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

Michael Brolley and David A. Cimon *

Accepted at *Journal of Financial and Quantitative Analysis*

*Brolley, mbrolley@wlu.ca, Wilfrid Laurier University (corresponding author); and Cimon, dcimon@wlu.ca, Wilfrid Laurier University. We thank an anonymous referee, Hendrik Bessembinder (the editor), Eric Budish, Sabrina Buti, Sarah Draus, Sean Foley, Corey Garriott, Michael Goldstein, 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 RSM Liquidity Conference, the 2017 Sustainable Architecture for Finance in Europe Market Microstructure Conference, the 2017 European Finance Association Annual Meeting, the 2017 Northern Finance Association Annual Meeting, the 2017 Canadian Economics Association Annual Meeting, and seminars at the Bank of Canada, University of Ontario Institute of Technology, University of Toronto, and Wilfrid Laurier University, for valuable discussions and comments. We also thank William Wootton for valuable research assistance. Michael Brolley acknowledges financial support by the Social Sciences and Humanities Research Council of Canada Insight Development Grant program, grant no. 430-2016-00279. All errors are our own.

Abstract

Latency delays intentionally slow order execution at an exchange, often to protect market-makers against latency arbitrage. We study informed trading in a fragmented market in which one exchange introduces a latency delay on market orders. Liquidity improves at the delayed exchange, as informed investors emigrate to the conventional exchange, where liquidity worsens. In aggregate, implementing a latency delay worsens total expected welfare. We find that the impact on price discovery depends on the relative abundance of speculators. If the exchange with delay technology competes against a conventional exchange, it implements a delay only if it has sufficiently low market share.

I Introduction

Liquidity providers prefer to intermediate uninformed trades, as these trades are unlikely to move prices against them. In a competitive market for order flow, many exchanges have introduced market features to appeal directly to these traders. Exchanges try to attract uninformed traders from other markets with innovations such as inverse pricing, dark trading, and retail order segmentation facilities, advertising that these innovations discourage informed trading. Recently, some exchanges have imposed latency delays—so-called "speed bumps"—as yet another way to appeal to uninformed order flow. Measured on the order of milliseconds, and even microseconds, latency delays extend the time between an order's receipt at the exchange and its execution.[1] In this paper, we study the impact of introducing such a delay.

The impact of latency delays on liquidity remains controversial among industry stakeholders. Exchanges advertise latency delays as a means of protecting market-makers from adverse selection by high-frequency traders (HFTs) who "snipe" stale quotes before market-makers can update them; exchanges argue that these savings will ultimately pass on to investors through a narrower spread.[2] Opponents claim that delays create an uneven playing field by allowing market-makers to "fade" quotes ahead of orders, executing them at worse prices than those available at order submission.[3] In our model, these behaviors arise endogenously: a latency delay induces market-makers to quote a better spread, but orders may be executed at a worse price if the arrival of information induces a quote update before an order is filled.

We construct a three-period model of informed trading in a fragmented market. We interpret private information as a fleeting arbitrage opportunity that becomes publicly available to all market participants in the second period. Some traders, denoted as speculators, may pay a cost to acquire the private information in period one. Our interpretation of private information is similar to the

[1] See Appendix VII.B for detailed descriptions of latency delay mechanics as implemented in practice.

[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`

latency arbitrage phenomenon (e.g., between New York and Chicago) studied in Budish, Cramton, and Shim (2015). Traders arrive sequentially to trade one unit of a risky security by submitting an order to one of three venues: i) a standard exchange, which fills orders upon receipt in period one, ii) a delayed exchange that imposes a latency delay between order receipt and execution such that orders may fill in period two with some probability, or iii) an off-exchange internalizer that fills orders in period two.[4] Any order filled in period two is filled only after news of the arbitrage opportunity is made public and impounded into prices.[5] In our model, speculators are motivated by information rents, while liquidity investors seek to minimize trading costs. In addition to paying the (half-)spread, liquidity investors face heterogeneous delay costs similar to Zhu (2014).

Exchanges have implemented latency delays of both fixed (e.g., IEX) and random lengths (e.g., TMX Alpha). By modeling a latency delay in terms of the probability that an information advantage is lost, our model reflects both prominent delay implementations. In practice, the difference between the time when a speculator acquires information and submits an order, and that when a market-maker learns of the latency arbitrage and prices it into its quotes (i.e., the time at which the private information becomes public) is random. Our model is well-suited to analyzing delays of both fixed and random lengths: in both cases, the latency period simply adds a fixed or random delay to the already random length of time for which information remains private, and during which quotes are effectively "stale."

In equilibrium, the latency delay increases execution risk at the delayed exchange, segmenting informed order flow to the standard exchange. As a result, liquidity at the conventional exchange worsens via the widening of the bid-ask spread, while delayed exchange liquidity improves. As informed speculators concentrate at the standard exchange, competition for information rents in-

[4]We use the off-exchange internalizer as a stylized representation for trading venues with slower execution rates (relative to the nature of latency arbitrage). The internalizer in our model functions similarly to a conventional midpoint crossing network (e.g., Sigma X's Reference Price Book), or a delayed midpoint crossing network (e.g., IntelligentCross, which sets minimum resting times and time-delayed execution).

[5]In practice, exchanges (e.g., IEX) that impose a delay between receipt and execution may also delay information transmission on completed trades. We focus on delay to incoming orders, both because it is universal among delayed exchanges and because it offers a lower bound for the segmenting effects of a delayed exchange. We posit that an exchange that also delays information transmission would only be more dissuasive to informed speculators relying on inter-market arbitrage, as they would require additional time to confirm order execution.

tensifies, reducing overall information acquisition by speculators. However, the reduction in adverse selection at the delayed exchange attracts the most relatively latency-sensitive uninformed order flow from the internalizer, leading to an increase in on-exchange volume. The net effect is an increase in total exchange-traded volume.

Given the mixed effects on liquidity across markets, we proceed to examine the overall effect of implementing a latency delay on the welfare of all market participants. We use a measure that reflects allocative efficiency, by taking expectation over the gains from trade that arise from the buyer and seller profit functions. We show that welfare simplifies to two costs: per-trade expected delay costs paid by liquidity investors, and information acquisition costs paid by speculators. We find that speculators commit fewer average resources to information acquisition in the presence of a delayed exchange, but that liquidity investors incur higher average delay costs. The increased delay costs outweigh the diminished information acquisition costs, such that the introduction of a delayed exchange worsens investor welfare.

Including private information and endogenous information acquisition in our model affords us the opportunity to study the contribution of a latency delay to price discovery. Using a root mean-squared proportional pricing error, we find that the ratio of speculators relative to liquidity investors plays an important role: the introduction of a delayed exchange worsens price discovery for securities with a relatively large speculator presence. The increased competition for information rents among relatively many informed speculators increases the expected price impact of informed speculators, but this effect is dominated by the countervailing reduction in information acquisition, ultimately worsening price discovery. If the relative number of speculators is low, the introduction of a short delay may improve price discovery, as the relatively many liquidity investors may sufficiently absorb the segmentation of informed order flow to the non-delayed exchange such that the increase in expected price impact dominates.

Since price discovery may improve or worsen depending on the length of delay implemented by the exchange, we examine the decision of an exchange to implement a delay. We assume that the exchange maximizes profits through volume. In our investigation, we consider two market

organization environments: i) a stand-alone delayed exchange that competes with a standard exchange (e.g., IEX) and ii) a delayed exchange that is a subsidiary of a standard exchange (e.g., TSX Alpha, NYSE American). We predict that a stand-alone exchange whose market share is sufficiently lower than that of a competing standard exchange will implement a delay. Such a stand-alone exchange selects a delay that balances investors' preference for liquidity and latency sensitivity to draw order flow from both the standard exchange and the off-exchange internalizer; as a result, informed speculators may not fully segment from the delayed exchange. A subsidiary exchange instead maximizes total exchange-traded volume across both exchanges. To do so, it is optimal to impose a maximal delay so that the delayed exchange effectively assumes the role of an off-exchange internalizer, attracting latency-insensitive liquidity investors to trade on-exchange. Here, the standard and delayed exchange jointly capture all volume in the market.

Related Literature. To our knowledge, our paper is the first to study latency delays as an order flow segmentation mechanism in a fragmented market. Existing models of latency delays focus on single-venue markets (e.g., Budish, Cramton, and Shim (2015), Rojcek and Ziegler (2016), Aldrich and Friedman (2019), and Aoyagi (2019)) or fragmented markets with identical delays. Closest to our work, Baldauf and Mollner (2018) analyze a fragmented market whose exchanges all impose an identical latency delay from which only limit order cancellations are exempt. As in our predictions on liquidity, the authors find that the quoted spread narrows when an exchange introduces a delay, an effect driven by the reduction in information acquisition. Complementary to Baldauf and Mollner (2018), we study the interaction between exchanges with delays and those without, providing a discussion on the migration of traders between exchanges; moreover, we examine a delayed exchange's optimal choice of delay magnitude.

Our work contributes to the broad literature on market segmentation, which includes (but is not limited to) studies on dark pools (e.g., Zhu (2014), and Menkveld, Yueshen, and Zhu (2017)), access fees (e.g., Colliard and Foucault (2012), and Malinova and Park (2015)), and broker order routing decisions (e.g., Battalio, Corwin, and Jennings (2016), and Cimon (2019)). Empirical work shows that fragmented markets improve liquidity (Foucault and Menkveld (2008)) and efficiency

(O'Hara and Ye (2011)). Market segmentation may also lead to "cream skimming", as alternative exchanges are able to divert desirable orders from primary exchanges. Empirical evidence on the topic is mixed (see e.g., Battalio (1997), and Hatheway, Kwan, and Zheng (2017)). In our paper, we predict that latency delays do play an order segmentation role, and may also allow for cream skimming, as delayed exchanges are able to capture a larger fraction of uninformed orders.

Our paper also relates to the literature on HFT, as we provide new predictions toward the impact that latency delays may have on the relationship between high-frequency (HF) arbitrageurs and HF liquidity providers, and consequently, their impact on liquidity and price discovery. Empirical evidence suggests that HF liquidity providers may improve liquidity (see e.g., Brogaard, Hagströmer, Nordén, and Riordan (2015), Subrahmanyam and Zheng (2015), and Brogaard and Garriott (2018)), whereas HF liquidity demanders may increase transaction costs (Chakrabarty, Jain, Shkilko, and Sokolov (2014), **?**). Further evidence suggests that HFTs may improve price discovery through both liquidity supply (Conrad, Wahal, and Xiang (2015), and Brogaard, Hendershott, and Riordan (2019)) and demand (Brogaard, Hendershott, and Riordan (2014)). Carrion (2013) finds that liquidity demanders may also improve market efficiency.

Theoretical studies on HFTs have examined their roles in modern markets, including: market making (Jovanovic and Menkveld (2015)), arbitrage (Wah and Wellman (2013)), and the incorporation of new information (Biais, Foucault, and Moinas (2015)).[6] Close to our paper, Menkveld and Zoican (2017) and Pagnotta and Philippon (2018) model the effects of exchange speed. Menkveld and Zoican (2017) focus on the processing latency within an exchange, versus latency in reaching the exchange. Pagnotta and Philippon (2018) focus on competition in exchange speed investment. We complement this body of work by examining a fragmented market where delays are asymmetric across exchanges and order types.

Our paper addresses the argument of latency delay proponents that delays may curb the "predatory" HFT practice of cross-market latency arbitrage. Critics have suggested that latency delays may also lead to quote fading. Existing evidence is mixed, as Malinova and Park (2016) document

[6]See Angel, Harris, and Spatt (2011), O'Hara (2015), and Menkveld (2016) for extensive surveys on work related to HFT.

evidence of predatory quote fading behavior by HFTs, while Latza, Marsh, and Payne (2014) find no evidence.[7] As these behaviours relate to latency delays, the evidence is also mixed. In studies of the latency delay implementation by Canadian exchange TMX Alpha, Chen, Foley, Goldstein, and Ruf (2017) find that liquidity demanders are able to access a lower proportion of posted liquidity following the introduction of a delay, whereas Anderson, Andrews, Devani, Mueller, and Walton (2018) find that market-wide liquidity does not deteriorate. In our paper, we abstract from arbitrary quote fading by assuming that market-makers update their quotes as a rational response to new information.

Finally, we acknowledge that our model abstracts from the phenomenon of queue-jumping, whereby liquidity providers circumvent time priority at one market by posting a limit order to another market at an economically insignificant price improvement. Because our model assumes competitive liquidity provision by the market-maker and limits the order placement strategy of speculators and liquidity investors to market orders only, there is no incentive for the market-maker to improve upon quotes that yield zero expected profits in equilibrium. For a theoretical analysis of queue-jumping in a fragmented market, we refer the reader to Buti, Consonni, Rindi, Wen, and Werner (2015), who study an environment in which a public limit order book with a positive tick size operates alongside a venue that permits price improvement on a finer grid than the limit order book.

II 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 $\sigma \in (0, 1]$. v is unknown by the public at $t = 1$, but is publicly announced at the beginning of $t = 2$. The asset is liquidated at $t = 3$.

Market Organization. There are two exchanges, Fast and Slow, and an off-exchange internalizer. Exchanges Fast and Slow operate limit order books, where posted limit orders are visible

[7] In related work, Gai, Yao, and Ye (2013) find evidence that high-frequency traders engage in the strategy known as "quote stuffing", which we do not address in this paper.

to all market participants. Market orders sent to Exchange Fast (the "standard exchange") at $t = 1$ fill immediately upon receipt. Market orders sent to Exchange Slow (the "delayed exchange") are subject to a random delay. With probability $\delta \in (0, 1)$ an order sent to Exchange Slow at $t = 1$ is delayed, to fill only after the public announcement of v at $t = 2$. Otherwise, the order is filled immediately at $t = 1$. Limit orders submitted to Exchange Slow are not subject to the delay.[8] The internalizer fills market orders that it receives with liquidity provided by a market-maker. These orders fill after the public announcement of v.

We define the latency delay at Exchange Slow in probabilistic terms to reflect the impact of the delay relative to the fleeting nature of an arbitrage opportunity. In practice, an exchange can impose a latency delay whose length is deterministic or random. Because traders competing for arbitrage opportunities across markets have a distribution of reaction times (e.g., different hardware, algorithms and other software, etc.,), events may reveal mispricing to liquidity providers before a trader's market order clears the delay and fills at the intended quote. As these reaction times are uncertain, the combined effect of latency differences between trader actions and an exchange-imposed latency delay leads a trader to interpret the event of a price change before an order executes as probabilistic. We illustrate the translation of a deterministic exchange latency delay into an interpretation in δ in Panel A of Figure 1. In this case (e.g., IEX), a deterministic delay slows the distribution of reaction times by a fixed amount, increasing the probability that liquidity providers move first. A random latency delay instead slows some orders for longer periods than others. Panel B of Figure 1 illustrates this interpretation, representing the delays currently implemented by Canadian venues TMX Alpha and Aequitas Neo.

We present the following simplified example to illustrate the impact of a latency delay on an arbitrageur's decision to submit an order to a delayed exchange. Consider an arbitrage opportunity that is exploitable by some market participants with an uncertain window of 5-10 milliseconds. Here, a latency delay of less than 5 milliseconds is unlikely to deter arbitrageurs from picking off

[8]Exchanges with latency delays have, generally, exempted liquidity providers from the latency delay. For example, IEX will update pegged orders in response to external factors (e.g., orders pegged to the midpoint, National Best Bid-Offer (NBBO), or their 'discretionary peg'). TMX Alpha requires that liquidity-providing orders meet a minimum size requirement bypass the delay. In general, it is insufficient to merely submit a limit order to bypass the delay.

stale quotes ($\delta = 0$), while a delay of greater than 10 milliseconds will negate the opportunity entirely ($\delta = 1$). A delay between 5-10 milliseconds will reduce the pick-off risk at the delayed exchange, but not eliminate it ($\delta \in (0, 1)$). The internalizer in our model effectively operates as a delayed venue that employs the most extreme delay, such that arbitrage is not possible via the internalizer.[9]

Market-Maker. A single risk-neutral market-maker supplies liquidity to all venues: Fast, Slow, and the internalizer. At each venue, the market-maker prices its limit orders competitively such that it earns zero expected profits at *each* exchange (i.e., in the manner of Glosten and Milgrom (1985)).[10] 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. The market-maker receives only the public signal v_0 at the beginning of $t = 0$. Upon the announcement of v at $t = 2$, the market-maker updates its $t = 1$ limit orders to the public value such that $\text{ask}_2^{\text{Fast}} = \text{ask}_2^{\text{Slow}} = \text{bid}_2^{\text{Fast}} = \text{bid}_2^{\text{Slow}} = v$. This update happens before orders that have been delayed at Exchange Slow are able to reach the exchange. Moreover, because all trades at the internalizer occur at $t = 2$ following the announcement of v, the market-maker prices all limit orders sent to the internalizer at v.

Market-Maker Payoff. As the market-maker sets prices at both exchanges such that its expected payoff from a buy or sell order is zero, its profit function is written as follows. We denote the market-maker with the short-hand 'MM'.

$$\pi_{MM}^{\text{Fast}}(\text{ask}_1^{\text{Fast}};\ \text{Buy at Fast at t=1}) = \text{ask}_1^{\text{Fast}} - \text{E}[v \mid \ \text{Buy at Fast at t=1}] = 0, \tag{1}$$

$$\pi_{MM}^{\text{Slow}}(\text{ask}_1^{\text{Slow}};\ \text{Buy at Slow at t=1}) = \text{ask}_1^{\text{Slow}} - \text{E}[v \mid \ \text{Buy at Slow at t=1}] = 0. \tag{2}$$

Finally, as the market-maker prices all limit orders at the (public) true value v in $t = 2$, the

[9] Menkveld, Yueshen, and Zhu (2017) find evidence that midpoint-crossing dark pools offer the lowest immediacy, when compared to other displayed and non-midpoint off-exchange venues.

