High-Frequency Trading and Market Performance

MARKUS BALDAUF and JOSHUA MOLLNER*

January 14, 2020

Journal of Finance, forthcoming

ABSTRACT

We study the consequences of, and potential policy responses to, high-frequency trading (HFT) via the tradeoff between liquidity and information production. Faster speeds facilitate HFT, with consequences for this tradeoff: information production decreases because informed traders have less time to trade before HFTs react, but liquidity (measured by the bid-ask spread) improves because informational asymmetries decline. HFT also pushes outcomes inside the frontier of this tradeoff. However, outcomes can be restored to the frontier by replacing the limit order book with one of two alternative mechanisms: delaying all orders except cancellations or implementing frequent batch auctions.

JEL classification: D47, D82, G14, G18

Keywords: high-frequency trading, order anticipation, quote fade, latency, information production, bid-ask spread, limit order book, non-cancellation delay, asymmetric delay, speed bump, frequent batch auctions

*Baldauf is with the Sauder School of Business, University of British Columbia. Mollner (Corresponding author) is with the Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University. E-mail: joshua.mollner@kellogg.northwestern.edu. First Version: September 18, 2014. We are indebted to our advisors Timothy Bresnahan, Gabriel Carroll, Jonathan Levin, Paul Milgrom, and Monika Piazzesi. We also thank Sandro Ambuehl, Sandeep Baliga, Robert Battalio, Dan Bernhardt, Philip Bond, Eric Budish, Darrell Duffie, Georgy Egorov, Liran Einav, Lorenzo Garlappi, Will Gornall, Joseph Grundfest, Terrence Hendershott, Peter Klibanoff, Fuhito Kojima, John Leahy, Alberto Teguia, Muriel Niederle, Ricardo Perez-Truglia, Matthew Pritsker, Alvin Roth, Ilya Segal, Andrzej Skrzypacz, Alireza Tahbaz-Salehi, Laura Tuttle, Xin Wang, Brian Weller, Glen Weyl, numerous seminar participants, various industry experts, and anonymous referees for valuable comments. We acknowledge financial support by the Kohlhagen Fellowship Fund and the Kapnick Fellowship Program through grants to the Stanford Institute for Economic Policy Research. Baldauf gratefully acknowledges financial support from the Social Science and Humanities Research Council of Canada. Mollner was a Postdoctoral Researcher at Microsoft Research while part of this research was completed, and he thanks them for their hospitality. We have read the *Journal of Finance*'s disclosure policy and have no conflicts of interest to disclose.

Financial markets have been transformed by faster speeds in recent years. For example, the BYX exchange slashed its order processing time sevenfold, from 445 microseconds in 2009 to 64 microseconds in 2018.[1] Likewise, the round-trip communication time between Nasdaq and the Chicago Mercantile Exchange (CME) has nearly halved, from over 14.5 milliseconds in 2010 to 7.9 milliseconds today.[2]

Another important feature of modern trading is that it is highly fragmented, with many stocks now traded at over 30 venues, considerably more than just a decade ago. Given this fragmentation, the faster speeds have increased the effectiveness of high-frequency trading (HFT) strategies, including order anticipation, which we describe below. A stylized fact is that latencies (the lag between when an order is sent to an exchange and when it is processed) are random, which implies that traders cannot ensure simultaneous processing of orders sent to several exchanges. Thus, if HFTs are sufficiently fast, they may observe the trade generated by the first such order to be processed by an exchange and react on the remaining exchanges before the orders of the original trader are processed at those exchanges and in such a way that the original trader fails to trade successfully. This can take two forms: (*i*) traders may cancel their remaining quotes, which we call "passive-side order anticipation," or (*ii*) traders other than the original trader may trade against the remaining quotes, which we call "aggressive-side order anticipation."[3]

In this paper, we present a model of order anticipation by HFTs, and we analyze both its consequences and potential policy responses through the lens of a well-known tradeoff between liquidity and information production. Our three main findings are as follows. First, faster speeds allow HFTs to be more successful at order anticipation, which improves liquidity in the sense of narrowing the bid-ask spread but also lessens information production. Second, order anticipation pushes outcomes inside the frontier of this tradeoff, which represents an inefficiency of HFT, albeit one that differs from the commonly voiced concern that it represents a socially costly arms race. Third, the inefficiency is due to aggressive-side (not passive-side) order anticipation, and hence, certain alternative trading mechanisms can eliminate the inefficiency by preventing aggressive-side order anticipation.

To obtain these results, we build a model that features random latency, multiple ex-

[1] See BATS (2009, 2018).

[2] See "Raging Bulls: How Wall Street Got Addicted to Light-Speed Trading," *Wired Magazine*, August 3, 2012 and Quincy Data (2019).

