.

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. Finally, we predict that as adverse selection increases, total exchange volume falls, while delayed exchange volume increases.

Of interest to policy makers, the impact of delays on price discovery and welfare depend on the relative concentration of speculators. The presence of a delay either decreases welfare when there are few speculators, or increases welfare when there are many. Results for price discovery are reversed: price discovery falls when there are many speculators, but may increase if there are few. Depending on whether a regulator prioritizes welfare or price discovery, the regulator may wish to allow delays for some assets, while disallowing for others.

References

Anderson, Lisa, Emad Andrews, Baiju Devani, Michael Mueller, and Adrian Walton, 2018, Speed segmentation on exchanges: Competition for slow flow, *Bank of Canada Staff Working Paper No. 2018-3*.

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, *SSRN Working Paper 2782692*.

Bessembinder, Hendrik, Jia Hao, and Kuncheng Zheng, 2015, Market making contracts, firm value, and the IPO decision, *Journal of Finance* 70, 1997–2028.

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, *SSRN Working Paper 2435999*.

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, *SSRN Working Paper 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, *SSRN Working Paper 2328231*.

Chen, Haoming, Sean Foley, Michael A Goldstein, and Thomas Ruf, 2016, The value of a millisecond: Harnessing information in fast, fragmented markets, *SSRN Working Paper 2890359*.

Cimon, David A, 2016, Broker routing decisions in limit order markets, *Bank of Canada Staff Working Paper No. 2016-50*.

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.

Glosten, Lawrence, and Paul R Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.

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

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, *SSRN Working Paper 2542184*.

Malinova, Katya, and Andreas Park, 2016, Does high frequency trading add noise to prices?, *Working Paper*.

Menkveld, Albert J, and Marius A Zoican, 2017, Need for speed? Exchange latency and liquidity, *Review of Financial Studies* 30, 1188–1228.

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, *SSRN Working Paper 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, *SSRN Working Paper 2066839*.

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

A Appendix

This appendix includes a notation index, a description of the mechanics underlying latency delays, and all proofs and figures not presented in-text.

A.1 List of Variables and Parameters

Variable	Description
v	asset fundamental value at period $t = 3$
v_0	public value at period t
δ	probability that an order is filled after v is announced
σ	absolute value of the innovation to the public prior at $t = 3$
μ	total mass of speculators
μ_I	mass of speculators who acquire information at $t = 0$
c_i	(private) costs of delay to liquidity investor i
k	universal scaling component of the costs of delay c_i
λ_i	private scaling component of the costs of delay c_i
K	cost paid by a liquidity investor who does not submit an order
γ_i	(private) information acquisition costs to speculator i
Fast	denotes the exchange without a latency delay
Slow	denotes the exchange with a latency delay
$\mathsf{ask}_t^{\mathsf{Fast}}$	ask price at Exchange Fast at period t
$\mathsf{ask}_t^{\mathsf{Slow}}$	ask price at Exchange Slow at period t
$\mathsf{bid}_t^{\mathsf{Fast}}$	bid price at Exchange Fast at period t
$\mathsf{bid}_t^{\mathsf{Slow}}$	bid price at Exchange Slow at period t
π_I	profit function for an informed investor
π_L	profit function for a liquidity investor
β	probability that an informed investor submits an order to Exchange Fast
α	probability that a liquidity investor submits an order to Exchange Fast
B	denotes the value for the benchmark case ($\delta = 1$)

A.2 Latency Delays

Broadly speaking, a latency delay is the imposition of an intentional delay on some or all incoming orders by a trading venue. 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 350 microseconds in advance. However, IEX allows traders to post "pegged" orders, which move instantaneously in response to external factors. The pegged orders at IEX are available in multiple forms, but the one most relevant to this paper is the "discretionary peg." This order type uses an algorithm to determine if a price movement is likely, a behaviour IEX refers to as a "crumbling quote."[8] 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. Since these pegged orders move instantaneously following trades at other exchanges, market makers using these orders receive some protection from multi-market trading strategies.

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 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 allow market makers to fade their quotes, removing liquidity before any large order reaches the exchange. This second form of delay is used on the Canadian exchange TSX Alpha, where orders entering the exchange are delayed by a period

[8]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`

of 1 to 3 milliseconds. 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.[9]

Finally, a third type of latency delay explicitly classifies traders into two groups. Some traders are affected by the delay, and have their orders delayed. Other traders are not affected and trade normally. Unlike the other two types of delays that rely on order types, this form requires the explicit division of traders into two types by the exchange. This form of delay is used on the Canadian exchange Aequitas NEO, which divides traders into Latency Sensitive Traders, who are affected by the delay, and non-Latency Sensitive Traders, who are not.[10] Those deemed to be "latency sensitive" are subjected to a randomized delay of between 3 to 9 milliseconds.