[10] We assume that a single market-maker prices competitively within-exchange to abstract from the Bertrand competition liquidity provision game that arises from multiple market-makers competing at each exchange. We assume that the market-maker earns zero profit at each exchange to eliminate the case in which the market-maker earns zero expected profits through a loss at one exchange and positive profits at the other, a case that would not arise in an environment where multiple market-makers compete.

market-maker earns zero profit on all trades at the internalizer.

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 a liquidity investor. Speculators are endowed with an information acquisition cost $\gamma_i \sim \text{U}[0, 1]$ which they can pay upon arrival at $t = 0$ to perfectly learn the random payoff v. We refer to those who acquire information as "informed speculators," 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. Speculators who do not acquire information are classified as "uninformed speculators."

With probability $(1 - \mu)$, a liquidity investor arrives and is a buyer or seller with equal probability. Liquidity investors have no private information about v and cannot acquire it, but they are endowed with liquidity needs that motivate them to trade. Liquidity investors also face a 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 $\lambda_i \sim U[0, 1]$ models an investor's private sensitivity to delay. Introducing a universal scaling parameter k allows us to normalize the distribution of the latency sensitivity parameter λ_i to the unit interval, which aids in the interpretation of our results. The liquidity investor delay cost is similar to that of Zhu (2014), which may represent a number of unmodelled factors, such as risk aversion or recapitalization costs. Both examples represent costs that liquidity investors face when the price moves away from them, but not if it moves in their favour.[11] Alternatively, a liquidity investor can elect not to trade, and incur a non-participation cost $K \in (\sigma, \infty)$. We assume that the cost of not trading is high enough that it induces liquidity investors to trade (i.e., $K > \max\{\frac{c_i}{2}\}$).

An investor sends a market order to the venue that will maximize expected profits. Speculators do so by capitalizing on informational advantage, whereas liquidity investors maximize profits by minimizing their total trading costs. An investor i may submit a single market order at $t = 1$ or not

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

trade.[12] An investor who submits a market order at $t = 1$ may select either one of the exchanges, or may send the order to the off-exchange internalizer. We assume that once an investor's order fills, any information acquired by the investor becomes public immediately, before any other trades occur. Finally, the structure of the model is known to all market participants. We illustrate the timing of the model in Figure 2.

Investor Payoffs. The expected payoff to an investor who submits a buy order at $t = 1$ is given by their knowledge of the true value of v minus the price paid and any information acquisition or delay costs incurred. Because the market-maker sets bid and ask prices competitively, conditioning on public information and the expected adverse selection of an incoming investor, any market order filled at $t = 1$ thus pays a premium above the public value known at $t = 1$. Hence, speculators who do not acquire information will not trade, as uninformed speculators know only the public value. Consequently, only two types of traders send orders at $t = 1$: informed speculators, and liquidity investors.

We denote liquidity investors as L, informed speculators as I, and uninformed speculators as U. The expected payoffs to a liquidity investor L from submitting a buy order to either Exchange $j \in \{\text{Fast}, \text{Slow}\}$, or the internalizer (Int) are given by:

$$\pi_L^{\text{Fast}}(\lambda_i;\ \text{Buy at t=1}) = v_0 - \text{ask}_1^{\text{Fast}}, \tag{3}$$

$$\pi_L^{\text{Slow}}(\lambda_i;\ \text{Buy at t=1}) = (1-\delta) \times (v_0 - \text{ask}_1^{\text{Slow}}) + \delta \times \left(v - v - \frac{k\lambda_i\sigma}{2} \right), \tag{4}$$

$$\pi_L^{\text{Int}}(c_i;\ \text{Buy at t=1}) = (v - v) - \frac{k\lambda_i\sigma}{2}. \tag{5}$$

[12] We make the single-order assumption to simplify the model, but its impact on our qualitative results is motivated by unmodelled trading costs (e.g., message costs) that, we argue, would prevent investors from regularly sending orders from which they expect only a very small probability of a non-zero payoff. We provide a detailed explanation in section VII.C of the Appendix.

Similarly, the payoffs to an informed speculator I are given by:

$$\pi_I^{\text{Fast}}(\gamma_i;\ \text{Buy at t=1}) = v - \text{ask}_1^{\text{Fast}} - \gamma_i, \tag{6}$$

$$\pi_I^{\text{Slow}}(\gamma_i;\ \text{Buy at t=1}) = (1-\delta) \times (v - \text{ask}_1^{\text{Slow}}) + \delta \times (v - v) - \gamma_i, \tag{7}$$

$$\pi_I^{\text{Int}}(\gamma_i;\ \text{Buy at t=1}) = (v - v) - \gamma_i = -\gamma_i < 0. \tag{8}$$

Finally, the payoff to an uninformed speculator, who does not trade, is given by:

$$\pi_U(\gamma_i; \text{no trade}) = 0. \tag{9}$$

Seller payoffs are similarly defined. 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 speculator who submits an order to the internalizer (or does not trade) recovers no value from the information, and pays acquisition cost $-\gamma_i$. A liquidity investor who submits a buy order to the internalizer purchases the asset at its true value v and pays the delay cost with probability $1/2$.

III Equilibrium

We begin by solving the model with two non-delayed exchanges to establish a benchmark against which to compare the setting in which Exchange Slow imposes a delay $\delta > 0$. We define our benchmark in this way—rather than as a single competitive exchange—to disentangle the impact of fragmentation. We maintain the exchange labeling convention Fast and Slow throughout the paper for consistency, acknowledging that in the benchmark case, both exchanges are identical.

In both the benchmark and delayed exchange settings, we search for a weak Perfect Bayesian equilibrium in which the market-maker chooses a quoting strategy that yields zero expected profits at each venue, and investors choose order submission strategies that maximize their profits. We also focus on equilibria where both exchanges receive positive order flow.[13] Because the set-up of

[13] While we allow for separating equilibria in which informed speculators and liquidity investors cluster at different

our model is symmetric for buyers and sellers, we focus our exposition on the decisions of buyers without loss of generality.

III.A Identical Fragmented Markets (No Latency Delay)

If no exchange imposes a processing delay ($\delta = 0$), investors' payoffs simplify considerably: all orders submitted to an exchange immediately fill at the posted quote. Informed speculator and liquidity investor payoffs to trading on Exchange j reduce to:

$$\pi_I^j(\gamma_i; \text{Buy at t=1}) = v - \text{ask}_1^j - \gamma_i, \tag{10}$$

$$\pi_L^j(\lambda_i; \text{Buy at t=1}) = v_0 - \text{ask}_1^j. \tag{11}$$

Because a market order fills immediately at the posted quote, a liquidity investor's latency sensitivity λ_i does not enter into their on-exchange payoff; instead, λ_i enters through the venue choice decision, which weighs immediate execution at an exchange against delayed execution at the internalizer.

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, pricing the expected adverse selection of an incoming buy (sell) order into the ask (bid) price at $t = 1$ on each exchange. We denote the ask prices at Exchanges Fast and Slow at $t = 1$ as $\text{ask}_1^{\text{Fast}}$ and $\text{ask}_1^{\text{Slow}}$, respectively, and write them below:

$$\text{ask}_1^{\text{Fast}} = \text{E}[v \mid \text{Buy at Exchange Fast}], \tag{12}$$

$$\text{ask}_1^{\text{Slow}} = \text{E}[v \mid \text{Buy at Exchange Slow}]. \tag{13}$$

Prices $\text{bid}_1^{\text{Fast}}$ and $\text{bid}_1^{\text{Slow}}$ are analogously determined using the symmetry of buyers and sellers. Upon the announcement of v at $t = 2$, the market-maker updates its buy orders on both exchanges

exchanges, the no-trade theorem of Milgrom and Stokey (1982) demonstrates why such an equilibrium cannot exist: a separating equilibrium of this type would perfectly reveal an informed speculator's private information to the market-maker, incentivizing the informed speculator to deviate to the exchange at which the liquidity investors participate, to hide amongst the uninformed order flow.

to $\text{ask}_2^{\text{Fast}} = \text{ask}_2^{\text{Slow}} = \text{bid}_2^{\text{Fast}} = \text{bid}_2^{\text{Slow}} = v$.

Each investor makes two decisions: i) whether to participate in the market at $t = 1$, and if so, ii) to which venue to submit an order. The participation decision for speculators depends on the profitability of private information relative to their private information acquisition cost γ_i. A participating speculator decides on a venue choice strategy to maximize profit. Similarly, liquidity investors receive their delay cost $c(\lambda_i)$ at $t = 0$ and weigh it against the cost of not trading K. A liquidity investor electing to trade chooses to which venue to send an order.

We characterize these decisions via backward induction. First, we characterize the profit from trading at the internalizer, which delays orders with certainty. At $t = 2$, a speculator (informed and otherwise) whose order is delayed has no information advantage, and thus their expected profit is zero. A liquidity investor who submits an order to the internalizer in $t = 1$ pays an average delay cost of $\frac{c_i}{2} = \frac{k\lambda_i\sigma}{2}$. Hence, it is always optimal for a liquidity investor to submit an order at $t = 1$, as the cost to abstaining $K > \max\{\frac{k\lambda_i\sigma}{2}\}$.

At $t = 1$, speculators who do not acquire information at $t = 0$ do not trade. A speculator who acquires knowledge of v, now knows that delaying an order until period $t = 2$ (i.e., via the internalizer) is unprofitable, so the informed speculator chooses an order submission strategy over Exchanges Fast and Slow. We denote the probability with which an informed speculator 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 at an exchange 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 exchanges at $t = 1$ is characterized by:

$$\text{(14)} \qquad \text{Informed Buyer: } \left\{\beta \mid \pi_I^{\text{Fast}}(\text{Buy t=1}) = \pi_I^{\text{Slow}}(\text{Buy t=1}) \iff \text{ask}_1^{\text{Fast}} = \text{ask}_1^{\text{Slow}}\right\},$$

$$\text{(15)} \qquad \text{Liquidity Buyer: } \left\{\alpha \mid \pi_L^{\text{Fast}}(\text{Buy t=1}) = \pi_L^{\text{Slow}}(\text{Buy t=1}) \iff \text{ask}_1^{\text{Fast}} = \text{ask}_1^{\text{Slow}}\right\}.$$

We note here that because both exchanges are identical, γ_i and λ_i do not directly impact an investor's venue choice; instead, an investor chooses a venue based on the available quotes. Hence, if quotes are not equal across the exchanges at $t = 1$, then investors pool at the best-priced ex-

change, which cannot be an equilibrium, as either: i) the high-priced exchange would have no volume, violating the equilibrium assumption of positive order flow at both exchanges, or ii) the high-priced exchange would improve its prices to attract order flow.

Given α and β, and prices $\text{ask}_1^{\text{Fast}}$ and $\text{ask}_1^{\text{Slow}}$ such that $\text{ask}_1^{\text{Fast}} = \text{ask}_1^{\text{Slow}}$, speculators and liquidity investors make participation decisions at $t = 0$ that determine: i) the profitability threshold for information acquisition, denoted $\bar{\gamma}$, below which they acquire information, and ii) the latency sensitivity threshold $\underline{\lambda}$ above which liquidity investors trade on exchange (rather than at the internalizer). To find $\bar{\gamma}$, we find the highest value of γ_i at which a speculator earns a non-negative expected profit from becoming informed:

$$\bar{\gamma} = \max\left\{v - \text{ask}_1^{\text{Fast}}, v - \text{ask}_1^{\text{Slow}}\right\} = v - \min\left\{\text{ask}_1^{\text{Fast}}, \text{ask}_1^{\text{Slow}}\right\} = v - \text{ask}_1^{\text{Fast}}. \tag{16}$$

Hence, any speculator with $\gamma_i \leq \bar{\gamma}$ will acquire information, and the mass of informed speculators at $t = 1$ is equal to: $\mu_I = \mu \text{Pr}\left(\gamma_i \leq \bar{\gamma}\right)$. Similarly, we characterize the latency sensitivity threshold above which a liquidity investor will trade on-exchange $\underline{\lambda}$ by comparing the total cost of trading via exchange to the cost of trading at the internalizer:

$$\min\left\{\pi_L^{\text{Fast}}(\underline{\lambda}), \pi_L^{\text{Slow}}(\underline{\lambda})\right\} = \pi_L^{\text{Int}}(\underline{\lambda}) \iff \underline{\lambda} = \frac{2}{k\sigma}\min\left\{\text{ask}_1^{\text{Fast}}, \text{ask}_1^{\text{Slow}}\right\} = \frac{2}{k\sigma}\text{ask}_1^{\text{Fast}}. \tag{17}$$

Therefore, a liquidity investor with a delay cost $\lambda_i \geq \underline{\lambda}$ chooses to trade at an exchange in $t = 1$. The probability that such a liquidity investor arrives is $(1 - \mu)\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right)$.

Given the participation and venue choice strategies of informed speculators and liquidity investors, we use Bayes' Rule to characterize the period 1 ask prices quoted by the market-maker at

Exchanges Fast and Slow (sell prices $\text{bid}_1^{\text{Fast}}$ and $\text{bid}_1^{\text{Slow}}$ are symmetric about v_0).

(18)
$$\text{ask}_1^{\text{Fast}} = v_0 + \frac{\Pr(\text{informed trade at Fast})}{\Pr(\text{trade at Fast})} \times \sigma = v_0 + \frac{\mu\bar{\gamma}\beta\sigma}{\mu\bar{\gamma}\beta + (1-\mu)\alpha\Pr(\lambda_i \geq \underline{\lambda})},$$

(19)
$$\text{ask}_1^{\text{Slow}} = v_0 + \frac{\Pr(\text{informed trade at Slow})}{\Pr(\text{trade at Slow})} \times \sigma = v_0 + \frac{\mu\bar{\gamma}(1-\beta)\sigma}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(1-\alpha)\Pr(\lambda_i \geq \underline{\lambda})}.$$

Finally, the existence of an equilibrium requires that liquidity investors with maximum latency sensitivity (i.e., $\lambda_i = 1$) have sufficiently high expected delay costs $c = k\sigma/2$ such that they strictly prefer immediate execution. This assumption assures that both exchanges have non-zero liquidity investor participation, as informed speculators will not trade at venues where no liquidity investors participate (see e.g., Milgrom and Stokey (1982)). We denote this value as $\underline{k}$. Formally, we assume the following.

Assumption 1 (Liquidity Investor Immediacy) $k > \underline{k} = \frac{4((1-\mu)+2\mu\sigma)}{(1-\mu)+4\mu\sigma}$.

For the identical exchange case, Assumption 1 is stronger than necessary (we require only that $k > 2$), as it reflects the level of k required to form an equilibrium in a market where one exchange imposes a delay. We have introduced it here for ease of exposition, to maintain consistency with Subsection III.B.

In summary, an equilibrium in our model is characterized by: (i) investor participation values, $\bar{\gamma}$ and $\underline{\lambda}$; (ii) investor venue strategies, α and β; and (iii) market-maker quotes at $t = 1$ for each exchange $j \in \{\text{Fast}, \text{Slow}\}$, ask_1^j and bid_1^j. These values solve the investor venue choice indifference conditions (14)-(15), participation conditions (16)-(17), and the market-maker quoting strategy (18)-(19). In the existence theorem to follow, we denote the equilibrium values of the benchmark case with the subscript 'B'.

Theorem 1 (Identical Fragmented Markets) *Let $\delta = 0$ and k satisfy Assumption 1. Then for any $\beta_B \in (0, 1)$, there exists a unique equilibrium consisting of participation constraints $\bar{\gamma}_B \in (0, 1)$,*

$\underline{\lambda}_B \in [0,1]$ *that solve (16)-(17), prices* $ask_{1,B}^{Fast}, ask_{1,B}^{Slow}$, $bid_{1,B}^{Fast}$ *and* $bid_{1,B}^{Slow}$ *that satisfy (18)-(19), and* $\alpha_B \in (0,1)$ *that solves (14)-(15) such that* $\alpha_B = \beta_B$.

Theorem 1 illustrates that, in equilibrium, identical fragmented markets may co-exist, and that they need not attract the same level of order flow despite offering identical prices. For example, in an equilibrium where $\alpha_B = \beta_B = 3/4$, α_B and β_B cancel out of the pricing equations (18)-(19) to yield $\text{ask}_{1,B}^{\text{Fast}} = \text{ask}_{1,B}^{\text{Slow}}$, despite Exchange Fast capturing three times the order flow of Exchange Slow. We summarize this in the Corollary below.

Corollary 1 (Equilibrium Prices) *In equilibrium, ask and bid prices at* $t = 1$ *are equal to* $ask_{1,B}^{Fast} = ask_{1,B}^{Slow} = v_0 + \frac{\mu\bar{\gamma}_B\sigma}{\mu\bar{\gamma}_B+(1-\mu)(1-\underline{\lambda}_B)}$ *and* $bid_{1,B}^{Fast} = bid_{1,B}^{Slow} = v_0 - \frac{\mu\bar{\gamma}_B\sigma}{\mu\bar{\gamma}_B+(1-\mu)(1-\underline{\lambda}_B)}$.