[3] Internet Appendix Section II reports evidence of order anticipation that comes from a variety of academic and industry sources. It also discusses the sources of randomness in latency and some of the magnitudes in question. The Internet Appendix is available in the online version of the article on the *Journal of Finance* website.

changes, and a single security that is traded by liquidity investors, information investors, and HFTs. Liquidity investors trade for exogenous hedging, saving, or borrowing motives. Information investors may, through costly research, obtain and subsequently trade on private information about the security. HFTs may trade for profit by speculating or by facilitating transactions with other traders. In the baseline, trading is conducted in limit order books (LOBs).

In equilibrium, as in prior literature (e.g., Budish, Cramton, and Shim (2015)), HFTs play two roles. One, the liquidity provider, facilitates trade by posting quotes at all exchanges. The others, snipers, wait to trade until order flow reveals a sufficiently strong signal of the value of the security. As is standard, the liquidity provider faces adverse selection: information investors and snipers trade against her quotes only when the quotes are mispriced. To offset the resulting losses, the liquidity provider must set a bid-ask spread. Also in equilibrium, and similar to, for example, Easley and O'Hara (1987), information investors trade larger quantities than liquidity investors, so that order flow signals the investor's type. Because of random latency, an information investor's orders are not processed simultaneously, which allows HFTs to act on this signal before he completes his trade. The liquidity provider reacts with passive-side order anticipation, that is, by sending cancellations for her remaining quotes after observing one or more trades. Snipers react with aggressive-side order anticipation, that is, by sending orders to trade against the remaining quotes after observing two or more trades. The resulting winner-take-all races may be won by the information investor, the liquidity provider, or a sniper.

Using the comparative statics of the model, we evaluate the consequences of recent improvements in HFT speed. Faster speeds enable HFTs to be more successful at order anticipation, with two primary economic effects. A negative effect is a reduction in information production. Intuitively, order anticipation reduces the amount of rent that informed traders can extract by trading on a piece of information, thereby weakening the incentive to obtain such information. Less fundamental research is then conducted so that markets provide less information about the fundamental value of the security, potentially generating further (unmodeled) distortions in the wider economy. However, a positive effect is an improvement in liquidity as measured by the bid-ask spread. Order anticipation achieves this by reducing the adverse selection faced by liquidity providers through two channels: (*i*) passive-side order anticipation is itself successful avoidance of adverse selection, and (*ii*) through its effect on research, order anticipation reduces the amount of asymmetric information available to create adverse selection. These two predicted consequences are in line with the conclusions

of many empirical studies, and as we discuss below, they suggest a need to reinterpret the conclusions of others.

In sum, when trading is conducted in LOBs, faster HFT speeds induce a tradeoff: information production declines, but the bid-ask spread narrows. Another question that arises is whether these outcomes lie on the frontier of this tradeoff. We show that they generally do not. Again, the reason is order anticipation—more precisely, aggressive-side order anticipation. To see why, a necessary condition for reaching the frontier is that all profits from informed trading accrue to the agents who actually produce the information. Aggressive-side order anticipation violates this condition, as it amounts to snipers profiting from information that they did not produce. We formalize this insight by following the spirit of mechanism design and optimizing over a general class of trading mechanisms so as to characterize the frontier. We also show that this frontier can be achieved by replacing the prevailing LOB with either of two plausible alternatives.

Noncancellation delay mechanisms (NDs) add a small delay between receipt at an exchange and processing for all order types but cancellations. The result is to eliminate aggressive-side order anticipation by allowing the liquidity provider to cancel mispriced quotes before they can be exploited by snipers. Adding randomness to the length of the delay leads to an additional effect: investors become less able to synchronize across exchanges, which amplifies passive-side order anticipation, and leads in turn to a smaller spread and lower research intensity. A set of points on the frontier of the tradeoff can thus be implemented by varying the amount of randomness, including points that dominate the LOB outcome along both dimensions, that is, with a smaller spread and more research.

Frequent batch auctions (FBAs) are uniform-price sealed-bid double auctions conducted at repeated intervals. Following Budish, Cramton, and Shim (2015), we consider batch intervals that are "long" relative to latency and that are synchronized across exchanges. This has an effect opposite that of a randomized ND: investors become *more* able to synchronize across exchanges. Thus, FBAs prevent not only aggressive-side but also passive-side order anticipation. FBAs therefore implement a point on the frontier of the tradeoff, yet one with a larger spread and more intensive research than either the LOB or any ND outcome. This reverses the result of Budish, Cramton, and Shim (2015), who study a model in which information is modeled as exogenous public news and find that FBAs implement a *smaller* spread than the LOB.

The remainder of the paper is organized as follows. Section I discusses related literature. Section II presents the model. Section III describes equilibrium in the baseline in which trad-

ing occurs in LOBs, and it assesses theoretically the consequences of recent improvements in HFT speed. Section IV describes the tradeoff between liquidity and information production, showing that the baseline equilibrium is not on its frontier. Section V characterizes the equilibria prevailing under the two alternative trading mechanisms. Section VI concludes.