A.3 Proofs

Proof (Theorem 1). The proof that follows focuses on the actions of buyers; sellers' decisions are symmetric. Informed (I) and liquidity (L) investors who submit an order at $t = 1$ to Exchange j have profit functions given by:

$$\pi_I(\gamma_i; \text{Buy at t=1}) = v - \mathsf{ask}_1^j - \gamma_i, \tag{29}$$

$$\pi_L(c_i; \text{Buy at t=1}) = v_0 - \mathsf{ask}_1^j. \tag{30}$$

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

[10] 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`

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

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = v_0 + \frac{\beta\mu_I}{\beta\mu_I + (1-\mu)\alpha\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma, \tag{31}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = v_0 + \frac{(1-\beta)\mu_I}{(1-\beta)\mu_I + (1-\mu)(1-\alpha)\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma. \tag{32}$$

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 $\underline{c}$:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = \mathsf{ask}_1^{\mathsf{Slow}}, \tag{33}$$

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

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

Given equilibrium prices in (31) and (32), we solve for μ_I and $\underline{c}$. Because $\mu_I = \mu\bar{\gamma}$, we can solve for $\bar{\gamma}$ and immediately obtain μ_I:

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

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

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 $\underline{c}$:

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

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

Proof (Theorem 2). The proof of Theorem 2 proceeds similarly to Theorem 1, except that we solve liquidity investor strategies, characterized by $\bar{\lambda}$ and $\underline{\lambda}$. We prove this theorem by examining the equilibrium through the informed investor's strategy, β.

Speculators use only Exchange Slow ($\beta = 0$): Consider the informed investor's incentive compatibility constraint, evaluated at $\beta = 0$.

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

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 system of equations from (19)-(22) for $\bar{\lambda}, \underline{\lambda}, \bar{\gamma}$ and β, for $\beta \in (0,1)$. We write them explicitly as:

$$\text{IC}_I\text{: } 1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\left(1 - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})}\right) = 0, \tag{44}$$

$$\text{PC}_I\text{: } \bar{\gamma} - \sigma\left(1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})}\right) = 0, \tag{45}$$

$$\text{IC}_L\text{: } \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} - \frac{\delta k\bar{\lambda}}{2} = 0, \tag{46}$$

$$\text{PC}_L\text{: } \frac{k\underline{\lambda}}{2} - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} = 0. \tag{47}$$

We begin first by rearranging (44) to solve for δ:

$$\delta = \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})}. \tag{48}$$

Substituting equation (48) into (46) and simplifying yields the solution $\bar{\lambda} = 2/k$.

Now, we show that $\bar{\gamma}^* \in [0,1]$ exists for all $(\beta, \underline{\lambda}) \in (0,1) \times [0, 2/k]$, by appealing to the intermediate value theorem:

$$\text{PC}_I \mid_{\bar{\gamma}=0} : 0 - \sigma < 0, \tag{49}$$

$$\text{PC}_I \mid_{\bar{\gamma}=1} : 1 - \sigma\left(1 - \frac{\mu\beta}{\mu\beta + (1-\mu)(1-\bar{\lambda})}\right) > 0, \tag{50}$$

where (50) holds by the fact that $\sigma < 1$. Thus, $\bar{\gamma}^* \in [0,1]$ exists. To show that $\bar{\gamma}^*$ is unique, we solve (45) for the non-negative root of $\bar{\gamma}$:

$$\bar{\gamma} = \frac{\sqrt{(1-\mu)^2(1-2/k)^2 + 4(1-\mu)(1-2/k)\mu\beta\sigma} - (1-\mu)(1-2/k)}{2\mu\beta}. \tag{51}$$

We can see that $\bar{\gamma}^*$ is unique, and is bounded within $[0,1]$, as the limit for $\mu = 0$ can be solved by inspection of (45), as $\mu \to 0 \implies \bar{\gamma}^* = \sigma < 1$.

We now appeal to the intermediate value theorem using (47) to show that $\underline{\lambda} \in [0, \bar{\lambda}]$ exists for all $\beta \in (0,1)$, given $\bar{\gamma}^*$. $\underline{\lambda}$ is bounded above by $\bar{\lambda} = 2/k$.

$$\text{PC}_L \mid_{\underline{\lambda}=0} = 0 - \frac{\mu\bar{\gamma}^*(1-\beta)}{\mu\bar{\gamma}^*(1-\beta) + (1-\mu) \times 2/k} < 0, \tag{52}$$

$$\text{PC}_L \mid_{\underline{\lambda}=2/k} = 0, \tag{53}$$

where $\bar{\gamma}^*$ is a function of β, and parameters. Hence, $\underline{\lambda} \in [0, \bar{\lambda}]$ exists.

To show that $\underline{\lambda}$ is unique, we take the first derivative of PC_L.

$$\frac{\partial}{\partial \underline{\lambda}}(\text{PC}_L) = \frac{k}{2} - \frac{\mu\bar{\gamma}^*(1-\beta)(1-\mu)}{(\mu\bar{\gamma}^*(1-\beta) + (1-\mu)(2/k - \underline{\lambda}))^2}. \tag{54}$$

Because we cannot sign (54), we take the second derivative to show that (47) crosses zero

from below at most once.

$$\frac{\partial^2}{\partial \underline{\lambda}^2}(\mathrm{PC}_L) = -\frac{2\mu\bar{\gamma}^*(1-\beta)(1-\mu)^2}{(\mu\bar{\gamma}^*(1-\beta)+(1-\mu)(2/k-\underline{\lambda}))^3} < 0. \tag{55}$$

Because (55) is negative, (47) must cross zero from below at most once on $\underline{\lambda} \in [0, 2/k]$. Hence, $\underline{\lambda}^*$ is unique.