III.B Slow Exchange Imposes a Latency Delay

Assume now that Exchange Slow imposes a random processing delay such that market orders sent to Exchange Slow fill after v is publicly announced at $t = 2$ with probability $\delta \in (0,1)$. The processing delay impacts payoffs to speculators and liquidity investors differently. A speculator receives the following payoffs to trading at Exchanges Fast and Slow:

$$\pi_I^{\text{Fast}}(\gamma_i; \text{Buy at t=1}) = v - \text{ask}_1^{\text{Fast}} - \gamma_i, \tag{20}$$

$$\pi_I^{\text{Slow}}(\gamma_i; \text{Buy at t=1}) = (1-\delta) \times \left(v - \text{ask}_1^{\text{Slow}}\right) - \gamma_i. \tag{21}$$

A liquidity investor's payoff functions simplify to:

$$\pi_L^{\text{Fast}}(c_i; \text{Buy at t=1}) = v_0 - \text{ask}_1^{\text{Fast}}, \tag{22}$$

$$\pi_L^{\text{Slow}}(c_i; \text{Buy at t=1}) = (1-\delta) \times (v_0 - \text{ask}_1^{\text{Slow}}) - \delta \times \frac{k\lambda_i\sigma}{2}, \tag{23}$$

$$\pi_L^{\text{Int}}(c_i; \text{Buy at t=1}) = -\frac{k\lambda_i\sigma}{2}. \tag{24}$$

When Exchange Slow imposes a processing delay, an investor weighs the cost of trading on Exchange Fast immediately, against the possibility of: a) losing their information if they are in-

formed, or b) paying a delay cost if they are a liquidity investor. An investor's order placement strategy has two equilibrium conditions: i) a participation constraint (PC), and ii) an indifference condition (IC) between orders to Exchanges Fast and Slow. For a speculator, the participation constraint PC_I is the maximum information acquisition cost γ_i at which it is profitable to become informed. Then, conditional on participation, the indifference condition IC_I represents the value of β such that an informed speculator is indifferent to submitting an order to Fast or Slow. We write these conditions as:

$$\text{IC}_I\text{: } (\sigma - \text{E}[\sigma \mid \text{Buy at Fast}]) - (1-\delta)(\sigma - \text{E}[\sigma \mid \text{Buy at Slow}]) = 0, \tag{25}$$

$$\text{PC}_I\text{: } \bar{\gamma} = \Pr(\gamma_i \leq \max\{\sigma - \text{E}[\sigma \mid \text{Buy at Fast}], (1-\delta)(\sigma - \text{E}[\sigma \mid \text{Buy at Slow}])\}). \tag{26}$$

A liquidity investor faces two conditions similar to (25) and (26). The participation constraint PC_L identifies the latency sensitivity $\underline{\lambda}$ at which a liquidity investor is indifferent to trading on an exchange or the internalizer. Then, conditional on trading at an exchange, the indifference condition IC_L characterizes the value $\bar{\lambda}$ such that a liquidity investor is indifferent to submitting an order to exchange Fast or Slow. We write these conditions as:

$$\text{IC}_L\text{: } \text{E}[\sigma \mid \text{Buy at Fast}] = (1-\delta)\text{E}[\sigma \mid \text{Buy at Slow}] + \delta \times \frac{k\bar{\lambda}\sigma}{2}, \tag{27}$$

$$\text{PC}_L\text{: } \underline{\lambda} = \min\left\{\frac{2\text{E}[\sigma \mid \text{Buy at Fast}]}{k\sigma}, \frac{2\text{E}[\sigma \mid \text{Buy at Slow}]}{k\sigma}\right\}. \tag{28}$$

Inferring the participation thresholds $\underline{\lambda}$ and $\bar{\gamma}$ and the indifference thresholds β and $\bar{\lambda}$ from the investor's problem, the market-maker sets its prices at $t = 1$ using Bayes' Rule:

$$\text{ask}_1^{\text{Fast}} = v_0 + \frac{\beta\mu\bar{\gamma}\sigma}{\beta\mu\bar{\gamma} + \Pr(\text{liquidity trade at Fast})}, \tag{29}$$

$$\text{ask}_1^{\text{Slow}} = v_0 + \frac{(1-\beta)\mu\bar{\gamma}\sigma}{(1-\beta)\mu\bar{\gamma} + \Pr(\text{liquidity trade at Slow})}. \tag{30}$$

Immediately upon the announcement of v at $t = 2$, the market-maker updates its prices to $\text{ask}_2^{\text{Fast}} = \text{ask}_2^{\text{Slow}} = v$.

Taken together, an equilibrium in our model with a standard exchange and a delayed exchange is thus characterized by: (i) ask prices (29) and (30) (and symmetric bid prices) set by the market-maker at Exchanges Fast and Slow, respectively, such that they earn zero profit in expectation; (ii) a solution to the speculator's order submission problem, (25)-(26); and (iii) a solution to the liquidity investor's order submission problem, (27)-(28). Finally, we require that Assumption 1 holds (i.e., $k > \underline{k} = \frac{4((1-\mu)+2\mu\sigma)}{(1-\mu)+4\mu\sigma}$). We can now state the following existence and uniqueness theorem.

Theorem 2 (Existence and Uniqueness) *Let k satisfy Assumption 1. If $\delta \in (0, 1]$, then there exist unique values $\beta^* \in (0, 1]$, $\bar{\gamma}^*$, $\underline{\lambda}^*$, $\bar{\lambda}^*$, and prices ask_1^{Fast*}, ask_1^{Slow*} given by (29)-(30) that solve equations (25)-(28). Moreover, there exists a unique $\delta^* \in (0, 1)$ such that: i) $\delta < \delta^* \Rightarrow \beta^* \in (0, 1)$, and ii) $\delta \geq \delta^* \Rightarrow \beta^* = 1$.*

Theorem 2 illustrates that the magnitude of the delay will impact the degree to which informed speculators will segment their orders away from the delayed exchange. For a delay of sufficiently small size, informed speculators will use both exchanges, exposing market-makers at the delayed exchange to some level of adverse selection. For a large enough delay ($\delta \geq \delta^*$) informed speculators will segment completely to the non-delayed exchange.

IV Impact of a Latency Delay in a Fragmented Market

IV.A Market Quality

A latency delay on market orders impacts the trading motives of informed speculators and liquidity investors differently, leading to degrees of order flow segmentation. All informed speculators face increased price risk on orders sent to Exchange Slow, while the impact to liquidity investors depend on their private delay cost λ_i: those who are less latency-sensitive (i.e., $\lambda_i \in (0, \bar{\lambda}^*)$) respond less to price risk, and hence prefer to trade at Exchange Slow or the internalizer to ensure execution close to the public value.

We begin by characterizing the delay length, denoted δ^*, such that the price risk at Exchange Slow for informed speculators is so high that they concentrate at Exchange Fast ($\beta^* = 1$) for any delay $\delta \geq \delta^*$. We refer to δ^* as the "segmentation point." For $\delta \geq \delta^*$, no informed trading occurs at Exchange Slow, and thus $\text{ask}_1^{\text{Slow}*} = 0$. Consequently, the cost to liquidity investors from trading at Exchange Slow is no greater than their delay costs, implying that all liquidity investors weakly prefer on-exchange trading to the internalizer at $t = 1$ ($\underline{\lambda}^* = 0$). Simplifying and solving equations (25)-(28) for δ^* yields the expression:

$$\delta^*(k, \mu, \sigma) = \frac{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + 4(1-\mu)(1-\frac{2}{k})\mu\sigma} - (1-\mu)(1-\frac{2}{k})}{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + 4(1-\mu)(1-\frac{2}{k})\mu\sigma} + (1-\mu)(1-\frac{2}{k})}. \tag{31}$$

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

Proposition 1 (Order Flow Segmentation) *Compared to the benchmark case, if Exchange Slow imposes a delay $\delta \in (0, 1)$, then orders sent to the internalizer by liquidity investors decrease ($\underline{\lambda}^* \downarrow$); for $\delta \geq \delta^*$, informed speculators use only Exchange Fast ($\beta^* = 1$) and no liquidity investors use the internalizer ($\underline{\lambda}^* = 0$). Moreover, for any delay $\delta \in (0, 1)$, informed trading at Exchange Slow declines ($\beta^* \uparrow$) in δ; for $\delta \geq \delta^*$, liquidity investors migrate from Exchange Slow to Exchange Fast ($\bar{\lambda}^* \downarrow$) as δ increases.*

We illustrate Proposition 1 in Figure 3: Fig. 3a describes venue choice by informed speculators in δ, and Fig. 3b depicts the venue choice for liquidity investors with delay cost λ_i as a function of δ. As the segmentation of informed order flow to Exchange Fast ($\beta = 1$) depends heavily on δ^*, we note here the role that the innovation to the security σ plays in determining the delay length δ^* required to achieve full segmentation.

Proposition 2 (Adverse Selection) *The segmentation point $\delta^*(\sigma)$ is increasing in σ.*

Adverse selection, which is also a proxy for fundamental volatility in our model, impacts only the information acquisition decision directly: as σ increases, information acquisition becomes

more profitable, which increases $\bar{\gamma}^*$. Thus, δ^* increases in σ: for high-volatility stocks, more information acquisition leads to greater adverse selection on Exchange Fast, reducing the migration of informed speculators from Exchange Slow at any δ. Though intuition suggests that delayed exchanges would provide greater protection for liquidity investors in higher volatility environments, we show that the increased competition for information acquisition requires a longer delay to fully segment informed trading to the standard exchange.

Liquidity and Exchange Volume. Proposition 1 predicts that the introduction of a delayed exchange segments informed order flow away from Exchange Slow as price risk increases, while relatively latency-insensitive liquidity investors remain at Exchange Slow (i.e., $\lambda \in (\underline{\lambda}^*, \bar{\lambda}^*)$) and the internalizer (i.e., $\lambda \in (0, \underline{\lambda}^*)$). The result is a widening of the spread at Exchange Fast. Moreover, the reduction in adverse selection at Exchange Slow siphons the most relatively latency-sensitive liquidity investors from the internalizer ($\underline{\lambda}^* \downarrow$; Fig. 3b), further pushing the quoted spread at Exchange Slow to zero as the delay length approaches the segmentation point ($\delta \to \delta^*$). Once the latency delay exceeds δ^*, informed order flow fully segments to Exchange Fast, and thus any greater delay narrows the spread at Exchange Fast as liquidity investors with lower latency sensitivity begin to migrate to Exchange Slow ($\bar{\lambda}^* \downarrow$; Fig. 3b). We summarize these liquidity effects in Proposition 3 below, and display the dynamics over δ in Figure 4.

Proposition 3 (Quoted Spreads) *Let Exchange Slow impose a delay, $\delta \in (0, 1)$. Compared to the benchmark case, the quoted spread is narrower at Exchange Slow ($ask_1^{Slow*} \leq ask_{1,B}^{Slow}$), but wider at Exchange Fast ($ask_1^{Fast*} \geq ask_{1,B}^{Fast}$). Moreover, ask_1^{Slow*} decreases in δ; ask_1^{Fast*} increases in δ for $\delta < \delta^*$, and decreases for $\delta \geq \delta^*$.*

Taken together, Propositions 1 and 3 detail the impact of a delay on overall exchange-traded volume. We define total exchange-traded volume as the probability that an investor who arrives at the market in $t = 0$ and submits an order to either Exchange Fast or Slow in $t = 1$:

$$\text{Volume} = \mu\bar{\gamma}^* + (1 - \mu)(1 - \underline{\lambda}^*). \tag{32}$$

Equation (32) divides volume into two components: informed speculator volume, and liquidity

investor volume. The introduction of a delay segments informed speculators to Exchange Fast, where increased competition for information rents (i.e., worsening of liquidity) reduces information acquisition by speculators. The speculators with the highest information acquisition costs γ_i leave the market, reducing informed speculator volume ($\mu\bar{\gamma}^* \downarrow$; Fig. 5a). Conversely, the segmentation of informed speculators to Exchange Fast improves liquidity at the delayed exchange, which then attracts the least latency-sensitive investors, who would otherwise trade at the internalizer ($\underline{\lambda}^* \downarrow$; Fig. 3b), increasing on-exchange volume from liquidity investors ($(1-\mu)(1-\underline{\lambda}^*) \uparrow$; Fig. 5a). In aggregate, the latter effect dominates, increasing total on-exchange volume (Fig. 5b).

Proposition 4 (Exchange Volume) *Compared to the benchmark case, if Exchange Slow imposes a delay $\delta \in (0,1)$, then: i) informed speculator participation falls, ii) liquidity investor participation increases, and; iii) total exchange-traded volume increases.*

Propositions 1, 3, and 4 yield several testable predictions on the impact to market quality resulting from an exchange's implementation of a latency delay.

Empirical Prediction 1 (Market Quality) *If an exchange introduces a latency delay:*

(i) *quoted spreads tighten at the delayed exchange, and widen at standard exchanges;*

(ii) *informed trading increases on standard exchanges relative to the delayed exchange, and decreases on the delayed exchange overall, and;*

(iii) *total exchange-traded volume increases, and total informed trading falls.*

Exchanges advertise the reduction in adverse selection to liquidity suppliers as a core benefit of a latency delay. We predict that introducing a latency delay achieves this goal by redistributing informed speculators from the delayed exchange to the standard exchange. Moreover, informed speculators optimally avoid exchanges with sufficiently long delays altogether, allowing liquidity suppliers to quote a very narrow spread. The latency delay allows the exchange to effectively "cream skim" uninformed orders orders from the conventional exchange. Our prediction that exchanges with latency delays facilitate fewer informed trades is supported empiricially by two

studies that examine the introduction of a delay at TSX Alpha (Chen et al. (2017) and Anderson et al. (2018)). Moreover, Chen et al. (2017) find that spreads worsen on standard exchanges, driven by the redistribution of adverse selection.

Chen et al. (2017) also find that liquidity worsens on TSX Alpha following the introduction of a delay, via widening spreads. We argue, though, that this does not directly contradict our predictions. TSX Alpha imposes a minimum order size for liquidity providers to bypass the latency delay to post limit orders.[14] Hence, the observed negative impact on liquidity may not arise from the introduction of the delay itself, but from the increased depth requirement. Anderson et al. (2018) find no impact on the quoted spread at TSX Alpha.

IV.B Price Discovery and Welfare

Information Acquisition and Price Discovery. Our model predicts that latency delays have the desired effect of segmenting informed order flow to the non-delayed exchanges (Proposition 1), leading to improved liquidity at the delayed exchange at the expense of the standard exchange (Proposition 3). Because we abstract from trading commissions and fees, the quoted spread at each exchange reflects the price impact of a trade at $t = 1$. The intensified competition for information rents reduces informed speculator participation (Proposition 4), and thus information acquisition. Taking these together, we arrive at the following Corollary.

Corollary 2 (Price Impact and Information Acquisition) *Compared to the benchmark case, if Exchange Slow imposes a delay $\delta \in (0, 1)$, then the price impact from orders filled at $t = 1$ increases at Exchange Fast and falls at Exchange Slow, and total information acquisition falls* $(\mu\bar{\gamma}^* \downarrow)$.

Corollary 2 indicates the ambiguous impact of a delay on price discovery: despite a fall in information acquisition, the delay has divergent effects on price impact across the exchanges. Hence, we seek a measure that captures price discovery from an ex-ante, per-trader perspective. To do so,

[14] As of May 2019, more than 85% of all symbols on TSX Alpha require a minimum post-only order size greater than 500 shares. `https://www.tmxmoney.com/en/research/post_only.html`

we measure price discovery using the proportional pricing error from Zhu (2014), defined as the root mean-squared error (RMSE) of the true asset value v and the expected permanent price impact at $t = 1$, conditional on v, scaled by the innovation (absolute) value σ. We denote price discovery at $t = 1$ (using the RMSE) by RMSE_1.

$$\text{RMSE}_1 = \frac{\sqrt{\text{E}[(v - \text{price impact})^2 \mid v]}}{\sigma}. \tag{33}$$

Equation (33) 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 the scaled RMSE explicitly:

$$\text{RMSE}_1 = \sqrt{1 - \frac{\mu\bar{\gamma}^*(\beta^* \times \text{ask}_1^{\text{Fast}*} + (1-\delta)(1-\beta^*) \times \text{ask}_1^{\text{Fast}*})}{\sigma}}. \tag{34}$$

We obtain the following result on price discovery numerically. We compute the RMSE for the case where a delayed exchange is introduced alongside a non-delayed exchange (i.e., $\delta \in (0, 1]$), and center it around the benchmark value by subtracting $\text{RMSE}(\delta = 0)$. Therefore, positive values imply that introducing a delayed exchange with delay δ produces a higher RMSE than the benchmark case, thereby worsening price discovery; negative values indicate an improvement in price discovery. We illustrate the result graphically in Figure 6a using three values for the measure of speculators $\mu = \{0.25, 0.5, 0.75\}$. Other values of μ provide similar intuition.

Numerical Observation 1 (Price Discovery) *Compared to the benchmark case, 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.*

We find that the impact of a latency delay on price discovery depends on both the relative population of speculators in the market and the length of the delay. When there are relatively more speculators (high μ), the result is an unambiguous fall in price discovery following the introduction

of a delayed exchange (Fig. 6a: the solid line is non-negative for all $\delta \in (0, 1]$). With a small population of speculators (low μ), the introduction of a delay that concentrates informed order flow to the standard exchange ($\delta \geq \delta^*$) may improve price discovery (Fig. 6a: the long-dash line is negative for all $\delta > \hat{\delta} > \delta^*_{0.25}$).

We predict that a delay may improve price discovery, despite lower information acquisition. Decomposing price discovery into its components illustrates the mechanism. An informed speculator's contribution to permanent price impact has two components: i) the level of information acquisition, and ii) the venue choice by the informed speculator. As information acquisition falls overall for any delay length, the first channel worsens price discovery, as price impacts fall at each exchange. The imposition of a delay, however, also re-routes informed order flow from Exchange Slow to Exchange Fast, and the effect of this second channel is ambiguous. It is not immediately apparent whether the expected price impact from an informed trade increases or decreases as informed trading concentrates on a single exchange, nor whether this effect dominates channel one.

Numerical Observation 1 suggests that μ plays an important role in a latency delay's impact on price discovery. From the perspective of empirical testing, we acknowledge that measuring μ is difficult, as it is not clear how to identify traders who *could* choose to acquire information. Toward a remedy of this issue, we note that our model predicts that price impact increases in μ in the identical exchange benchmark setting.

Proposition 5 (Speculators and Price Impact) *For $\delta = 0$, ask_1^{Fast*} and ask_1^{Slow*} are increasing in μ.*

Proposition 5 suggests that if an exchange implements a uniform delay across a basket of securities (e.g., as implemented at IEX, TMX Alpha, Aequitas NEO, etc.,), pre-implementation differences in permanent price impact levels may proxy for differences in μ.[15]

[15]Empirically, price impact has been measured using the change in mid-quote over a specified time horizon (e.g., 30 seconds, 1 minute). In their work on the TMX Alpha speed bump introduction, Chen, Foley, Goldstein, and Ruf (2017) use a 20-second price impact; Anderson, Andrews, Devani, Mueller, and Walton (2018) use a 1-second price impact.