Lastly, we show that there is a unique $\beta^* \in (0,1)$ that solves (43), given $(\bar{\gamma}^*, \bar{\lambda}^*, \underline{\lambda}^*)$.

$$\mathrm{IC}_I\,|_{\beta=0}: 1-(1-\delta)\frac{\mu\sigma}{\mu\sigma+(1-\mu)(2/k-\underline{\lambda}^*)} > 0, \tag{56}$$

$$\mathrm{IC}_I\,|_{\beta=1}: \delta - \frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^*+(1-\mu)(1-2/k)} < 0,\ \forall \delta < \frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^*+(1-\mu)(1-2/k)} = \bar{\delta}. \tag{57}$$

Thus, by the intermediate value theorem, for all $\delta < \bar{\delta}$, there exists a $\beta \in (0,1)$ that satisfies (44). To show that β^* is unique, we differentiate (44) with respect to β.

$$\frac{\partial}{\partial\beta}(\mathrm{IC}_I) = -\frac{2(1-\mu)\sigma^2(k-2)^2\left(((1-\delta)/2)\times\sqrt{h(\beta)}+r(\beta)\right)\mu}{\sqrt{h(\beta)}(\sqrt{h(\beta)}+(k-2)(1-\mu))^2} < 0, \tag{58}$$

$$\text{where: } h(\beta) = (1-\mu)(k-2)(4\mu\beta k\sigma+(1-\mu)(k-2)) > 0, \tag{59}$$

$$r(\beta) = k(1-\delta)\left(\frac{(1-\mu)}{2}+\sigma(1+\beta)\mu\right)+(1+\delta)(1-\mu) > 0. \tag{60}$$

Thus, β^* is unique. Finally, we show that $\beta \in (0,1) \Rightarrow \delta < \bar{\delta}$. Let $\delta \geq \bar{\delta}$, and suppose $\beta \in (0,1)$. We know that $\beta \in (0,1) \Rightarrow \bar{\lambda} = 2/k$. Because equation (44) is decreasing in β and increasing in δ, there cannot be a solution $\beta \in (0,1)$ to the right of $\bar{\delta}$, given that $\beta(\bar{\delta}) = 1$. Thus, $\beta \in (0,1)$ if and only if $\delta < \bar{\delta}$. Moreover, by simplifying (57), we can write $\bar{\delta}$ in terms of parameters only:

$$\bar{\delta} = \frac{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2+(1-\mu)(1-\frac{2}{k})\mu\sigma}-(1-\mu)(1-\frac{2}{k})}{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2+(1-\mu)(1-\frac{2}{k})\mu\sigma}+(1-\mu)(1-\frac{2}{k})}. \tag{61}$$

Speculators use only Exchange Fast ($\beta = 1$): Here, we solve equations (19)-(22) for

the case where $\beta = 1$. Inputting $\beta = 1$, we have:

$$\text{IC}_I\text{: } \delta - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} \geq 0, \tag{62}$$

$$\text{PC}_I\text{: } \bar{\gamma} - \sigma\left(1 - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})}\right) = 0, \tag{63}$$

$$\text{IC}_L\text{: } \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} - \frac{\delta k\bar{\lambda}}{2} = 0, \tag{64}$$

$$\text{PC}_L\text{: } \frac{\delta k\underline{\lambda}}{2} = 0. \tag{65}$$

First, by inspection of (65), we see that $\underline{\lambda}^* = 0$. To prove the existence of a unique $\bar{\gamma}$, we solve equation (63) for the non-negative root of $\bar{\gamma}$:

$$\bar{\gamma}^* = \frac{\sqrt{(1-\mu)^2(1-\bar{\lambda})^2 + 4(1-\mu)(1-\bar{\lambda})\mu\sigma} - (1-\mu)(1-\bar{\lambda})}{2\mu}. \tag{66}$$

By inspection, $\bar{\gamma}^*$ exists and is unique as long as the limit $\mu \to 0$ exists, and is in the interval [0,1]. By simply setting $\mu = 0$, (63) admits the limit $\bar{\gamma} = \sigma$. Thus, $\bar{\gamma}^*$ is unique.

Lastly, we show that there exists a unique $\bar{\lambda} \in [0, 2/k]$ that solves (64) for all $\delta \geq \underline{\delta}$. We can bound $\bar{\lambda} \in [0, 2/k]$ because for any $\bar{\lambda} > 2/k$, (64) would be negative if the required inequality in (62) holds. First, we show that $\bar{\lambda}$ exists by evaluating $\bar{\lambda}$ at 0 and $2/k$:

$$\text{IC}_L\mid_{\bar{\lambda}=0}\text{: } \frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^* + (1-\mu)} - 0 > 0, \tag{67}$$

$$\text{IC}_L\mid_{\bar{\lambda}=2/k}\text{: } \frac{\mu\bar{\gamma}^*(2/k)}{\mu\bar{\gamma}^*(2/k) + (1-\mu)(1-2/k)} - \delta < 0, \tag{68}$$

$$\forall\delta > \frac{\mu\bar{\gamma}^*(2/k)}{\mu\bar{\gamma}^*(2/k) + (1-\mu)(1-2/k)}.$$

Hence, by the continuity of (64) in $\bar{\lambda}$, $\bar{\lambda}^*$ exists for all $\delta \geq \frac{\mu\bar{\gamma}^*(2/k)}{\mu\bar{\gamma}^*(2/k)+(1-\mu)(1-2/k)} = \underline{\delta}$. To show that $\bar{\lambda}^*$ is unique, we show that IC_L is decreasing in $\bar{\lambda}$, which ensures that IC_L crosses zero from above only once for any $\delta > \underline{\delta}$. Differentiating (64) with respect to $\bar{\lambda}$:

$$\frac{\partial}{\partial\bar{\lambda}}(IC_L) = \frac{\mu\bar{\gamma}(1-\mu)}{(\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial\bar{\gamma}}{\partial\bar{\lambda}} \cdot \frac{\mu(1-\bar{\lambda})(1-\mu)}{(\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda}))^2} - \frac{\delta k}{2}. \tag{69}$$

For (69) to be negative, first note that condition (62) holds only for $\delta \geq \frac{\mu\bar{\gamma}}{\mu\bar{\gamma}+(1-\mu)(1-\bar{\lambda})}$. Thus, input $\delta = \frac{\mu\bar{\gamma}}{\mu\bar{\gamma}+(1-\mu)(1-\bar{\lambda})}$ into (69). Computing the derivative and simplifying, we achieve the inequality:

$$\frac{\partial}{\partial\bar{\lambda}}(IC_L) < -\frac{2(1-\mu)(1-\bar{\lambda}) \times \left((k \cdot v(\bar{\lambda}) - 2(1-\mu)\right)\mu\sigma}{v(\bar{\lambda})\left((1-\mu)^2(1-\bar{\lambda})^2 + v(\bar{\lambda})\right)^2}, \tag{70}$$

where $v(\bar{\lambda}) = \sqrt{(1-\mu)(1-\bar{\lambda})\left((1-\mu)(1-\bar{\lambda}) + 4\mu\sigma\right)}$. Then, expression (70) is negative if and only if $v(\bar{\lambda}) > \frac{2}{k} \cdot (1-\mu) \iff k > \frac{4(1-\mu+2\mu\sigma)}{1-\mu+4\mu\sigma} = \underline{k}$, which is our assumed lower bound on k. This result yields that for all $k > \underline{k}$, $\underline{\delta}$ is the lowest such δ that a solution to (64) exists. Thus, for all $k > \underline{k}$, $\bar{\lambda}^*$ exists and is unique if and only if $\delta \in [\underline{\delta}, 1]$. Moreover, by inspection, $\underline{\delta} = \bar{\delta} = \delta^*$. ■

Proof (Proposition 1). We begin by showing that $\beta \geq \beta_{\delta=0}$ and $\underline{\lambda} \leq \underline{\lambda}_{\delta=0}$ for all $\delta \in (0, \delta^*)$. For $\delta \in (0, \delta^*)$, we know that $\frac{\partial\beta}{\partial\delta} > 0$ from the proof of Theorem 2.

To show that $\underline{\lambda}$ is decreasing in δ, we note that $\frac{\partial\bar{\gamma}}{\partial\beta} > 0$ from the proof of Theorem 2. Thus, by the product of the partial derivatives, we know that $\bar{\gamma}$ is decreasing in δ. We also know that $\frac{\partial \mathsf{ask}^{\mathsf{Slow}}}{\partial\delta} < 0$. To show this, we note that given the existence of unique equilibrium, the following condition in $\underline{\lambda}$ must be satisfied:

$$\frac{k\underline{\lambda}}{2} - \mathsf{ask}^{\mathsf{Slow}} = 0 \iff \frac{k\underline{\lambda}}{2} - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda}-\underline{\lambda})} = 0. \tag{71}$$

For this condition to hold, it must be decreasing in $\underline{\lambda}$ for a decrease in $\bar{\gamma}$ and an increase in β, which follow from an increase in δ.

Now, let $\delta \in [\delta^*, 1]$. We obtain $\beta = 1$ through the proof of Theorem 2. To show $\bar{\lambda}$ is decreasing in δ, note that $\bar{\gamma}$ is now a function of $\bar{\lambda} \leq 2/k$, and $\frac{\partial\bar{\gamma}}{\partial\delta} = \frac{\partial\bar{\gamma}}{\partial\bar{\lambda}}\frac{\partial\bar{\lambda}}{\partial\delta}$. We know that $\frac{\partial\bar{\gamma}}{\partial\bar{\lambda}} < 0$ by substituting the value for $\mathsf{ask}^{\mathsf{Fast}}$ from IC_L into the speculator's information acquisition indifference condition:

$$\bar{\gamma} - \sigma\left(1 - \frac{\delta\bar{\lambda}k}{2}\right) = 0. \tag{72}$$

Hence, $\bar{\gamma}$ moves inversely to $\bar{\lambda}$. Then, because IC_L is decreasing in δ and $\bar{\lambda}$, it must be that if δ increases, $\bar{\lambda}$ must decline in δ for $\delta \in [\delta^*, 1]$. ■

Proof (Proposition 2). For the half spread at Exchanges Fast and Slow, $\mathsf{ask}^{\mathsf{Fast}}$ and $\mathsf{ask}^{\mathsf{Slow}}$, we prove the two cases, $\delta \in (0, \delta^*)$ and $\delta \in [\delta^*, 1]$, separately. Let $\delta = 0$. From the proof of Theorem 1, we know that $\mathsf{ask}^{\mathsf{Fast}} = \mathsf{ask}^{\mathsf{Slow}} = \frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^* + (1-\mu)(1-\underline{\lambda}^*)} > 0$. Now, consider $\mathsf{ask}^{\mathsf{Fast}}$. We know from the speculator's information acquisition indifference condition that $\mathsf{ask}^{\mathsf{Fast}}$ moves inversely to $\bar{\gamma}$, as $\bar{\gamma} = \left(\sigma - \mathsf{ask}^{\mathsf{Fast}}\right)$.