We stress that our model examines the "speed" of price discovery: as information is eventually impounded into prices (through trades or quotes) by the assumption that the asset value v is made public in $t = 2$, our model studies to what extent introducing a latency delay at one exchange impacts the expected level of information reflected in prices before $t = 2$. We assume that even if no latency arbitrageurs ferry information from one market to another via market orders, the information eventually reaches the market-maker who impounds the information through limit order quote revision.

Investor Welfare. Our liquidity results suggest that a latency delay benefits latency-insensitive liquidity investors, to the detriment of speculators and latency-sensitive liquidity investors. The ambiguity as to the aggregate effect raises questions about the effect of latency delays on overall investor welfare. To study investor welfare in our setting, we construct a measure that reflects allocative efficiency (similar to Bessembinder, Hao, and Zheng (2015)). Our measure aggregates the total profits from all market participants across all venues (Int denotes the internalizer). We write the welfare function W explicitly as the expected net gains from trade to an investor who enters the market at $t = 0$:

$$
\begin{aligned}
(35) \qquad W = \Pr(\text{liquidity investor}) \times \Bigg(& \int_{\bar{\lambda}^*}^{1} \pi_L^{\text{Fast}}(\lambda_i; \text{Buy Order}) + \pi_{MM}^{\text{Fast}}(\text{Sell Order}) \mathrm{d}\lambda \\
& + \int_{\underline{\lambda}^*}^{\bar{\lambda}^*} \pi_L^{\text{Slow}}(\lambda_i; \text{Buy Order}) + \pi_{MM}^{\text{Slow}}(\text{Sell Order}) \mathrm{d}\lambda \\
& + \int_{0}^{\underline{\lambda}^*} \pi_L^{\text{Int}}(\lambda_i; \text{Buy Order}) + \pi_{MM}^{\text{Int}}(\text{Sell Order}) \mathrm{d}\lambda \Bigg) \\
& + \Pr(\text{speculator}) \times \Bigg(\beta^* \int_0^{\bar{\gamma}^*} \left(\pi_I^{\text{Fast}}(\gamma_i; \text{Buy Order}) + \pi_{MM}^{\text{Fast}}(\text{Sell Order}) \right) \mathrm{d}\gamma \\
& + (1 - \beta^*) \int_0^{\bar{\gamma}^*} \left(\pi_I^{\text{Slow}}(\gamma_i; \text{Buy Order}) + \pi_{MM}^{\text{Slow}}(\text{Sell Order}) \right) \mathrm{d}\gamma \Bigg)
\end{aligned}
$$

Equation (35) simplifies considerably.[16] First, the profit to the uninformed speculator (who does not trade) is zero. Next, note that the quotes, public values, and true values net out to zero in any trade, as the transaction simply transfers these values between counterparties. Because the market-

[16] We provide a step-by-step simplification of (35) in the Appendix, Section VII.D.

maker and the internalizer have no additional costs or private values, their contributions to welfare beyond wealth transfer are zero. The delay costs paid by the liquidity investor remain, as do the information acquisition costs paid by the informed speculator. Equation (35) thus simplifies to,

$$W = \mu\left(-\int_0^{\bar{\gamma}^*}\gamma \mathrm{d}\gamma\right) + (1-\mu)\left(\int_{\bar{\lambda}^*}^{1} 0\mathrm{d}\lambda + \int_{\underline{\lambda}^*}^{\bar{\lambda}^*} -\delta\times\frac{k\sigma}{2}\lambda\mathrm{d}\lambda + \int_0^{\underline{\lambda}^*} -\frac{k\sigma}{2}\lambda\mathrm{d}\lambda\right), \tag{36}$$

$$= -\frac{\mu\bar{\gamma}^{*2}}{2} - (1-\mu)\left(\delta(\bar{\lambda}^{*2}-\underline{\lambda}^{*2}) + \underline{\lambda}^{*2}\right)\frac{k\sigma}{4}. \tag{37}$$

Our expression for welfare in (37) simplifies to two costs: delay costs incurred by liquidity investors, and resources spent on information acquisition by speculators. Liquidity investors face a rise in delay costs as they migrate to slower venues to avoid adverse selection costs generated by informed speculators at Exchange Fast, and through the implementation of the delay itself at Exchange Slow. Information acquisition costs negatively impact welfare when less efficient speculators (i.e., those with higher private information acquisition costs) find it profitable to become informed. Taking these together, we evaluate the impact of the introduction of delay at Exchange Slow on welfare by comparing welfare in the environment with a delayed exchange W, to welfare in the benchmark environment with identical exchanges W_B. We do so by computing W centered about the benchmark value $W - W_B$; hence, positive values indicate a welfare improvement from the introduction of a delayed exchange. We display our numerical result graphically in Figure 6b.

Numerical Observation 2 (Expected Welfare) *Compared to the benchmark case, any delay length* $\delta \in (0,1)$ *lowers expected welfare.*

We find that expected welfare declines for any non-maximal delay length, a result driven by the disproportionate increase in delay costs borne by liquidity investors relative to the reduction in information acquisition costs paid by speculators. We compute average information costs and average delay costs for the environment where Exchange Slow implements a delay $\delta \in (0,1]$ in Figures 7a and 7b, respectively. Similar to our numerical welfare result, our figures for average information costs and average delay costs are centered about their benchmark values. Thus, positive values indicate that these costs worsen with a delay, and negative values indicate improvement.

In equilibrium, exchange Slow offers a narrower bid-ask spread than Exchange Fast, and a lower rate of delay than the internalizer. Though this drives some liquidity investors to move on-exchange from the internalizer, these investors are among the least latency-sensitive in the market; thus average delay costs do not meaningfully reduce. The larger effect stems from investors who would submit orders to Exchange Slow in the absence of a delay. These investors are more latency-sensitive than those who move to the delayed exchange from the internalizer, and the implementation of any delay increases the rate at which they incur delay costs. The implementation of a delay does reduce average spending on information acquisition through a reduction in the profitability of informed trading at Exchange Slow, and through the overall increase in competition for information rents as informed speculators migrate to Exchange Fast. But, the increase in delay costs dominates the reduction in resources spent, leading to a decline in total investor welfare.

V Exchange Competition and Optimal Latency Delays

When one exchange in a fragmented market implements a delay, liquidity impacts are mixed, but the overall effect on welfare is negative. Because the impact on price discovery depends largely on the delay, we provide insight into what delay lengths may be implemented by exchanges, in equilibrium. Toward this goal, we examine the partial-equilibrium implementation decision of the exchange that has access to a delay technology. That is, what delay length δ would an exchange choose to implement, if any?

We assume that an exchange with access to a delay technology (Exchange Slow) strategically selects a delay $\delta \in [0, 1]$ to maximize its profits by maximizing trading volume. We contend that volume correlates with profit through per-trade access fees (abstracting from data sales, co-location services, etc.). We enrich our treatment of optimal delay length selection by considering whether the organizational relationship between the delayed exchange and the non-delayed exchange may impact their volume-maximization strategy. In Canada and U.S., for example, IEX operates as stand-alone venue, while Aequitas NEO, TSX Alpha, and NYSE American operate as subsidiaries

or partners of non-delayed exchanges. We account for these differences through assumptions on the profit motive: a stand-alone venue imposes a delay (if any) that maximizes its own volume, while a subsidiary exchange imposes a delay (if any) that maximizes the total volume across both exchanges. We assume that a subsidiary exchange would be reluctant to impose a delay that siphoned order flow away from its parent exchange, if the result were fewer trades in aggregate.

First, we assume that Exchange Slow operates as a stand-alone venue, and hence selects the delay length that maximizes its own-venue volume. We know from Proposition 1 that any delay length $\delta \geq \delta^*$ segments informed order flow to Exchange Slow, and thus the quoted spread at Exchange Slow does not improve for any delay longer than δ^*. Then, because any delay $\delta > \delta^*$ increases the rate at which liquidity investors incur delay costs, with no improvement in the spread, liquidity investors at Exchange Slow who are the most latency-sensitive will migrate to Exchange Fast for any $\delta \in (\delta^*, 1]$. The result is lower volume for any $\delta \in (\delta^*, 1]$ compared to δ^*. Hence, an optimal delay length (if any) must be in $\delta \in [0, \delta^*]$.

Second, for any positive delay length to be optimal, delayed exchange volume *conditional on a delay* must be higher than in the benchmark case (Theorem 1). Therefore, Exchange Slow will impose a delay only if doing so increases its volume. Taking this into consideration, we arrive at the following result.

Proposition 6 (Stand-alone Delayed Exchange) *Let $\delta = 0$ and $(\beta_B, \bar{\lambda}_B, \underline{\lambda}_B, \bar{\gamma}_B) \in (0,1)^4$ form an equilibrium that satisfies Theorem 1. If Exchange Slow operates independently of Exchange Fast, then it is optimal for Exchange Slow to impose a delay $\delta = (0, \delta^*]$ for any $\beta_B > \frac{1-2/k}{1-\underline{\lambda}_B}$. Moreover, any $\delta = (0, \delta^*]$ yields the same post-delay market share.*

Proposition 6 suggests that a sufficiently small stand-alone exchange will impose a short delay to provide an option for liquidity investors who would sacrifice some level of price improvement to secure faster order execution. In effect, the stand-alone exchange creates an alternative in the latency "product space" that limits adverse selection, while providing quicker order fill than the internalizer. Interestingly, we observe that delayed exchange volume is constant for all $\delta \in (0, \delta^*]$, as any outflow of informed speculators to the standard exchange is exactly offset by an inflow

of liquidity investors from the internalizer. In light of this, we posit that an exchange may elect to set the delay at the segmentation point $\delta = \delta^*$ to align with the aforementioned goal of some exchanges (e.g., Aequitas NEO and IEX) to "level the playing field" between natural investors and high-frequency traders.

Chakrabarty, Huang, and Jain (2018) find evidence that IEX's implementation of a 350 microsecond delay correlates with an increase in its volume share. The authors find no contemporaneous impact on the exchanges ICE and NASDAQ, but a -4.90% decline in volume at the CBOE. They conjecture that the difference in effects relates the responses of each exchange to IEX's delay: while ICE and NASDAQ had expressed plans to introduce their own delays, the CBOE had not.

Now, consider a delayed exchange that operates as a subsidiary of a standard exchange. Since any cross-exchange migration has a net-zero effect on profits, an optimal delay maximizes total volume by encouraging the maximal combination of speculator participation (i.e., information acquisition), and emigration of liquidity investors from the internalizer. We find that, while information acquisition is lower for all $\delta \in [0, \delta^*]$, the liquidity investor order flow siphoned from the internalizer exceeds the loss of informed order flow. Moreover, for $\delta > \delta^*$, the migration of liquidity investors from Exchange Slow to Fast reduces adverse selection on the standard exchange, incentivizing an increase in speculator information acquisition and informed order flow. Thus, the optimal delay length is $\delta = 1$.

Proposition 7 (Subsidiary Delayed Exchange) *Let $\delta = 0$ and $(\beta^*, \bar{\lambda}^*, \underline{\lambda}^*, \bar{\gamma}^*) \in (0,1)^4$ form an equilibrium that satisfies Theorem 1. If Exchange Slow operates as a subsidiary of Exchange Fast, then Exchange Slow will impose a delay $\delta = 1$ for any $\beta^* \in (0,1)$.*

If the delayed exchange operates as a subsidiary of the standard exchange, the optimal delay differs substantially from Proposition 6. In this case, Exchange Slow chooses the (effectively) maximal delay, $\delta = 1$, such that all orders fill after the market-maker updates its quotes at $t = 2$. The exchange is motivated to effectively operate an internalizer of its own, providing an on-exchange option for latency-insensitive investors. The firm prefers the maximum delay over any shorter delay, as it not only incentivizes latency-insensitive liquidity investors to migrate to the delayed

exchange, but also maintains the presence of latency-sensitive liquidity investors at its standard exchange, thus ensuring maximum speculator participation. Together, these effects maximize total on-exchange volume (Fig. 3b).

Propositions 6 and 7 predict that both stand-alone and subsidiary delayed exchanges will select some positive level of delay as part of a volume-maximization strategy. As is common practice in Canada and the U.S., delayed exchanges apply a uniform delay to all securities (e.g., 350 microsecond fixed delay at IEX; 1-3 millisecond random delay at TSX Alpha). Proposition 2 suggests, though, that fundamental volatility σ may impact the effectiveness of a uniform delay, as σ impacts the segmentation point δ^*.

Corollary 3 (Adverse Selection and Exchange Volume) *For $\delta \leq \delta^*$, an increase in σ leaves delayed volume unchanged; for $\delta > \delta^*$, delayed exchange volume increases in σ.*

While the volatility-sensitivity of δ^* does not impact the delay choice of a subsidiary exchange that invariably sets $\delta = 1$, it encourages a stand-alone exchange to set a shorter delay for securities where σ is low. If a stand-alone exchange insists on implementing a uniform delay, Corollary 3 suggests that the exchange cannot maximize volume while simultaneously minimizing the bid-ask spread (adverse selection).

Propositions 6 and 7 also focus our result on price discovery. Because a subsidiary exchange optimally selects the maximal delay ($\delta = 1$), the resulting environment is effectively unchanged from the benchmark case. A stand-alone exchange, however, will select an interior delay $\delta \in (0, \delta^*]$, if such a delay will improve their market share. Because all delays $\delta \in (0, \delta^*]$ maximize the profit of a stand-alone exchange, market organization alone does not yield an unambiguous price discovery prediction. Consider that delayed exchanges may also seek to minimize adverse selection from latency arbitrageurs as a secondary goal. Formally, if we assume that, conditional on maximizing profit, a stand-alone exchange will select the delay that minimizes the bid-ask spread at its venue, then the optimal delay length for a stand-alone exchange is $\delta = \delta^*$. With this assumption, our model yields the following numerical result, shown graphically in Figure 8a.

Numerical Observation 3 *Assume that a stand-alone delayed exchange selects the delay length* $\delta = \delta^*$*. Then, there exists a* $\hat{\mu} \in (0, 1)$ *such that price discovery improves (worsens) for* $\mu \leq \hat{\mu}$ *(*$\mu > \hat{\mu}$*).*

Numerical Observation 3 suggests that availability of a stand-alone delayed venue may improve the speed of price discovery for assets with lower ex-ante price impact (Proposition 5). Moreover, a delayed exchange may improve the efficiency of price discovery for these assets, as any improvement in price discovery occurs despite a reduction in resources spent on information acquisition (Corollary 2).

VI Conclusion

Delayed exchanges advertise latency delays as a way to "level the playing field" between fast and slow traders. As latency arbitrageurs require a speed advantage to pick off stale quotes, a latency delay can reduce this advantage, to the benefit of liquidity investors who are less latency-sensitive. Our paper analyzes the impact that introducing a latency delay has on the overall market quality of a fragmented market. We show that delayed exchange liquidity does improve, but that this comes at the expense of the standard exchange. The overall impact is a worsening of aggregate investor welfare.

The organizational relationship between the delayed exchange and the standard exchange also plays an important role on the set of optimally implementable delays. An exchange that focuses on its own profit will impose a delay that reduces overall investor welfare, while a subsidiary of a conventional exchange will essentially replicate the model of an off-exchange internalizer, leaving market quality unchanged. We show that exchanges that implement a delay are relatively small by volume share, and specialize to provide a venue that offers a middle-ground between price improvement and latency sensitivity.

Because latency delays impact the possibility of cross-market arbitrage, our study does not examine price discovery from the perspective of impounding "new" fundamental information into

prices. Instead, we examine the speed at which mispricings are corrected. We show that a delay reduces market-wide investment in identifying these pricing errors, which in most cases leads to a decline in the speed of price discovery. We find, though, that the speed of price discovery may improve for securities with a low speculator presence relative to liquidity traders—that is, securities with lower price impact before the implementation of a delay.

References

Aldrich, Eric M, and Daniel Friedman, 2019, Order protection through delayed messaging, *SSRN Working Paper 2999059*.

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.

Aoyagi, Jun, 2019, Strategic speed choice by high-frequency traders under speed bumps, *ISER Discussion Paper No. 1050*.

Baldauf, Markus, and Joshua Mollner, 2018, High-frequency trading and market performance, *SSRN Working Paper 2674767*.

Battalio, Robert, Shane A Corwin, and Robert Jennings, 2016, Can brokers have it all? on the relation between make-take fees and limit order execution quality, *Journal of Finance* 71, 2193–2238.

Battalio, Robert H, 1997, Third market broker-dealers: Cost competitors or cream skimmers?, *The Journal of Finance* 52, 341–352.

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, 2018, High-frequency trading competition, *Journal of Financial and Quantitative Analysis* (forthcoming).

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.

——— , 2019, Price discovery without trading: Evidence from limit orders, *Journal of Finance* (forthcoming).

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.

Buti, Sabrina, Francesco Consonni, Barbara Rindi, Yuanji Wen, and Ingrid M. Werner, 2015, Sub-penny and queue-jumping, *Dice Center Working Paper 2013-18*.

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

Chakrabarty, Bidisha, Jianning Huang, and Pankaj Jain, 2018, Effects of a speed bump on market quality and exchange competition, *SSRN eLibrary*.

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, 2017, The value of a millisecond: Harnessing information in fast, fragmented markets, *SSRN Working Paper 2890359*.

Cimon, David A, 2019, 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.

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

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.

Hatheway, Frank, Amy Kwan, and Hui Zheng, 2017, An empirical analysis of market segmentation on U.S. equity markets, *Journal of Financial and Quantitative Analysis* 52, 2399–2427.

Jovanovic, Boyan, and Albert J Menkveld, 2015, Middlemen in limit order markets, *SSRN eLibrary No. 1624329*.

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

Malinova, Katya, and Andreas Park, 2015, Subsidizing liquidity: The impact of make/take fees on market quality, *Journal of Finance* 70, 509–536.

——— , 2016, Does high frequency trading add noise to prices?, *Working Paper*.

Menkveld, Albert, 2016, The economics of high-frequency trading: Taking stock, *Annual Review of Financial Economics* 8, 1–24.

Menkveld, Albert J., Bart Zhou Yueshen, and Haoxiang Zhu, 2017, Shades of darkness: A pecking order of trading venues, *Journal of Financial Economics* 124, 503–534.