For $\delta \in (0, \delta^*)$, we know that $\bar{\gamma}$ is decreasing in δ from the proof of Proposition 1, which implies that $\mathsf{ask}^{\mathsf{Fast}}$ is increasing in δ. Now, let $\delta \in [\delta^*, 1]$. Recall from Proposition 1 that $\beta = 1$, and $\bar{\lambda}$ is decreasing in δ on $[\delta^*, 1]$. Then, because $\mathsf{ask}^{\mathsf{Fast}} |_{\delta=0} = \mathsf{ask}^{\mathsf{Fast}} |_{\delta=1}$, it must be that $\mathsf{ask}^{\mathsf{Fast}}(\delta) > \mathsf{ask}^{\mathsf{Fast}} |_{\delta=0}$ for all $\delta \in (0, 1)$. Now, consider $\mathsf{ask}^{\mathsf{Slow}}$. Let $\delta \in [\delta^*, 1]$. By the proof of Theorem 2, $\underline{\lambda} = 0$, and thus $\mathsf{ask}^{\mathsf{Slow}} = 0 < \mathsf{ask}^{\mathsf{Slow}} |_{\delta=0}$. For $\delta \in (0, \delta^*)$, $\frac{\partial \mathsf{ask}^{\mathsf{Slow}}}{\partial \delta} < 0$ follows from Proposition 1, as $\underline{\lambda}$ declines in δ. ■

Proof (Proposition 3). Total exchange volume is given by the expression:

$$\mathsf{Volume}_{t=1} = \mu\bar{\gamma} + (1-\mu) \times (1 - \underline{\lambda}). \tag{73}$$

For $\delta \in (\delta^*, 1)$, we know that $\underline{\lambda} = 0$, and thus $\frac{\partial \bar{\gamma}}{\partial \delta} > 0$ implies that $\mathsf{Volume}_{t=1}$ increases in δ on $[\delta^*, 1]$. Now let $\delta \in (0, \delta^*]$. Then, by (21), we have that:

$$\frac{k}{2} \times \underline{\lambda}\sigma - \mathsf{ask}^{\mathsf{Slow}} = 0. \tag{74}$$

Because in equilibrium $\bar{\lambda} = 2/k$, we can rewrite (74) as:

$$\underline{\lambda}(\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda}))\sigma = \bar{\lambda}\mu\bar{\gamma}(1-\beta) \iff \underline{\lambda}(1-\mu) = \mu\bar{\gamma}(1-\beta).$$

Using this fact, we can rewrite (73) as:

$$\mathsf{Volume}_{t=1} = \mu\bar{\gamma}\beta + (1-\mu), \tag{75}$$

which is necessarily increasing in δ, because $\bar{\gamma}\beta$ is increasing in δ. ■

Proof (Proposition 4). We begin with the subsidiary delayed exchange. Because the subsidiary exchange will set a delay δ, such that the sum of all volume across the delayed and non-delayed exchanges is maximized, we appeal to Proposition 3. Proposition 3 proves that total exchange volume is increasing for all δ, implying that the optimal delay is $\delta = 1$.

For an independently operated exchange, the delayed exchange maximizes its own volume, which is given by:

$$\mathsf{Volume}_{t=1}^{\mathsf{Slow}} = \mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda}). \tag{76}$$

From the proof of Proposition 3, we know that $\mu\bar{\gamma}(1-\beta) = (1-\mu)\underline{\lambda}$ for $\delta \in (0, \delta^*]$, which implies that $\mathsf{Volume}_{t=1}^{\mathsf{Slow}} = (1-\mu)\bar{\lambda}$, a constant. Then, for $\delta \in (\delta^*, 1]$, we know $\underline{\lambda} = 0$ and $\beta = 1$, implying that again that $\mathsf{Volume}_{t=1}^{\mathsf{Slow}} = (1-\mu)\bar{\lambda}$, which is decreasing in $\bar{\lambda}$ for all $\delta \in (\delta^*, 1]$. Thus, any $\delta \in (0, \delta^*]$ maximizes delayed exchange volume. ■

Proof. (Proposition 5). We show that $\delta^*(k, \mu, \sigma)$ is increasing in σ by differentiating δ^*, as δ^* is defined in terms of parameters:

$$\delta^*(k,\mu,\sigma) = \frac{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + (1-\mu)(1-\frac{2}{k})\mu\sigma} - (1-\mu)(1-\frac{2}{k})}{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + (1-\mu)(1-\frac{2}{k})\mu\sigma} + (1-\mu)(1-\frac{2}{k})}. \tag{77}$$

First, for ease of exposition, we denote the following term $x = \big((1-\mu)(1-\frac{2}{k})\big)$. Then, differentiating (77) by σ and simplifying:

$$\frac{\partial \delta^*}{\partial \sigma} = \frac{\mu x^2}{(\sqrt{x^2 + x\mu\sigma} + x)^2} > 0. \tag{78}$$

Thus, δ^* is increasing in σ. ■

Figure 2: Length of Latency Delay

Figure 2 illustrates how to interpret δ in the context of a fixed or random delay. Panels A and B depict a speculator with a distribution of possible reaction times, competing against a market maker with a known reaction time. In Panel A, if no latency delay is present (upper figure), the speculator is able to move before the market maker with certainty. With a fixed delay (lower figure), the distribution of the speculator's reaction time is slowed down by a fixed value, such that the speculator no longer moves before the market maker with certainty. In the presence of the delay, the speculator either moves before the market maker with probability $1 - \delta$, or is delayed until after the market maker with probability δ. Similarly, in Panel B, we depict a random delay. In this case, the distribution of speculator reaction times widens, instead of shifting.