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

Milgrom, Paul, and Nancy Stokey, 1982, Information, trade and common knowledge, *Journal of Economic Theory* 26, 17–27.

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

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

Pagnotta, Emiliano S, and Thomas Philippon, 2018, Competing on speed, *Econometrica* 86, 1067–1115.

Rojcek, Jakub, and Alexandre Ziegler, 2016, High-frequency trading in limit order markets: Equilibrium impact and regulation, *SSRN Working Paper 2624639*.

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.

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

VII Appendix

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

VII.A List of Variables, Parameters, and Notation

Variable	Description
v	asset fundamental value at period $t = 3$
v_0	public value at period $t = 0$
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
$\text{ask}_t^{\text{Fast}}$	ask price at Exchange Fast at period t
$\text{ask}_t^{\text{Slow}}$	ask price at Exchange Slow at period t
$\text{bid}_t^{\text{Fast}}$	bid price at Exchange Fast at period t
$\text{bid}_t^{\text{Slow}}$	bid price at Exchange Slow at period t
π_I	profit function for an informed speculator
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 ($\delta = 1$)

VII.B 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, as well as all data leaving the exchange, by 350 microseconds. This alone would not prevent multi-market strategies, as traders could simply send their orders to IEX 350 microseconds in advance. However, IEX allows traders to post "pegged" orders, which move instantaneously in response to external factors. 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 behavior IEX refers to as a "crumbling quote."[17] 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 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, updating can occur before liquidity-demanding orders traverse the latency delay. For example, Canadian exchange TSX Alpha imposes a minimum order size requirement on liquidity providing orders that wish to bypass their random delay of 1 to 3 milliseconds. Liquidity providers submitting limit orders called "post-only orders" satisfy a minimum size requirement based on the price of the security, which range from 100 shares for high-priced to

[17] 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`

20,000 shares for lower-priced securities.[18]

Finally, a third type of latency delay explicitly classifies traders into two groups. Some traders are slowed by the delay, while other traders trade normally. Unlike the other types of delays that rely on order types, this form requires the explicit division of traders into two classes by the exchange. An example is the latency delay imposed by 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.[19] In the case of Aequitas Neo, those deemed to be "latency sensitive" are subjected to a randomized delay of between 3 to 9 milliseconds.

VII.C Multiple Orders from Informed Speculators

We simplify our model by assuming that investors may submit a single market order to only one exchange. In this section, we briefly discuss the possibility of multiple orders.

While it is standard in the literature to assume that liquidity investors have a fixed trading size, it is not uncommon to allow speculators to realize their information rents to the fullest extent possible. We argue that the single-order assumption plays a similar role to message and access fees, or technology costs associated with employing a multi-market strategy (e.g., co-location, order-routing). In the context of a speed bump, replacing the single-order assumption with a cost for multi-market orders yields qualitative similar results.

Suppose an informed speculator may play a multi-market latency arbitrage strategy. Informed speculators may choose to submit a market order to the standard exchange, as well as a market order to the delayed exchange.[20] The market order to the delayed exchange earns positive profit with probability $1 - \delta$, or zero profit otherwise, and hence there is no cost to submitting an extra

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

[19]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`

[20]In practice, as an alternative to sending a market order to the delayed exchange, investors may elect to use an immediate-or-cancel/fill-or-kill order, so as to transact only if liquidity is available at the price observed at time of submission.

order. In this case, informed speculators would always choose to play a multi-market strategy.

In practice, investors pay additional costs per order. For example, the Investment Industry Regulatory Organization of Canada (IIROC) mandates a messaging cost to be passed on to dealers[21], and NYSE MKT implements a tiered-fee schedule to members based on quarterly message traffic for its obligation to the Consolidated Audit Trail NMS plan.[22] Several other fees for data access, monitoring, and order routing would introduce costs to investors who send messages to many exchanges.[23] These costs can be introduced as heterogeneous messaging costs for informed speculators $\eta_i \in U[0, \infty]$ (or some finite upper bound). With this feature, informed speculators continue to submit order to the standard exchange, but only submit to the delayed-exchange if their messaging cost is sufficiently low. Investors whose messaging cost exceeds their expected profit on the delayed exchange, adopt a single-market strategy at the standard exchange. Similar to our existing model, as δ increases, the delayed exchange becomes increasingly unprofitable and informed trading becomes concentrated on the standard exchange.

[21] See sections 23a,b of the IIROC Notice:
`http://www.iiroc.ca/Documents/2016/2b56885a-9932-433c-99a2-766be291c2ce_en.pdf`

[22] See "Consolidated Audit Trail Funding Fees" in
`https://www.nyse.com/publicdocs/nyse/markets/nyse-american/NYSE_MKT_Equities_Price_List.pdf`

[23] TSX and TSX Alpha list several "Common Technology and Other Fees" associated with connecting to each exchange: `https://www.tsx.com/trading/tsx-alpha-exchange/fee-schedule`.

VII.D Simplification of Welfare Function

In this section, we provide a step-by-step simplification of Equation (35) to the form in (37).

(38)

$$
\begin{aligned}
W = \Pr(\text{liquidity investor}) \times \Bigg(& \int_{\bar{\lambda}^*}^{1} \left(\pi_L^{\text{Fast}}(\lambda_i; \text{Buy Order}) + \pi_{MM}^{\text{Fast}}(\text{Sell Order})\right) \mathrm{d}\lambda \\
& + \int_{\underline{\lambda}^*}^{\bar{\lambda}^*} \left(\pi_L^{\text{Slow}}(\lambda_i; \text{Buy Order}) + \pi_{MM}^{\text{Slow}}(\text{Sell Order})\right) \mathrm{d}\lambda \\
& + \int_{0}^{\underline{\lambda}^*} \left(\pi_L^{\text{Int}}(\lambda_i; \text{Buy Order}) + \pi_{MM}^{\text{Int}}(\text{Sell Order})\right) \mathrm{d}\lambda \Bigg) \\
& + \Pr(\text{speculator}) \times \Bigg(\beta^* \int_{0}^{\bar{\gamma}^*} \left(\pi_I^{\text{Fast}}(\gamma_i; \text{Buy Order}) + \pi_{MM}^{\text{Fast}}(\text{Sell Order})\right) \mathrm{d}\gamma \\
& + (1-\beta^*) \int_{0}^{\bar{\gamma}^*} \left(\pi_I^{\text{Slow}}(\gamma_i; \text{Buy Order}) + \pi_{MM}^{\text{Slow}}(\text{Sell Order})\right) \mathrm{d}\gamma \Bigg)
\end{aligned}
$$

(39)

$$
\begin{aligned}
= (1-\mu) \Bigg(& \frac{(1-\bar{\lambda}^*)}{2} \left(v_0 - \text{ask}_1^{\text{Fast}*} + \text{ask}_1^{\text{Fast}*} - v_0\right) \\
& + (1-\delta)(\bar{\lambda}^* - \underline{\lambda}^*)(v_0 - \text{ask}_1^{\text{Slow}*} + \text{ask}_1^{\text{Slow}*} - v_0) - \delta \frac{k\sigma(\bar{\lambda}^{*2} - \underline{\lambda}^{*2})}{4} \\
& + \left(v - v - \frac{k\sigma\underline{\lambda}^*}{4} + (v - v)\right) \frac{\underline{\lambda}^*}{2} \Bigg) \\
& + \mu\bar{\gamma}^* \Bigg(\beta^* (v - \text{ask}_1^{\text{Fast}*} - \frac{\bar{\gamma}^*}{2} + \text{ask}_1^{\text{Fast}*} - v) \\
& + (1-\beta^*) \left((1-\delta)(v - \text{ask}_1^{\text{Slow}*} - \frac{\bar{\gamma}^*}{2} + \text{ask}_1^{\text{Slow}*} - v) + \delta(v - v - \frac{\bar{\gamma}^*}{2} + (v - v)) \right) \Bigg)
\end{aligned}
$$

(40)

$$
= -\frac{\mu\bar{\gamma}^{*2}}{2} - (1-\mu)\left(\delta(\bar{\lambda}^{*2} - \underline{\lambda}^{*2}) + \underline{\lambda}^{*2}\right)\frac{k\sigma}{4}
$$

VII.E Proofs

Proof (Theorem 1). The proof that follows focuses on the actions of buyers; sellers' decisions are symmetric. As in the main text, we will use the subscript 'B' to denote benchmark equilibrium values (e.g., β_B). Informed (I) and liquidity (L) investors who submit an order at $t = 1$ to

Exchange j have profit functions given by:

$$\pi_I^j(\gamma_i; \text{Buy at t=1}) = v - \text{ask}_1^j - \gamma_i, \tag{41}$$

$$\pi_L^j(\lambda_i; \text{Buy at t=1}) = v_0 - \text{ask}_1^j. \tag{42}$$

Moreover, the profits investors to submitting an order to the internalizer are given by:

$$\pi_I^{\text{Int}}(\gamma_i) = v - v - \gamma_i, \tag{43}$$

$$\pi_L^{\text{Int}}(\lambda_i) = v - v - \frac{k\sigma\lambda_i}{2}. \tag{44}$$

Because exchanges are identical by assumption, it must be that in any equilibrium, their ask and bid prices are identical. Recall that prices are given by (18)-(19):

$$\text{ask}_1^{\text{Fast}} = \text{E}[v \mid \text{Buy at Fast}] = v_0 + \frac{\beta\mu\bar{\gamma}\sigma}{\beta\mu\bar{\gamma} + (1-\mu)\alpha\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right)}, \tag{45}$$

$$\text{ask}_1^{\text{Slow}} = \text{E}[v \mid \text{Buy at Slow}] = v_0 + \frac{(1-\beta)\mu\bar{\gamma}\sigma}{(1-\beta)\mu\bar{\gamma} + (1-\mu)(1-\alpha)\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right)}. \tag{46}$$

We then solve $\text{ask}_1^{\text{Fast}} = \text{ask}_1^{\text{Slow}}$ for $(\alpha_B, \beta_B) \in (0,1)^2$, for all $\bar{\gamma}$ and $\underline{\lambda}$:

$$\frac{\beta_B\mu\bar{\gamma}\sigma}{\beta_B\mu\bar{\gamma} + (1-\mu)\alpha_B\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right)} = \frac{(1-\beta_B)\mu\bar{\gamma}\sigma}{(1-\beta_B)\mu\bar{\gamma} + (1-\mu)(1-\alpha_B)\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right)},$$

$$\iff \beta_B(1-\alpha_B) = (1-\beta_B)\alpha_B \Rightarrow \beta_B = \alpha_B. \tag{47}$$

Given that equilibrium prices in (45) and (46) are equal, we need only solve for $\bar{\gamma}_B$ such that (41) is zero for either the Fast or Slow exchange. Letting $j = \text{Fast}$ in (41), we have:

$$\bar{\gamma}_B - (v - \text{ask}_1^{\text{Fast}}) = 0, \tag{48}$$

We now show that there exists a unique $\bar{\gamma}_B \in [0,1]$ that solves (48).

$$\bar{\gamma} = 0 : 0 - (\sigma - 0) < 0, \tag{49}$$

$$\bar{\gamma} = 1 : 1 - \sigma\left(1 - \frac{\mu}{\mu + (1-\mu)\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right)}\right) > 0, \tag{50}$$

where (50) is positive $\forall\sigma \leq 1$. Hence, $\bar{\gamma}_B$ exists. Then, differentiate equation (48) by $\bar{\gamma}$:

$$\frac{\partial}{\partial\bar{\gamma}}(\bar{\gamma} - (v - \text{ask}_1^{\text{Fast}})) = 1 + \sigma\left(\frac{(1-\mu)\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right)}{(\mu + (1-\mu)\text{Pr}\left(\lambda_i \geq \underline{\lambda}\right))^2}\right) > 0, \tag{51}$$

Then, as (48) crosses zero from below at most once, implying that $\bar{\gamma}_B$ is unique $\forall\underline{\lambda} \in [0,1]$.

We now search for the latency sensitivity value $\underline{\lambda}$ such that a liquidity investor is indifferent to trading on-exchange or at the internalizer. We show that a unique $\underline{\lambda}_B \in [0,1]$ exists by setting equal the liquidity investor profit functions (42) and (44):

$$\pi_L^{\text{Fast}}(\lambda_i) = \pi_L^{\text{Int}}(\lambda_i) \iff \frac{k\sigma\underline{\lambda}}{2} - (v_0 - \text{ask}_1^{\text{Fast}}) = 0 \tag{52}$$

Evaluating (52) at the endpoints of $\underline{\lambda}$, we have:

$$\underline{\lambda} = 0 : 0 - \frac{\mu\bar{\gamma}_B(0) \times \sigma}{\mu\bar{\gamma}_B(0) + (1-\mu)\text{Pr}\left(\lambda_i \geq 0\right)} < 0, \tag{53}$$

$$\underline{\lambda} = 1 : \frac{k\sigma}{2} - \sigma > 0, \tag{54}$$

where $\bar{\gamma}_B(\underline{\lambda} = 0) > 0$ because $\text{ask}_1^{\text{Fast}} > 0$, and (54) is positive by the assumption $k > \underline{k} > 2$. Then, because the solution to (52) solves a quadratic equation in $\underline{\lambda}$, the solution must be unique. Thus, a unique equilibrium exists for all $\beta_B = \alpha_B \in (0,1)$. ■

Proof (Theorem 2). We prove this theorem by partitioning the informed speculator's venue choice variable β into three cases.

Speculators use only Exchange Slow ($\beta^* = 0$): Consider the informed speculator's information indifference condition, where $\beta^* = 0$.

$$\text{IC}_I\text{: } \sigma - 0 - (1-\delta)\left(\sigma - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})}\right) = \delta\sigma + \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})} > 0. \tag{55}$$

Hence, there is always an incentive for an informed investor to submit an order to Exchange Fast, implying that $\beta^* \neq 0$.

Speculators use both exchanges ($\beta^* \in (0,1)$): We now solve the system of characterizing equations from (25)-(28) for $\underline{\lambda}^*, \bar{\lambda}^*, \bar{\gamma}^*$ and β^*, given the assumption that $\beta^* \in (0,1)$. We write the

characterizing equations explicitly below:

$$\text{IC}_I\text{: } 1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\left(1 - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda}-\underline{\lambda})}\right) = 0, \tag{56}$$

$$\text{PC}_I\text{: } \bar{\gamma} - \sigma\left(1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})}\right) = 0, \tag{57}$$

$$\text{IC}_L\text{: } \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda}-\underline{\lambda})} - \frac{\delta k\bar{\lambda}}{2} = 0, \tag{58}$$

$$\text{PC}_L\text{: } \frac{k\underline{\lambda}}{2} - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda}-\underline{\lambda})} = 0. \tag{59}$$

First, we rearrange (56) to solve for δ:

$$\delta = \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda}-\underline{\lambda})}. \tag{60}$$

Then, substituting equation (60) into (58) and simplifying yields the solution $\bar{\lambda}^* = 2/k$.

Now, we show that $\bar{\gamma}^* \in [0,1]$ exists for all $(\beta, \underline{\lambda}) \in (0,1) \times [0, \bar{\lambda}^*]$ ($\underline{\lambda}$ is bounded above by $\bar{\lambda}^*$ by construction). Evaluating (57) at its endpoints, we have:

$$\text{PC}_I\mid_{\bar{\gamma}=0} : 0 - \sigma < 0, \tag{61}$$

$$\text{PC}_I\mid_{\bar{\gamma}=1} : 1 - \sigma\left(1 - \frac{\mu\beta}{\mu\beta + (1-\mu)(1-\bar{\lambda})}\right) > 0, \tag{62}$$

where (62) holds by the fact that $\sigma \leq 1$. Thus, by the intermediate value theorem, $\bar{\gamma}^* \in [0,1]$ exists. To show that $\bar{\gamma}^*$ is unique, we solve (57) for the non-negative root of $\bar{\gamma}$:

$$\bar{\gamma}^* = \frac{\sqrt{(1-\mu)^2(1-2/k)^2 + 4(1-\mu)(1-2/k)\mu\beta\sigma} - (1-\mu)(1-2/k)}{2\mu\beta}. \tag{63}$$

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 (57), as $\mu \to 0 \implies \bar{\gamma}^* = \sigma \leq 1$.

We now appeal to the intermediate value theorem using (59) to show that $\underline{\lambda}^* \in [0, \bar{\lambda}^*]$ exists for

all $\beta \in (0,1)$, given $\bar{\gamma}^*$.

$$\text{PC}_L \mid_{\underline{\lambda}=0} = 0 - \frac{\mu\bar{\gamma}^*(1-\beta)}{\mu\bar{\gamma}^*(1-\beta)+(1-\mu)\times 2/k} < 0, \tag{64}$$

$$\text{PC}_L \mid_{\underline{\lambda}=2/k} = \frac{k}{2}\times\frac{2}{k} - 1 = 0, \tag{65}$$

where $\bar{\gamma}^*$ is a function of β and parameters, but not $\underline{\lambda}$. Hence, $\underline{\lambda}^* \in [0, \bar{\lambda}^*]$ exists.

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

$$\frac{\partial}{\partial\underline{\lambda}}(\text{PC}_L) = \frac{k}{2} - \frac{\mu\bar{\gamma}^*(1-\beta)(1-\mu)}{(\mu\bar{\gamma}^*(1-\beta)+(1-\mu)(2/k-\underline{\lambda}))^2}. \tag{66}$$

Because we cannot sign (66), we take the second derivative to show that (59) crosses zero from below at most once.

$$\frac{\partial^2}{\partial\underline{\lambda}^2}(\text{PC}_L) = -\frac{2\mu\bar{\gamma}^*(1-\beta)(1-\mu)^2}{(\mu\bar{\gamma}^*(1-\beta)+(1-\mu)(2/k-\underline{\lambda}))^3} < 0. \tag{67}$$

Because (67) is negative, (59) must cross zero from below at most once on $\underline{\lambda} \in [0, 2/k]$. Hence, $\underline{\lambda}^*$ is unique for all $\beta \in (0,1)$.

Lastly, we show that there is a unique $\beta^* \in (0,1)$ that solves (55), given $(\bar{\gamma}^*, \bar{\lambda}^*, \underline{\lambda}^*)$.

$$\text{IC}_I \mid_{\beta=0} : 1 - (1-\delta)\frac{\mu\sigma}{\mu\sigma+(1-\mu)(2/k-\underline{\lambda}^*)} > 0, \tag{68}$$

$$\text{IC}_I \mid_{\beta=1} : \delta - \frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^*+(1-\mu)(1-2/k)} < 0,\ \forall\delta < \frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^*+(1-\mu)(1-2/k)} = \bar{\delta}. \tag{69}$$

Thus, by the intermediate value theorem, for all $\delta < \bar{\delta}$, there exists a $\beta \in (0,1)$ that satisfies (56). To show that β^* is unique, we insert the expression for $\bar{\gamma}^*$ into (55) and differentiate (56) with

respect to β.

$$\frac{\partial}{\partial\beta}(\text{IC}_I) = -\frac{2(1-\mu)\sigma^2(k-2)^2\left(((1-\delta)/2)\times\sqrt{h(\beta)}+r(\beta)\right)\mu}{\sqrt{h(\beta)}(\sqrt{h(\beta)}+(k-2)(1-\mu))^2} < 0, \tag{70}$$

$$\text{where: } h(\beta) = (1-\mu)(k-2)(4\mu\beta k\sigma + (1-\mu)(k-2)) > 0, \tag{71}$$

$$r(\beta) = k(1-\delta)\left(\frac{(1-\mu)}{2}+\sigma(1+\beta)\mu\right)+(1+\delta)(1-\mu) > 0. \tag{72}$$

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 (56) 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 (69), we can write $\bar{\delta}$ in terms of parameters only:

$$\bar{\delta} = \frac{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2+4(1-\mu)(1-\frac{2}{k})\mu\sigma}-(1-\mu)(1-\frac{2}{k})}{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2+4(1-\mu)(1-\frac{2}{k})\mu\sigma}+(1-\mu)(1-\frac{2}{k})}. \tag{73}$$

Speculators use only Exchange Fast ($\beta^* = 1$): Here, we solve equations (25)-(28) for the case where $\beta^* = 1$. Inputting $\beta = 1$, we have the following characterizing equations:

$$\text{IC}_I\text{: } \delta - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma}+(1-\mu)(1-\bar{\lambda})} \geq 0, \tag{74}$$

$$\text{PC}_I\text{: } \bar{\gamma} - \sigma\left(1-\frac{\mu\bar{\gamma}}{\mu\bar{\gamma}+(1-\mu)(1-\bar{\lambda})}\right) = 0, \tag{75}$$

$$\text{IC}_L\text{: } -\frac{\mu\bar{\gamma}}{\mu\bar{\gamma}+(1-\mu)(1-\bar{\lambda})}+\frac{\delta k\bar{\lambda}}{2} = 0, \tag{76}$$

$$\text{PC}_L\text{: } \frac{\delta k\underline{\lambda}}{2} = 0. \tag{77}$$

First, by inspection of (77), it must be that $\underline{\lambda}^* = 0$. To prove the existence of a unique $\bar{\gamma}^*$, we solve equation (75) for the non-negative root of $\bar{\gamma}$:

$$\bar{\gamma}^* = \frac{\sqrt{(1-\mu)^2(1-\bar{\lambda})^2+4(1-\mu)(1-\bar{\lambda})\mu\sigma}-(1-\mu)(1-\bar{\lambda})}{2\mu}. \tag{78}$$

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 $\mu = 0$, (75) admits the limit $\bar{\gamma} = \sigma$. Thus, $\bar{\gamma}^*$ is unique.

Next, we show that there exists a unique $\bar{\lambda}^* \in [0, 2/k]$ that solves (76) for all $\delta \geq \underline{\delta}$. We can bound $\bar{\lambda}^* \in [0, 2/k]$ because for any $\bar{\lambda}^* > 2/k$, (76) would be negative if the required inequality in (74) holds. First, we show that $\bar{\lambda}^*$ exists by evaluating $\bar{\lambda}$ at 0 and $2/k$:

$$\text{IC}_L \mid_{\bar{\lambda}=0} : \frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^* + (1-\mu)} - 0 > 0, \tag{79}$$

$$\text{IC}_L \mid_{\bar{\lambda}=2/k} : \frac{\mu\bar{\gamma}^*(2/k)}{\mu\bar{\gamma}^*(2/k) + (1-\mu)(1-2/k)} - \delta < 0, \tag{80}$$

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

Hence, by the continuity of (76) 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 $\delta > \underline{\delta}$ on $\bar{\lambda} \in [0, 2/k]$. Differentiating (76) with respect to $\bar{\lambda}$:

$$\frac{\partial}{\partial\bar{\lambda}}(IC_L) = \frac{\mu\bar{\gamma}^*(1-\mu)}{(\mu\bar{\gamma}^* + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial\bar{\gamma}^*}{\partial\bar{\lambda}} \times \frac{\mu(1-\bar{\lambda})(1-\mu)}{(\mu\bar{\gamma}^* + (1-\mu)(1-\bar{\lambda}))^2} - \frac{\delta k}{2}. \tag{81}$$

To see that (81) is negative, note that condition (74) 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 (81). Computing $\frac{\partial\bar{\gamma}^*}{\partial\bar{\lambda}}$ and simplifying, we obtain the inequality:

$$\frac{\partial}{\partial\bar{\lambda}}(IC_L) < -\frac{2(1-\mu)(1-\bar{\lambda})\left(k \times v(\bar{\lambda}) - 2(1-\mu)\right)\mu\sigma}{v(\bar{\lambda})\left((1-\mu)^2(1-\bar{\lambda})^2 + v(\bar{\lambda})\right)^2}, \tag{82}$$

where $v(\bar{\lambda}) = \sqrt{(1-\mu)(1-\bar{\lambda})\left((1-\mu)(1-\bar{\lambda}) + 4\mu\sigma\right)}$. Then, evaluating (82) at $\bar{\lambda} = 2/k$, the equation (82) is negative if and only if $v(\bar{\lambda} = 2/k) > \frac{2}{k}(1-\mu) \iff k > \frac{4(1-\mu+2\mu\sigma)}{1-\mu+4\mu\sigma} = \underline{k}$, which is satisfied by Assumption 1. Hence, for all $k > \underline{k}$, $\underline{\delta}$ is the lowest δ for which a solution to (76) 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, $\underline{\delta} = \bar{\delta} = \delta^*$.

Finally, we show that condition (74) is satisfied by $\bar{\lambda}^*$, $\underline{\lambda}^*$, and $\bar{\gamma}^*$. Inputting $\bar{\gamma}^*$ and $\underline{\lambda}^*$ into

condition (74) and differentiating by $\bar{\lambda}$, we obtain:

$$\frac{\partial \text{IC}_I}{\partial \bar{\lambda}} = \frac{4\mu\sigma(1-\mu)^2(1-\bar{\lambda})}{v(\bar{\lambda})(v(\bar{\lambda})-(1-\mu)(1-\bar{\lambda}))^2} > 0, \tag{83}$$

where $v(\bar{\lambda})$ is as above. Then, because the second term of (74) is maximized at $\bar{\lambda}$ and equal to δ^*, condition (74) is satisfied for all $\delta \geq \delta^*$. ■

Proof (Proposition 1). We begin by showing that $\beta^* \geq \beta_B^*$ and $\underline{\lambda}^* \leq \underline{\lambda}_B$ 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 $\frac{\partial \text{ask}_1^{\text{Slow}*}}{\partial \delta} < 0$. To show this, we note that given the existence of unique equilibrium, the following condition in $\underline{\lambda}^*$ must be satisfied:

$$\frac{k\underline{\lambda}^*}{2} - \text{ask}_1^{\text{Slow}*} = 0 \iff \frac{k\underline{\lambda}^*}{2} - \frac{\mu\bar{\gamma}^*(1-\beta^*)}{\mu\bar{\gamma}^*(1-\beta^*)+(1-\mu)(\bar{\lambda}^*-\underline{\lambda}^*)} = 0. \tag{84}$$

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 $\text{ask}_1^{\text{Fast}*}$ from IC_L into the speculator's information acquisition condition PC_I, which yields the expression $\bar{\gamma}^* = \frac{\sigma(2-\delta\bar{\lambda}^*k)}{2}$. 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). To show that δ^* is increasing in σ, we take the derivative of (31) with respect to σ, and show that $\frac{\partial \delta^*}{\partial \sigma} > 0$. Computing the derivative yields $\frac{\partial \delta^*}{\partial \sigma} = \frac{\mu x^2}{(\sqrt{x^2+x\mu\sigma}+x)^2} > 0$, where $x = (1-\mu)(1-\frac{2}{k})$. Thus, δ^* is increasing in σ. ■

Proof (Proposition 3). For the half spread at Exchanges Fast and Slow, $\text{ask}_1^{\text{Fast}*}$ and $\text{ask}_1^{\text{Slow}*}$, 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 $\text{ask}_1^{\text{Fast}*} = \text{ask}_1^{\text{Slow}*} = \frac{\mu\bar{\gamma}^*\sigma}{\mu\bar{\gamma}^*+(1-\mu)(1-\underline{\lambda}^*)} > 0$. Now, consider $\text{ask}_1^{\text{Fast}*}$. We know from the speculator's information acquisition condition PC_I that $\text{ask}_1^{\text{Fast}*}$ moves inversely to $\bar{\gamma}^*$, as $\bar{\gamma}^* = \left(\sigma - \text{ask}_1^{\text{Fast}*}\right)$.

For $\delta \in (0, \delta^*)$, we know that $\bar{\gamma}^*$ is decreasing in δ from the proof of Proposition 1, which implies that $\text{ask}_1^{\text{Fast}*}$ 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 $\text{ask}_1^{\text{Fast}*} \mid_{\delta=0} = \text{ask}_1^{\text{Fast}*} \mid_{\delta=1}$, it must be that $\text{ask}_1^{\text{Fast}*}(\delta) > \text{ask}_1^{\text{Fast}*} \mid_{\delta=0}$ for all $\delta \in (0, 1)$. Now, consider $\text{ask}_1^{\text{Slow}*}$. Let $\delta \in [\delta^*, 1]$. By the proof of Theorem 2, $\underline{\lambda}^* = 0$, and thus $\text{ask}_1^{\text{Slow}*} = 0 < \text{ask}_1^{\text{Slow}*} \mid_{\delta=0}$. For $\delta \in (0, \delta^*)$, $\frac{\partial \text{ask}_1^{\text{Slow}*}}{\partial \delta} < 0$ follows from Proposition 1, as $\underline{\lambda}$ declines in δ. ■

Proof (Proposition 4). Total exchange volume is given by the expression:

$$\text{Volume} = \mu\bar{\gamma}^* + (1-\mu) \times (1 - \underline{\lambda}^*). \tag{85}$$

For $\delta \in (\delta^*, 1)$, we know that $\underline{\lambda}^* = 0$, and thus $\frac{\partial \bar{\gamma}^*}{\partial \delta} > 0$ implies that Volume increases in δ on $[\delta^*, 1]$. Now let $\delta \in (0, \delta^*]$. Recall that $\bar{\lambda}^* = 2/k$. Thus, we can simplify (27) to obtain:

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

Using this fact, we can simplify (85) to $\text{Volume} = \mu\bar{\gamma}^*\beta^* + (1-\mu)$. Recall that from the proof of Proposition (3), $\text{ask}_1^{\text{Fast}*}(\delta < \delta^*) = \frac{\mu\bar{\gamma}^*\beta^*}{\mu\bar{\gamma}^*\beta^*+(1-\mu)(1-2/k)}$ is increasing in δ. By computing this derivative, we have that $\frac{\partial \text{ask}_1^{\text{Fast}*}}{\partial \delta} = \frac{\partial \text{ask}_1^{\text{Fast}*}}{\partial(\bar{\gamma}^*\beta^*)} \times \frac{\partial(\bar{\gamma}^*\beta^*)}{\partial \delta} > 0$. Then, $\frac{\partial \text{ask}_1^{\text{Fast}*}}{\partial(\bar{\gamma}^*\beta^*)} = \frac{\bar{\gamma}^*\beta^*(1-\mu)\sigma}{(\bar{\gamma}^*\beta^*+(1-\mu)(1-2/k))^2} > 0$, which implies that $\frac{\partial(\bar{\gamma}^*\beta^*)}{\partial \delta} > 0$. ■

Proof (Proposition 5). In the equilibrium described by Theorem 1, price impact at either exchange when $\delta = 0$ is given by the half-spread, $\text{ask}_{1,B}^{\text{Fast}} = \text{ask}_{1,B}^{\text{Slow}}$. The expression for $\text{ask}_{1,B}^{\text{Fast}}$ is only a function of μ, σ, $\bar{\gamma}_B$ and $\underline{\lambda}_B$, as β_B and α_B cancel out. Thus, we have:

$$\text{ask}_1^{\text{Fast}} = \frac{\mu\bar{\gamma}_B\sigma}{\mu\bar{\gamma}_B + (1-\mu)\text{Pr}(\lambda_i \geq \underline{\lambda}_B)}. \tag{86}$$

Next, recall the participation constraints from the identical market benchmark:

$$\text{PC}_L : \frac{\underline{\lambda}_B k\sigma}{2} - \text{ask}_{1,B}^{\text{Fast}} = 0, \tag{87}$$

$$\text{PC}_I : \bar{\gamma}_B - (\sigma - \text{ask}_{1,B}^{\text{Fast}}) = 0. \tag{88}$$

Solve (87) for $\text{ask}_{1,B}^{\text{Fast}}$, and substitute into (88) to solve for $\bar{\gamma}_B$ as a function of $\underline{\lambda}_B$: $\bar{\gamma}_B = \sigma - \frac{\underline{\lambda}_B k\sigma}{2}$. Then, (86) becomes,

$$\text{ask}_{1,B}^{\text{Fast}} = \frac{\mu\left(\sigma - \frac{\underline{\lambda}_B k\sigma}{2}\right)\sigma}{\mu\left(\sigma - \frac{\underline{\lambda}_B k\sigma}{2}\right)) + (1-\mu)\text{Pr}(\lambda_i \geq \underline{\lambda}_B)}. \tag{89}$$

Differentiating $\text{ask}_{1,B}^{\text{Fast}}$ by μ, we obtain:

$$\frac{\partial \text{ask}_{1,B}^{\text{Fast}}}{\partial \mu} = \frac{\left(\sigma - \frac{\underline{\lambda}_B k\sigma}{2}\right)\text{Pr}(\lambda_i \geq \underline{\lambda}_B)\sigma}{\mu\left(\sigma - \frac{\underline{\lambda}_B k\sigma}{2}\right) + (1-\mu)\text{Pr}(\lambda_i \geq \underline{\lambda}_B)} + \frac{\partial \text{ask}_{1,B}^{\text{Fast}}}{\partial \underline{\lambda}_B} \times \frac{\partial \underline{\lambda}_B}{\partial \mu}. \tag{90}$$

Next, we differentiate the participation constraint (87) to provide an expression for $\frac{\partial \underline{\lambda}_B}{\partial \mu}$:

$$\frac{\partial \underline{\lambda}_B}{\partial \mu} = \frac{\partial \text{ask}_{1,B}^{\text{Fast}}}{\partial \mu}. \tag{91}$$

Inputting the expression for $\frac{\partial \underline{\lambda}_B}{\partial \mu}$ obtained from (91) into (90) and solving for $\frac{\partial \text{ask}_{1,B}^{\text{Fast}}}{\partial \mu}$ and simplifying, we obtain:

$$\frac{\partial}{\partial \mu}(\text{ask}_{1,B}^{\text{Fast}}) = -\frac{\mu k(1-\mu)(k-2)\sigma^3}{\left((1-\mu) + \left(\sigma - \frac{\underline{\lambda}_B k\sigma}{2}\right)\mu - 2(1-\mu)\frac{\underline{\lambda}_B k\sigma}{2}\right)^2} < 0. \tag{92}$$

Hence, the price impact of trades when $\delta = 0$ is increasing in μ. ■

Proof (Proposition 6). To determine the maximum level of delayed exchange market share such that implementing a delay is a profit-maximizing action, we characterize the equilibrium values for β, $\bar{\gamma}$, $\bar{\lambda}$, and $\underline{\lambda}$ that, in the limit as $\delta \to 0$, simultaneously satisfy Theorem 1 and 2 in the following Lemma.

Lemma 1 (Limit Equilibrium) *Let $\delta \to 0$. Then, the equilibrium values in Theorem 2:*

- $(\lim_{\delta\to 0}\beta, \lim_{\delta\to 0}\bar{\lambda}) \to \left(\frac{1-2/k}{1-\underline{\lambda}}, \frac{2}{k}\right)$
- $\lim_{\delta\to 0}\bar{\gamma} \to \frac{\sqrt{((1-\underline{\lambda})x)^2+4x(1-\underline{\lambda})(1-2/k)\mu\sigma}-x}{2\mu(1-2/k)}$, *where* $x=(1-\mu)(1-2/k)$
- $\lim_{\delta\to 0}\underline{\lambda} \to \frac{1-\mu(1-2\sigma)-\sqrt{(1-\mu)((1-\mu)+4\mu(1-2\sigma/k)\sigma)}}{2+(k-2)\mu}$

Proof (Lemma 1). We begin by recalling the indifference conditions from Theorem 2:

$$\text{IC}_I\text{: } 1-\frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta+(1-\mu)(1-\bar{\lambda})}-(1-\delta)\left(1-\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta)+(1-\mu)(\bar{\lambda}-\underline{\lambda})}\right)=0, \tag{93}$$

$$\text{PC}_I\text{: } \bar{\gamma}-\sigma\left(1-\frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta+(1-\mu)(1-\bar{\lambda})}\right)=0, \tag{94}$$

$$\text{IC}_L\text{: } \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta+(1-\mu)(1-\bar{\lambda})}-(1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta)+(1-\mu)(\bar{\lambda}-\underline{\lambda})}-\frac{\delta k\bar{\lambda}}{2}=0, \tag{95}$$

$$\text{PC}_L\text{: } \frac{k\underline{\lambda}}{2}-\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta)+(1-\mu)(\bar{\lambda}-\underline{\lambda})}=0. \tag{96}$$

For this proof, we take some $\delta > 0$ such that $\delta < \delta^*$ (which exists, for all μ, σ and $k > \underline{k}$, because $\delta^* > 0$). Thus, the equilibrium is such that $\beta^* \in (0,1)$. By simplifying (95) using (93), we arrive at the solution for $\bar{\lambda}^* = \frac{2}{k}$, which is constant in the limit as $\delta \to 0$.