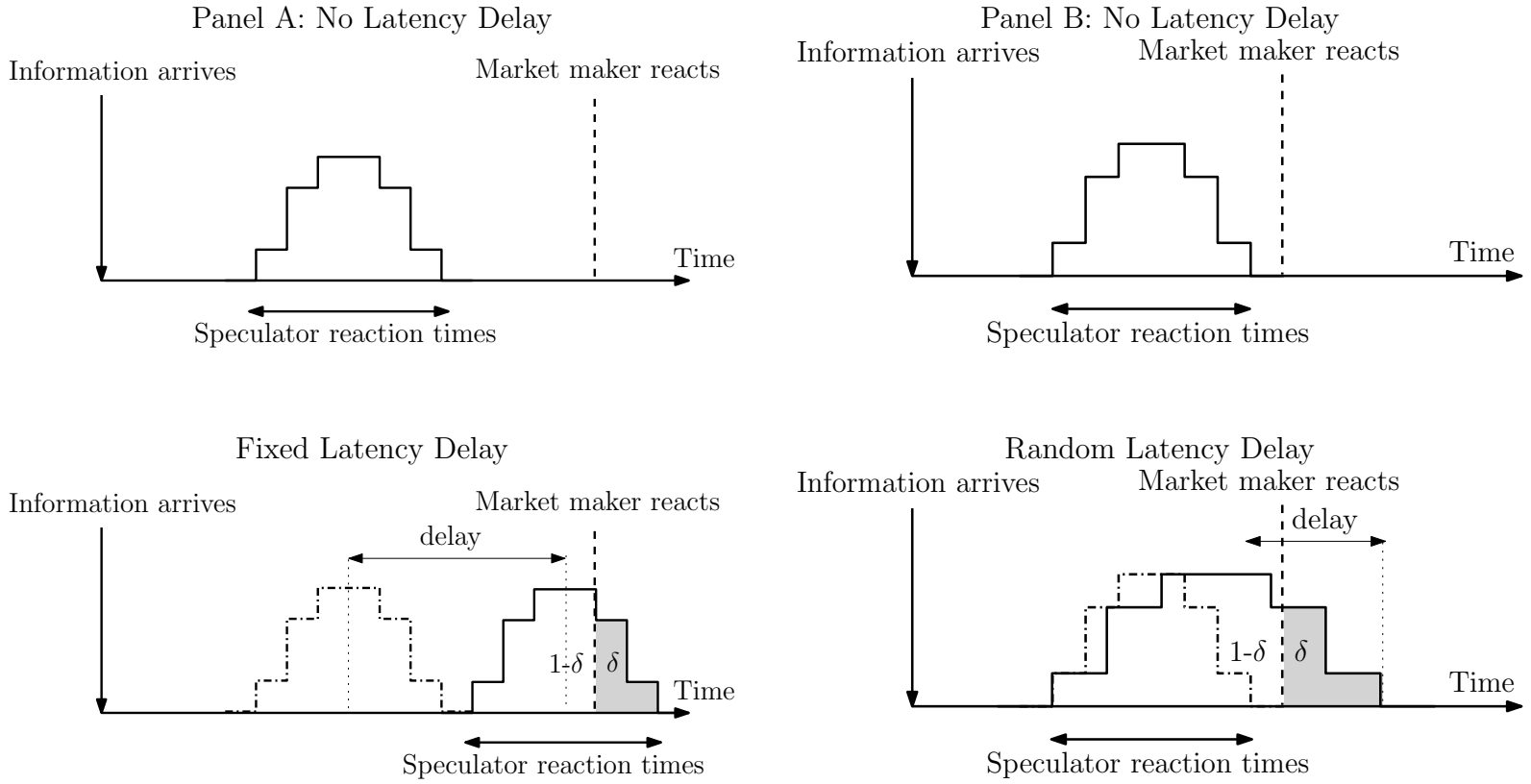


Figure 3: Market Participation by Investor Type

The left panel below depicts the unconditional probabilities of a speculator's action prior to $t = 2$, 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 $\mu = 0.5$. Results for other values of μ are qualitatively similar.

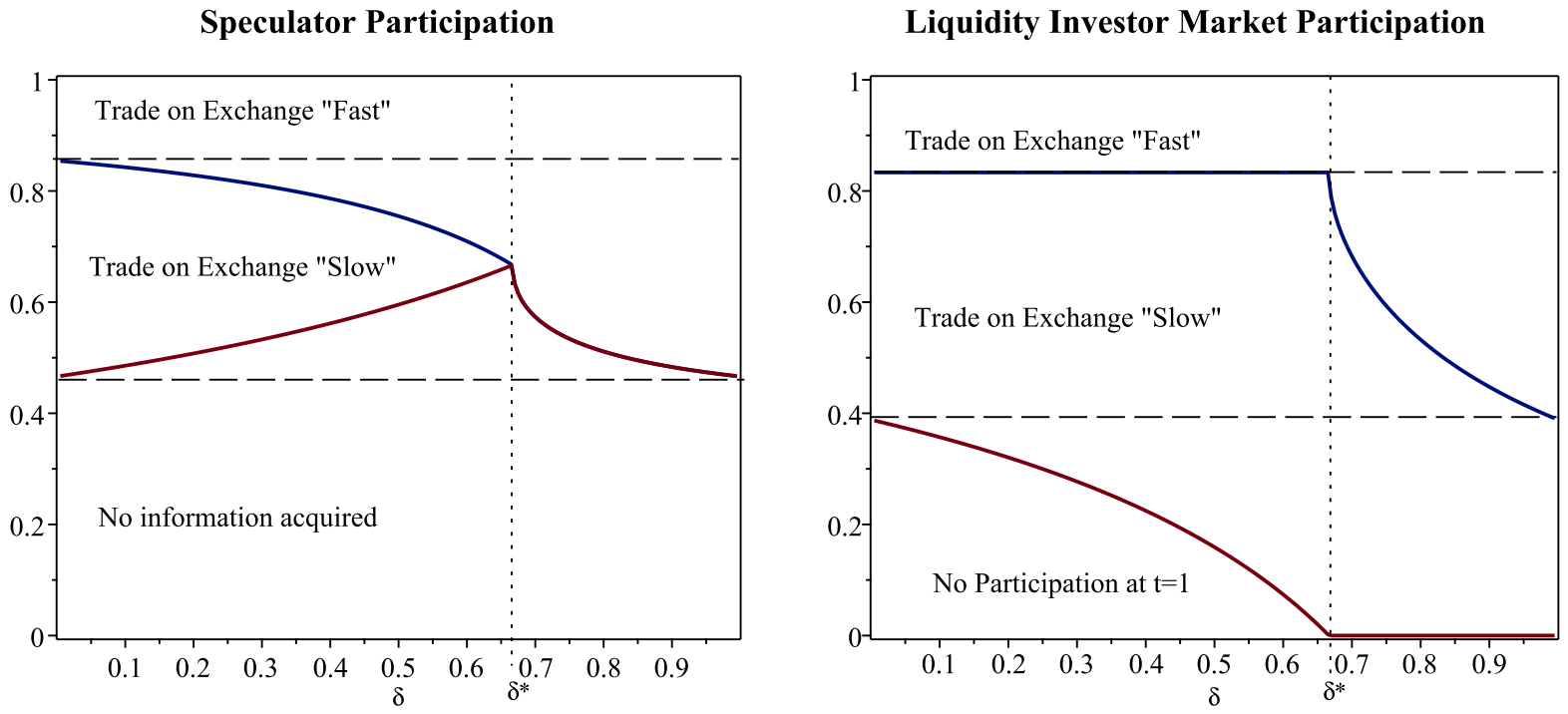


Figure 4: Order Submissions, Trades, and Market Participation

The left panel below depicts three volume figures: total exchange volume, delayed, and non-delayed volume, as a function of the Exchange Slow latency delay δ. The right panel illustrates order submission to exchanges 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 $\mu = 0.5$. Results for other values of μ are qualitatively similar.