Next, consider β^* from (93). Taking the limit $\lim_{\delta\to 0}\text{IC}_I$ yields that $\text{ask}_1^{\text{Fast}*} = \text{ask}_1^{\text{Slow}*}$, which holds if and only if $\beta^* = \frac{1-\bar{\lambda}^*}{1-\underline{\lambda}^*} = \frac{1-2/k}{1-\underline{\lambda}^*}$, given that $\lim_{\delta\to 0}\underline{\lambda}^*$ exists.

Now, consider the solution for $\bar{\gamma}^*$, which obtains from (94). Using the solution from the proof of Theorem 2, we see that $\bar{\gamma}^*$ is not directly a function of δ, but instead is a function of β^* and $\bar{\lambda}^*$, which have limiting values as above, conditional on the existence of $\lim_{\delta\to 0}\underline{\lambda}^*$. Hence, $\bar{\gamma}^* = \frac{\sqrt{((1-\underline{\lambda}^*)x)^2+4x(1-\underline{\lambda}^*)(1-2/k)\mu\sigma}-x}{2\mu(1-2/k)}$, where $x=(1-\mu)(1-2/k)$.

Lastly, take (96), and simplify using the limiting expression for β^*. We obtain:

$$\text{PC}_L\text{: } \frac{k\underline{\lambda}^*}{2}-\frac{\mu\bar{\gamma}^*}{\mu\bar{\gamma}^*+(1-\mu)(1-\underline{\lambda}^*)}=0. \tag{97}$$

Then, taking the limit $\lim_{\delta\to 0}\text{PC}_L$ and solving for $\underline{\lambda}^*$, we obtain a single non-negative root in $[0, \bar{\lambda}^*]$, $\underline{\lambda}^* = \frac{1-\mu(1-2\sigma)-\sqrt{(1-\mu)((1-\mu)+4\mu(1-2\sigma/k)\sigma)}}{2+(k-2)\mu}$. ■

We now show that any $\delta \in (0, \delta^*]$ is an optimal delay length to maximize volume at Exchange Slow, conditional on $\beta > \frac{1-2/k}{1-\underline{\lambda}^*}$. Let Exchange Slow operate as a stand-alone exchange, and as

such, maximizes only its own volume:

$$\text{Volume}^{\text{Slow}} = \mu\bar{\gamma}^*(1-\beta^*) + (1-\mu)(\bar{\lambda}^* - \underline{\lambda}^*). \tag{98}$$

From the proof of Proposition 4, we know that $\mu\bar{\gamma}^*(1-\beta^*) = (1-\mu)\underline{\lambda}^*$ for $\delta \in (0, \delta^*]$, which implies that $\text{Volume}^{\text{Slow}} = (1-\mu)\bar{\lambda}^*$, a constant. Then, for $\delta \in (\delta^*, 1]$, we know $\underline{\lambda}^* = 0$ and $\beta^* = 1$, implying that again that $\text{Volume}^{\text{Slow}} = (1-\mu)\bar{\lambda}^*$, which is decreasing in $\bar{\lambda}^*$ for all $\delta \in (\delta^*, 1]$. Thus, any $\delta \in (0, \delta^*]$ maximizes delayed exchange volume.

Then, by Lemma 1, $\bar{\lambda}_B = 2/k = \bar{\lambda}^* \mid_{\delta>0}$, and $\beta^* = \frac{1-2/k}{1-\underline{\lambda}_B}$. Thus, a stand-alone exchange elects to introduce a delay for any $(1-\beta^*) < \frac{2/k-\underline{\lambda}_B}{1-\underline{\lambda}_B} \iff \beta^* > \frac{1-2/k}{1-\underline{\lambda}_B}$. ■

Proof (Proposition 7). Let Exchange Slow operate as a subsidiary of Exchange Fast. Then, Exchange Slow will set a delay δ, such that the sum of all volume across the delayed and standard exchanges is maximized. From Proposition 4, recall that total exchange volume is increasing for all δ. Thus, we have that the optimal delay is $\delta = 1$ for all $\delta \in (0, 1]$.

Now, let $\delta = 0$. Recall that total exchanged-traded volume in the benchmark case Volume_B is invariant for all $\beta_B = \alpha_B$, as:

$$\text{Volume}_B = \mu\bar{\gamma}_B(\beta_B + (1-\beta_B)) + (1-\mu)(\alpha_B \text{Pr}(\lambda_i \geq \underline{\lambda}_B) + (1-\alpha_B)\text{Pr}(\lambda_i \geq \underline{\lambda}_B)) \tag{99}$$

$$= \mu\bar{\gamma}_B + (1-\mu)(1-\underline{\lambda}_B). \tag{100}$$

Then, because total volume is invariant in the initial market shares of Exchanges Fast and Slow, and there exists a unique β_B such that the limit of $\beta^* \to \beta_B$ as $\delta \to 0$, it must be that $\text{Volume}_B <$ Volume $\mid_{\delta=1}$. Hence, a subsidiary exchange would always impose $\delta = 1$ to maximize total exchange-traded volume. ■

Figure 1: Length of Latency Delay

Figure 1 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.

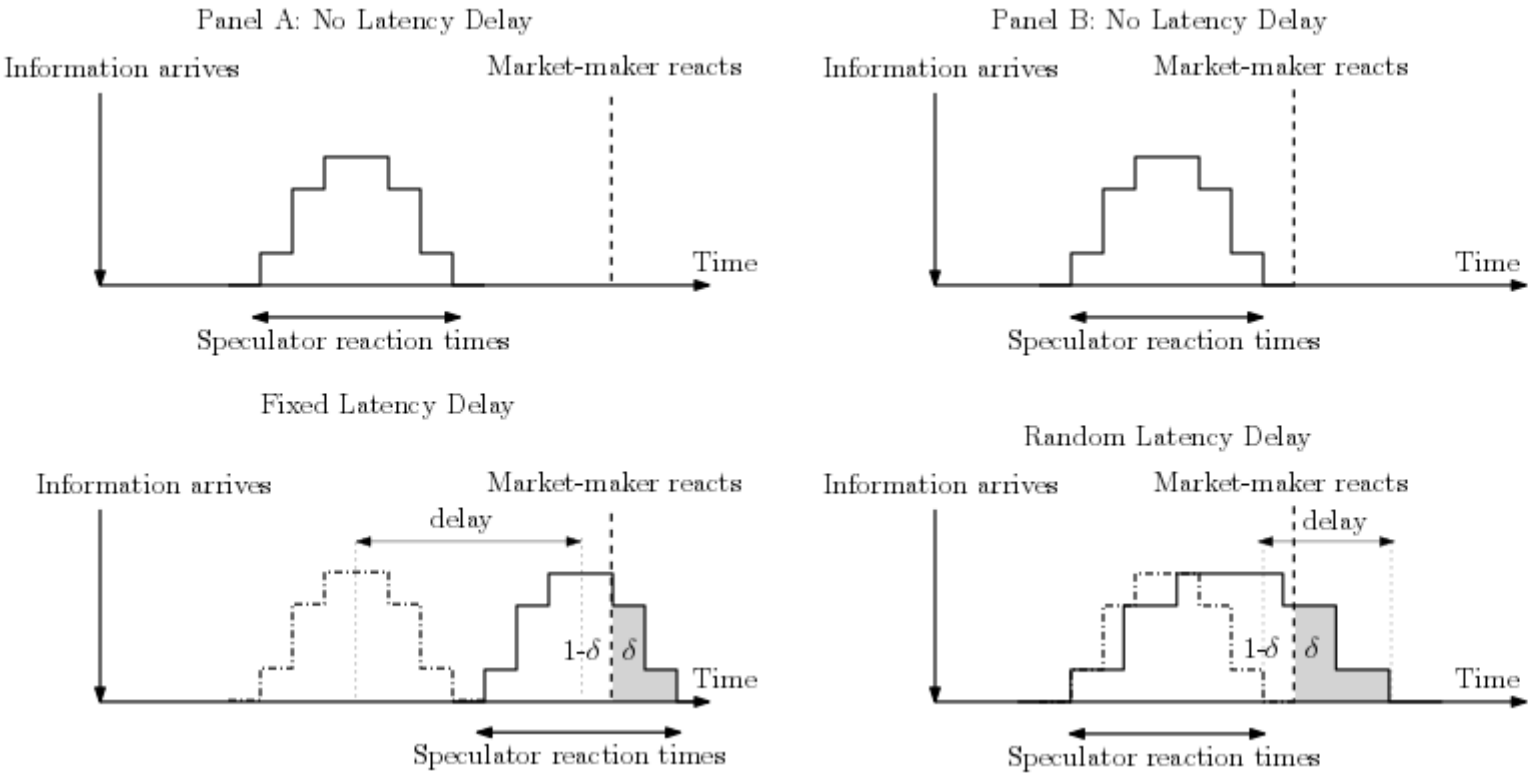


Figure 2: Model Timeline

This figure illustrates the timing of events upon an investor's arrival at $t = 0$, until their payoff is realized at $t = 3$.

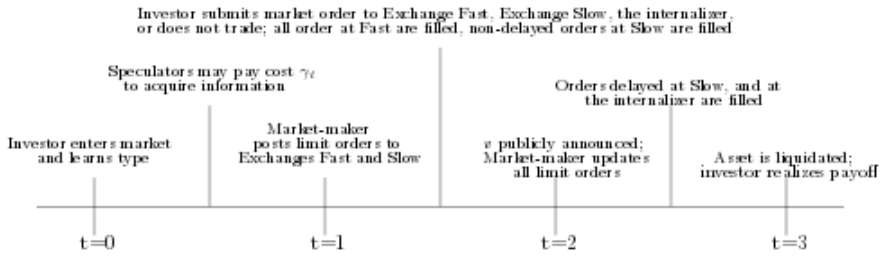
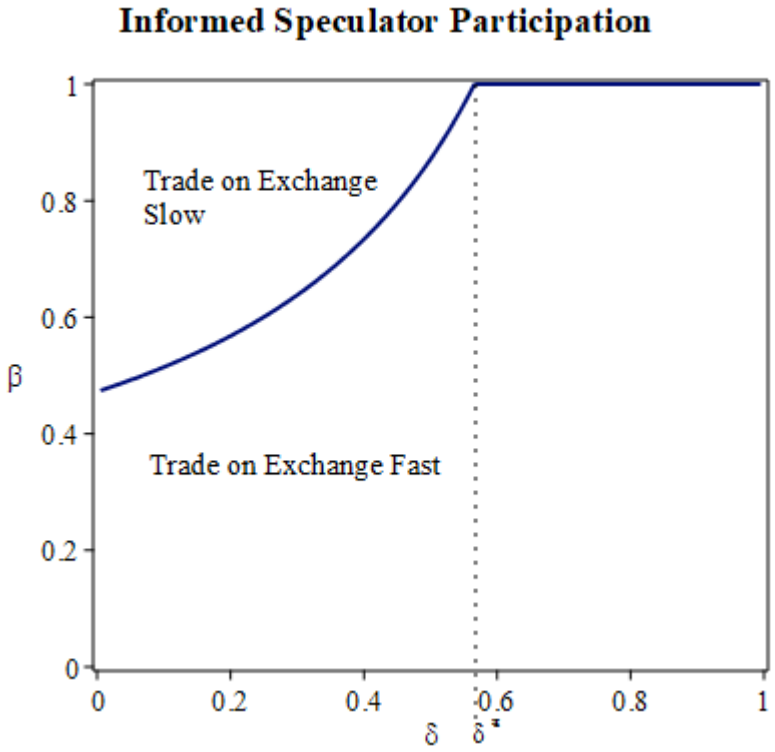
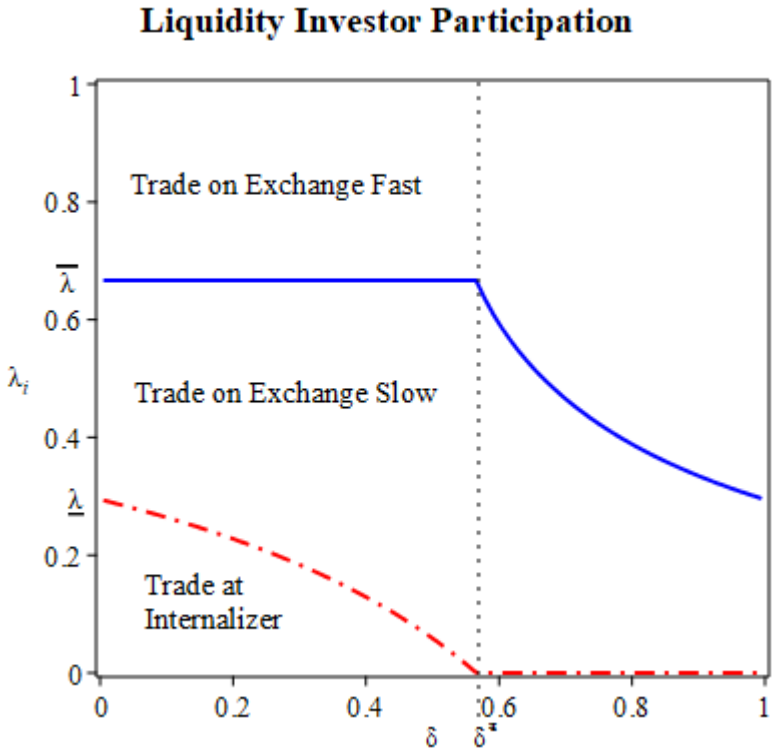


Figure 3: Market Participation by Investor Type

Panel (a) below depicts the probability that an informed speculator sends an order to Exchange Fast (solid line), β, as a function of the latency delay δ; the difference between the β plot and 1 marks the probability that an informed speculator sends an order to Exchange Slow. Panel (b) illustrates the venue choice of liquidity investors, as a function of δ. For a value of λ_i such that: i) $\lambda_i \in [\bar{\lambda}, 1]$, a liquidity investor submits an order to Exchange Fast, ii) $\lambda_i \in [\underline{\lambda}, \bar{\lambda})$ a liquidity investor submits an order to Exchange Slow, and iii) $\lambda_i \in [0, \underline{\lambda})$, a liquidity investor submits an order to the internalizer. A vertical dotted line marks δ^*: for all $\delta > \delta^*$, informed speculators use only Exchange Fast. Horizontal wide-spaced dashed lines mark values for the benchmark case. Parameters $\mu = 0.5, k = 3, \sigma = 1$. Results for other values of μ, k and σ are qualitatively similar.



(a)



(b)

Figure 4: Quoted Spreads

The plot below presents the quoted half-spreads for Exchanges Fast and Slow at $t = 1$, as a function of the latency delay δ. The solid line graphs $\text{ask}_1^{\text{Fast}*}$, while the dash-dot line graphs $\text{ask}_1^{\text{Slow}*}$. A vertical dotted line marks δ^*: for all $\delta > \delta^*$, informed speculators use only Exchange Fast. A horizontal wide-spaced dashed line marks the benchmark value. Parameters $\mu = 0.5, k = 3, \sigma = 1$. Results for other values of μ, k and σ are qualitatively similar.