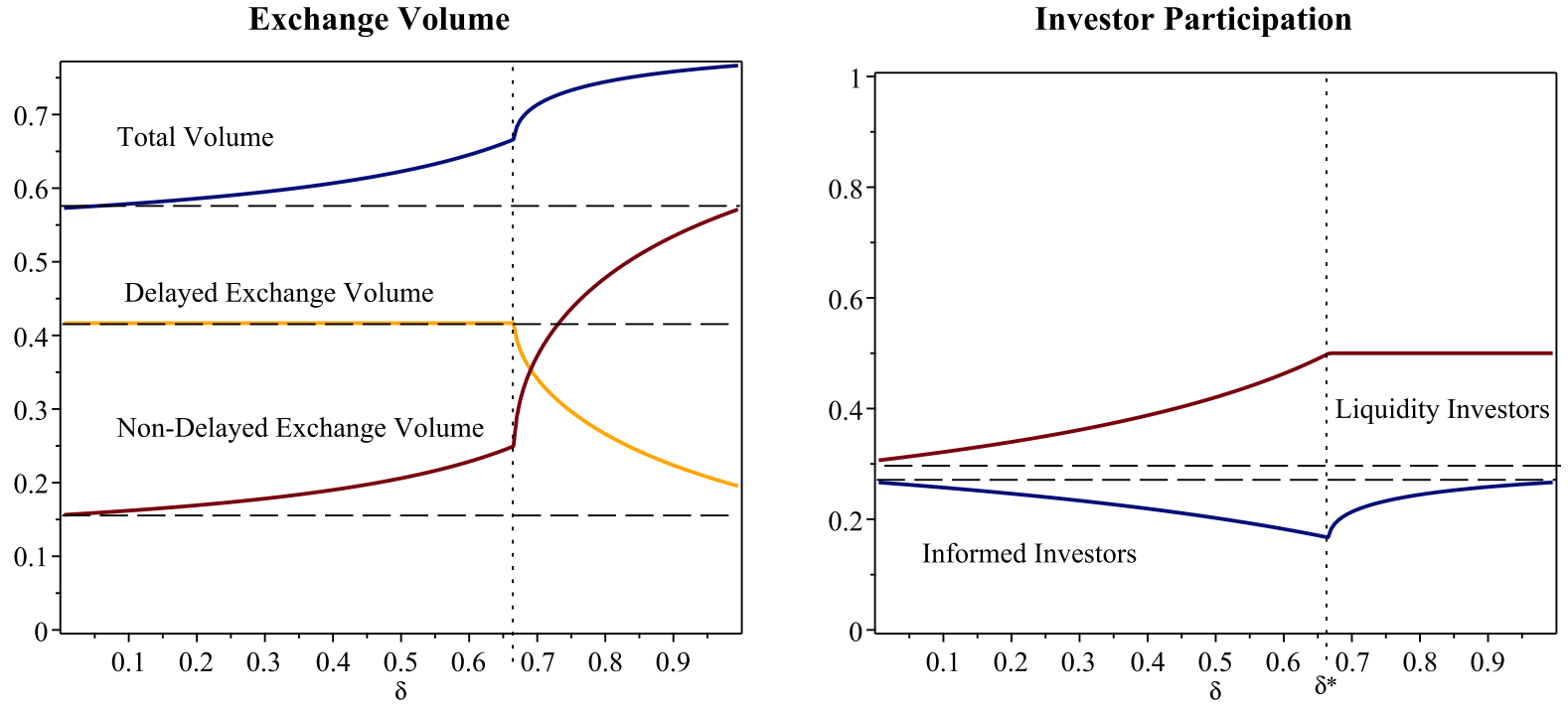


Figure 5: Quoted Spreads and Pricing Error

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. 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 $\mu = 0.5$. Results for other values of μ are qualitatively similar. The right panel depicts a measure of pricing error at $t = 1$. Our measure, root-mean-squared error, is an inverse measure of price discovery, and thus higher values indicate worse price discovery. The measure is centred about the benchmark level ($\delta = 0$), marked by the dotted black line. Line colours blue, green and red reflect parameter values $\mu = \{0.25, 0.5, 0.75\}$, respectively.

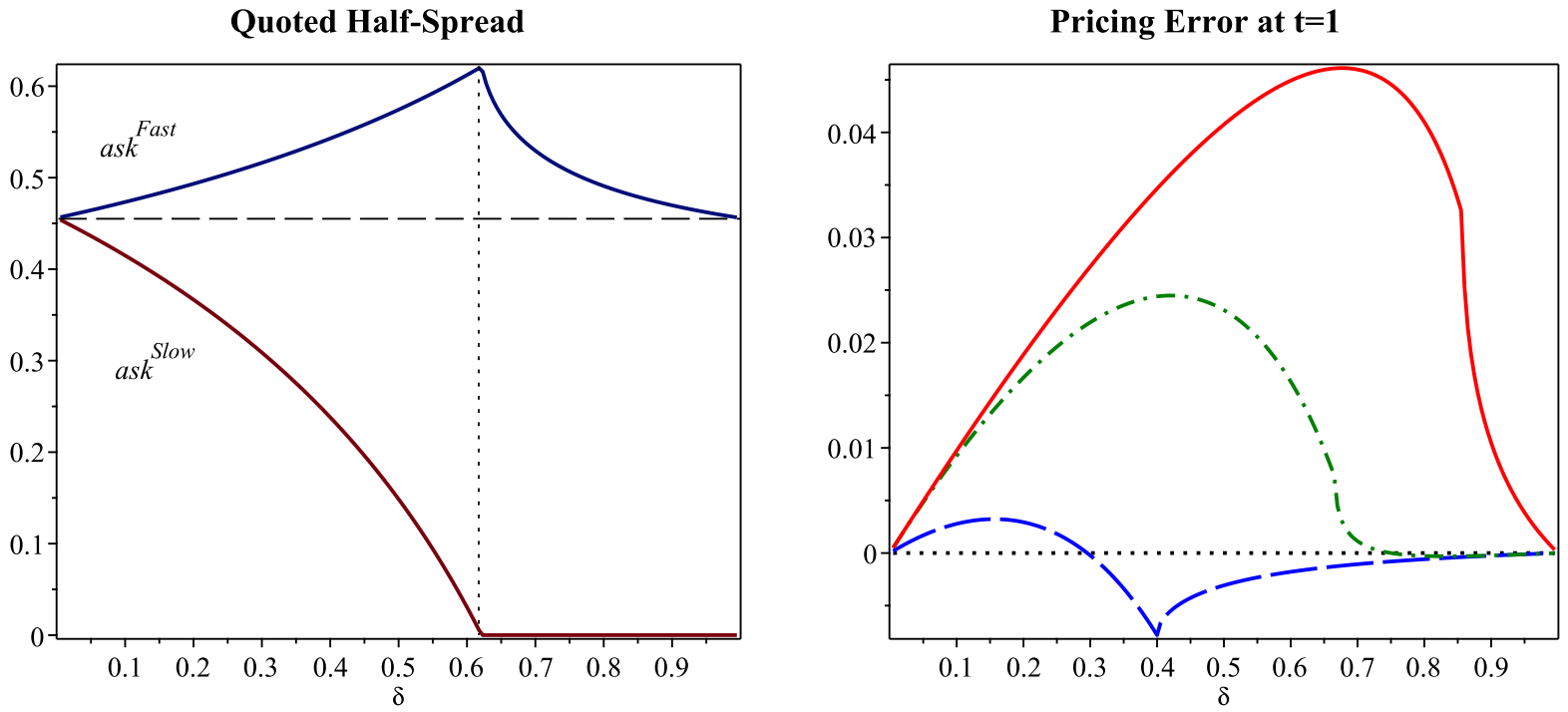


Figure 6: Expected Welfare

The left panel below illustrates the expected welfare realized by an investor who enters the market at $t = 0$. In the right panel, we present the expected information acquisition costs for informed investors (lower), and expected delay costs for liquidity investors (upper) separately. We present these costs as a function of the latency delay δ. The measure is centred about the benchmark level ($\delta = 0$), marked by the dotted black line. Line colours blue, green and red reflect parameter values $\mu = \{0.25, 0.5, 0.75\}$, respectively.