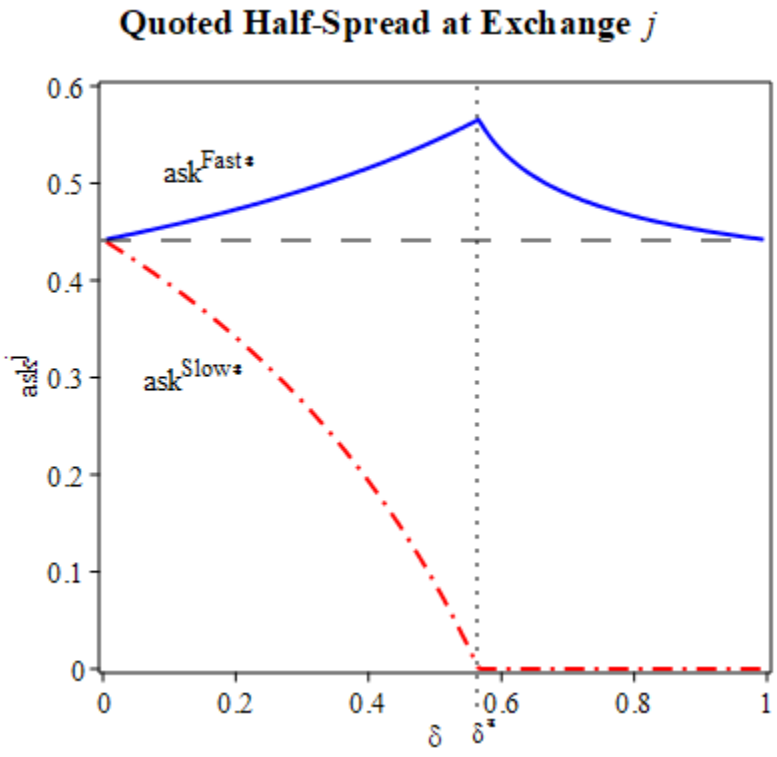
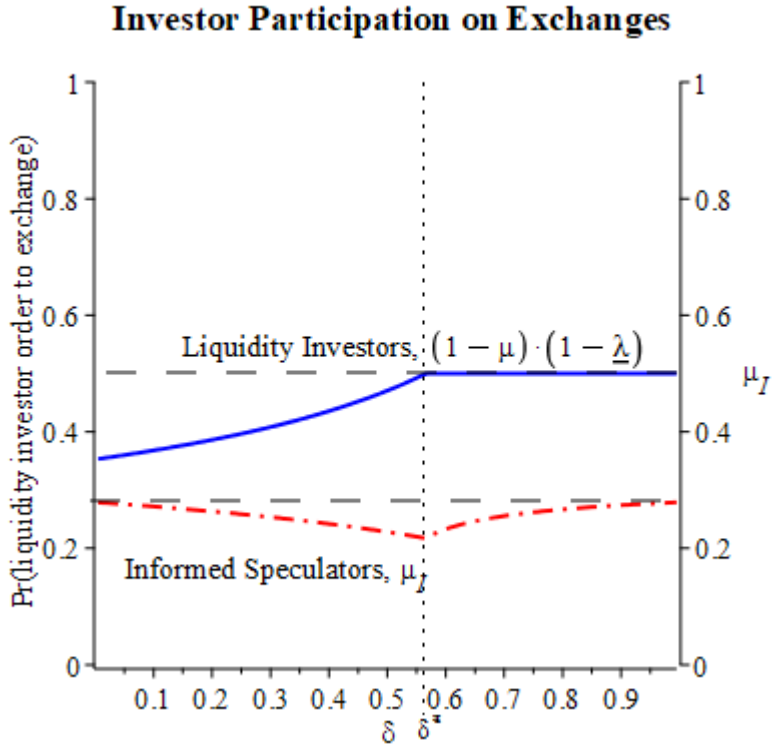
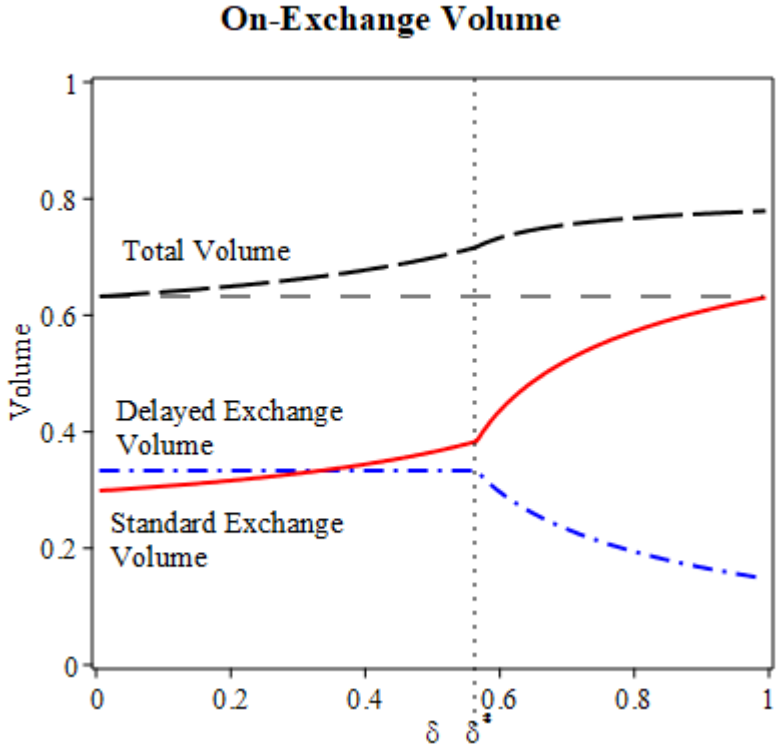


Figure 5: Order Submissions, Trades, and Market Participation

Panel (a) illustrates the probability of an order submission to any exchange by i) speculators (μ_I, dash-dot line), and ii) liquidity investors ($(1-\mu)(1-\underline{\lambda})$, solid), as a function of the Exchange Slow latency delay δ. Panel (b) provides three volume figures: total exchange volume (long-dash line), delayed (dash-dot), and standard exchange volume (solid), as a function of δ. A vertical dotted line marks δ^*: for all $\delta > \delta^*$, informed speculators use only Exchange Fast. Horizontal wide-spaced dashed lines mark values for the benchmark case. We note that there are no benchmark values indicated for the Delayed Exchange Volume and Standard Exchange volume plots, as pre-implementation volume values ($\delta = 0$) depend on β_B and α_B. Parameters $\mu = 0.5, k = 3, \sigma = 1$. Results for other values of μ, k and σ are qualitatively similar.



(a)



(b)

Figure 6: Price Discovery (Scaled RMSE at $t = 1$) and Expected Welfare

Panel (a) below depicts our numerical results related to price discovery. We plot the root-mean-squared proportional pricing error at $t = 1$ (deviation from the true value at $t = 1$) against the delay length δ, centred about the benchmark level ($\delta = 0$). Thus, positive values indicate that a positive delay worsens price discovery. Panel (b) illustrates the expected welfare realized by an investor who enters the market at $t = 0$, also centred about the benchmark level ($\delta = 0$). Line styles long-dash, dash-dot and solid reflect parameter values $\mu = \{0.25, 0.5, 0.75\}$, respectively. The benchmark value ($\delta = 0$) is marked with a horizontal wide-spaced dashed line. Vertical dotted lines mark δ^* at each plotted value of μ, indicated by $\delta^*_{0.25}, \delta^*_{0.5}$, and $\delta^*_{0.75}$. Parameters $k = 3, \sigma = 1$. Results for other values of k and σ are qualitatively similar.

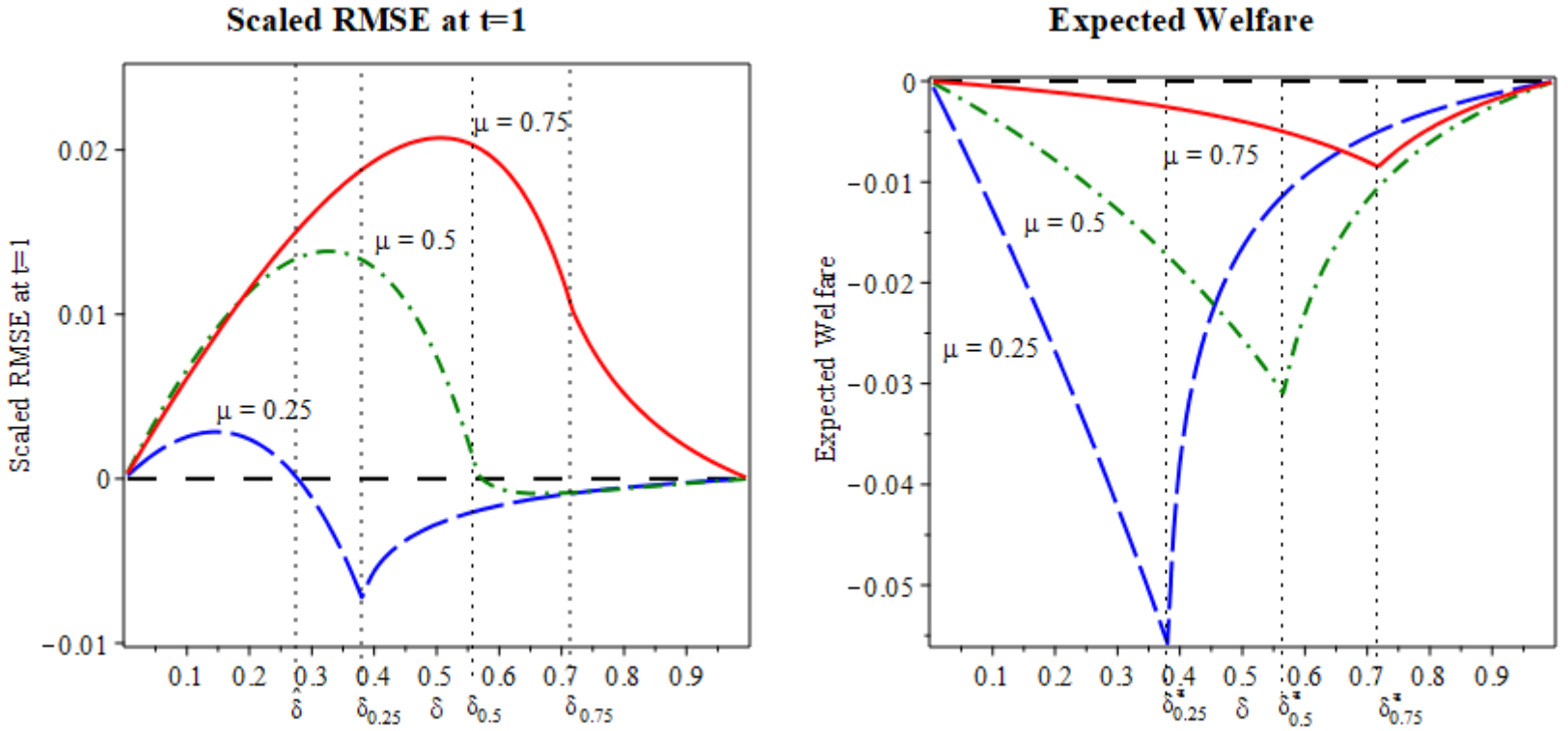


Figure 7: Components of Expected Welfare

Panel (a) plots the average information acquisition costs paid by informed speculators; panel (b) plots the average delay costs incurred by liquidity investors. We present these costs as a function of the latency delay δ. Both plots are centred about the benchmark level ($\delta = 0$), marked by the horizontal wide-spaced dashed line. Thus, negative values indicate an improvement in costs over the non-delayed environment, and positive values indicate a worsening. Line styles long-dash, dash-dot and solid reflect parameter values $\mu = \{0.25, 0.5, 0.75\}$, respectively. Vertical dotted lines indicate δ^* at each plotted value of μ, indicated by $\delta^*_{0.25}, \delta^*_{0.5}$, and $\delta^*_{0.75}$. Parameters $k = 3, \sigma = 1$. Results for other values of k and σ are qualitatively similar.

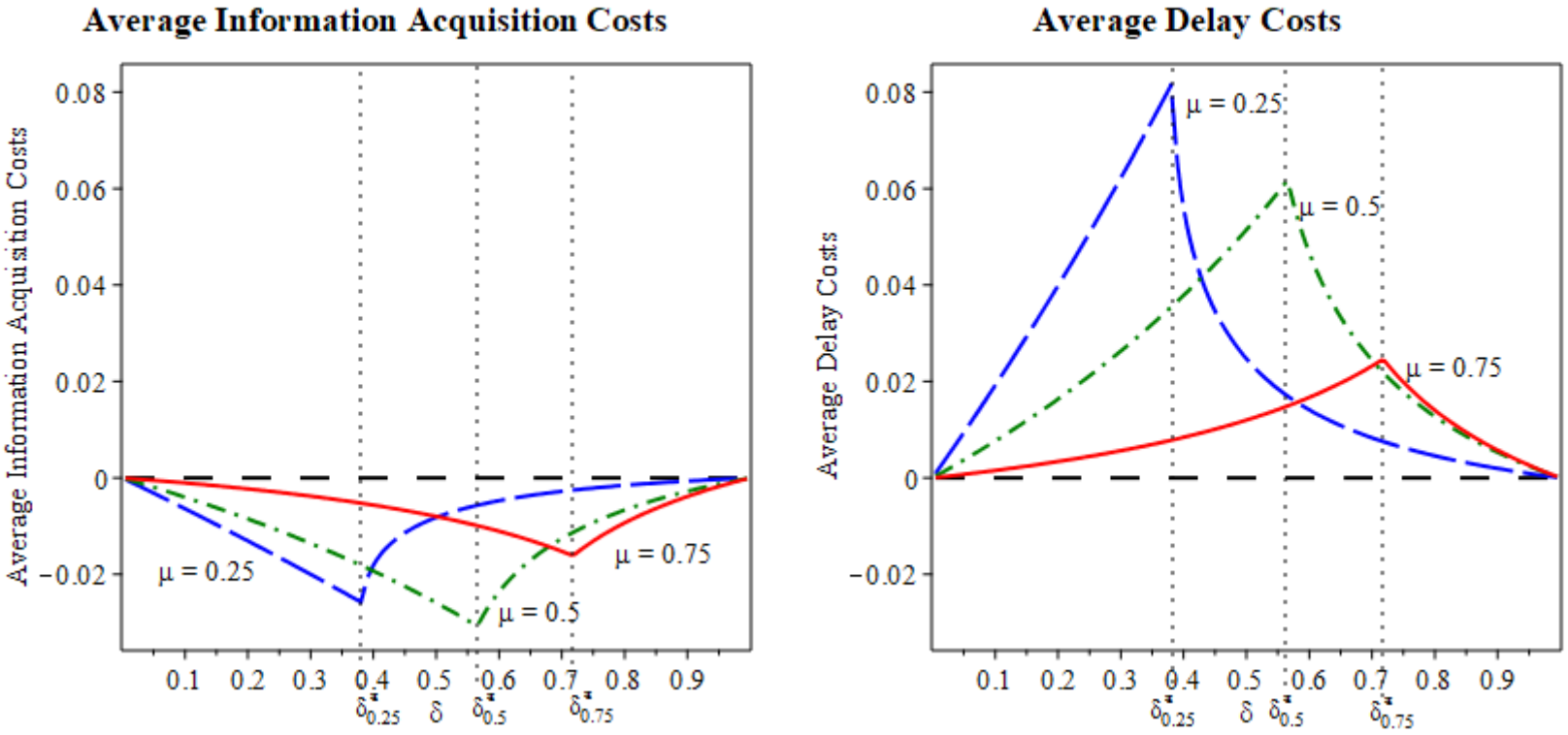
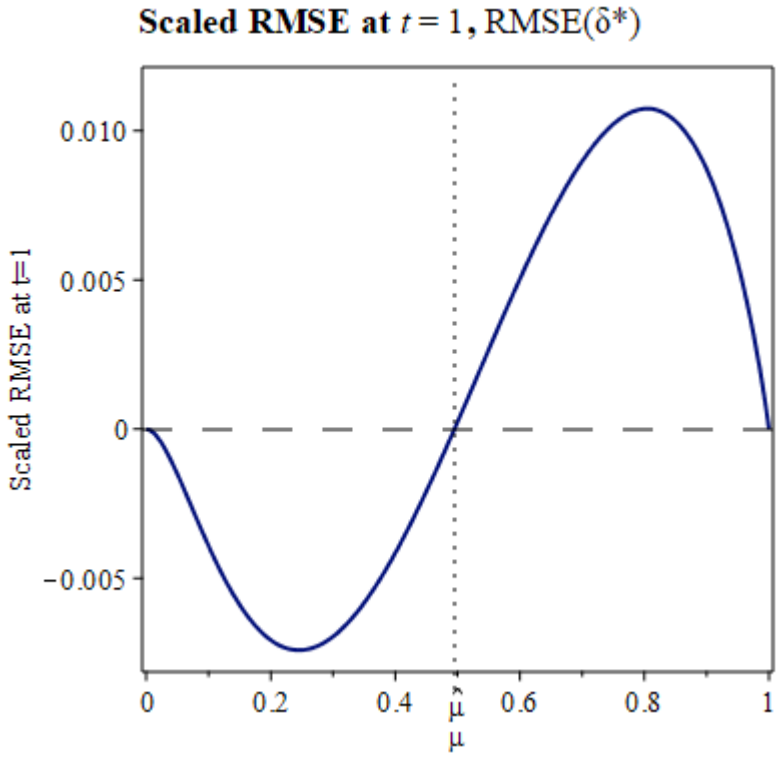
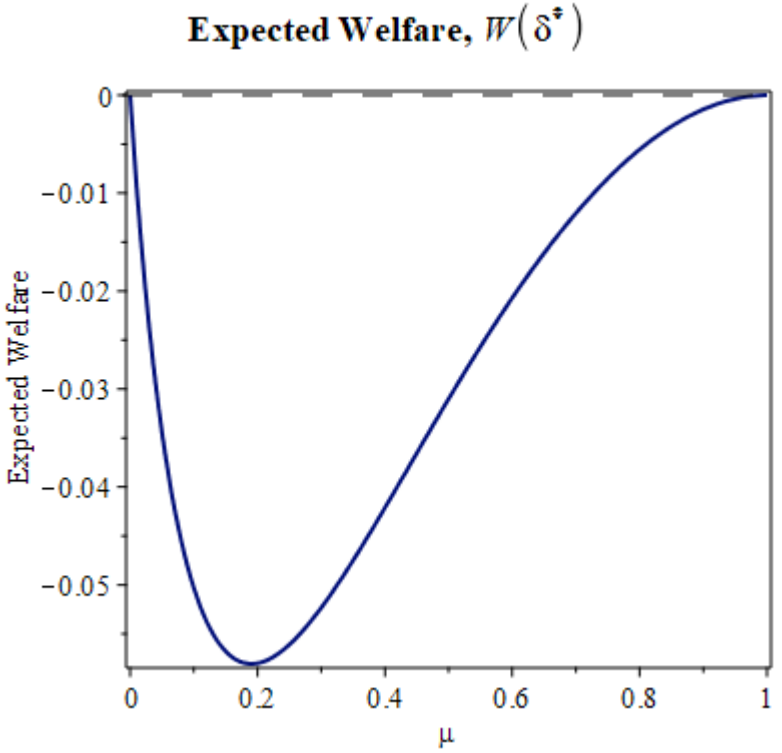


Figure 8: Price Discovery and Expected Welfare at the Segmentation Point

Panels (a) and (b) below provide numerical results on price discovery and expected welfare, respectively, in a fragmented market with a delayed exchange that imposes delay length δ^*, for different levels of speculators, μ. Panel (a) graphs the root-mean-squared proportional pricing error at $t = 1$ (deviation from the true value at $t = 1$) conditional on delay length δ^* centered about the benchmark value RMSE_B, against the measure of speculators, μ. Panel (b) illustrates expected welfare conditional on delay length δ^* centered about the benchmark value W_B, plotted against μ. RMSE_B and W_B are marked by a horizontal wide-spaced dashed line in their respective figures. Thus, negative values for price discovery indicate price discovery improvement, and negative values for expected welfare indicate worsening welfare. A vertical dotted line in panel (a) indicates $\hat{\mu}$. Parameters $k = 3, \sigma = 1$. Results for other values of k and σ are qualitatively similar.



(a)



(b)