I. Related Literature

Our model connects first to the literature on market microstructure, specifically the strand focusing on liquidity and asymmetric information (Glosten and Milgrom (1985), Kyle (1985), Easley and O'Hara (1987), Glosten (1994), Back and Baruch (2004)). Other work demonstrates that public information may give rise to similar forces when trading takes place in LOBs (Foucault (1999), Goettler, Parlour, and Rajan (2009), Budish, Cramton, and Shim (2015), Foucault, Kozhan, and Tham (2017)), with the more recent contributions also highlighting connections to HFT. The aspects of our model that pertain to trading build on some of these papers by embedding communication latency within a multi-exchange financial system. Our model also connects to the literature on information acquisition in financial markets. Central is Grossman and Stiglitz (1980), who study incentives to acquire information and repercussions for informational efficiency. Combining aspects of these two literatures, our paper embeds information acquisition in a realistic model of modern trading. This allows our model to capture the tradeoff between liquidity and information production. What is more, it allows us to study how this tradeoff is influenced by various features of market microstructure, including trading speed and the trading mechanism.

Order anticipation lies at the heart of our model: HFTs attempt to infer an investor's information as he trades, and they respond by canceling their quotes (i.e., passive-side order anticipation) or by trading themselves (i.e., aggressive-side order anticipation). Versions of this behavior have been considered in earlier literature. Indeed, one might interpret the market maker's behavior in Kyle (1985)—adjusting prices in response to information inferred from order flow—as a form of passive-side order anticipation. In subsequent work, Yang and Zhu (forthcoming) add aggressive-side order anticipation to a two-period Kyle (1985) model by including HFTs that are endowed with superior ability (relative to the market maker) to extract information from order flow, which affects how the insider works his order over time. In contrast, our information investors split their orders not over time but rather across exchanges. While this precludes our ability to speak to dynamic considerations, it allows us to capture the microstructure of the trading environment in more detail and thus to formulate

an explicit model of the source of signals that permit aggressive-side order anticipation.

Other models of HFT include those of Biais, Foucault, and Moinas (2015) and Foucault, Hombert, and Roşu (2016), which, like Yang and Zhu (forthcoming), limit HFTs to aggressive trading. Consequently, faster speeds increase adverse selection against liquidity providers, reducing liquidity. In Aït-Sahalia and Sağlam (2017a, 2017b), HFTs are limited to passive trading and thus are liquidity providers. Consequently, faster speeds decrease adverse selection and improve liquidity. In contrast to these models, ours includes both aggressive and passive HFTs, and as a result we capture both of the intuitive effects mentioned above. Interestingly, we nevertheless find that the latter effect dominates, so that faster speeds improve liquidity on net. Other models that contain both aggressive and passive HFTs include Foucault, Kadan, and Kandel (2013), Jovanovic and Menkveld (2016), Foucault, Kozhan, and Tham (2017), Menkveld and Zoican (2017), and Budish, Cramton, and Shim (2015). The details of our model build most heavily on the latter. Our main incremental contribution can be thought of as endogenizing the signals to which HFTs react. In Budish, Cramton, and Shim (2015), these signals are exogenous. In our model, however, they are patterns in order flow that arise endogenously from informed trading in fragmented markets. As we explain in the text, this difference proves important.

Finally, other studies use models of HFT to evaluate alternative trading mechanisms or other interventions. Wah and Wellman (2013), Budish, Cramton, and Shim (2015), Bongaerts and Van Achter (2016), Jovanovic and Menkveld (2016), Rojček and Ziegler (2016), Aït-Sahalia and Sağlam (2017a), Du and Zhu (2017), Brolley and Cimon (2018), Aldrich and Friedman (2019), and Bernales (2019) variously consider Tobin taxes, cancellation fees, minimum resting times, pro-rata matching, batch auctions, and order processing delays.

II. Baseline Model

A single security is traded on multiple exchanges. An investor, who may be either liquidity-motivated or information-motivated, arrives at a random point in time. There are two types of HFTs.

Limit order book. There are $X \geq 1$ exchanges. In the baseline, each exchange operates a separate LOB in which prices are continuous and shares are divisible; below we consider alternative trading mechanisms. Our model of LOB trading is standard, but it is useful to highlight two special order types, both common in practice. An *immediate-or-cancel order*

is a limit order that is automatically cancelled if it does not lead to an immediate trade (i.e., cancelled if it is nonmarketable).[4] In contrast, a *post-only order* is automatically cancelled if it does lead to an immediate trade (i.e., cancelled if marketable).

Orders are processed sequentially in the usual way.[5] If multiple orders arrive at the same time, ties among traders are broken uniformly at random (as by a small amount of random latency). Traders observe orders when they are processed, but trading is anonymous in that traders do not observe identities behind orders they did not themselves submit.

Time. Fix ε to be some positive infinitesimal.[6] The set of time periods is $\mathcal{T} = \{0, \varepsilon, 2\varepsilon, \ldots, 1\}$. See footnote 6 for a discussion of the mathematical tools used in the background to make this construction rigorous. In a period $t \in \mathcal{T}$, events occur in the following order: (*i*) the investor may arrive, (*ii*) previously sent orders are processed, and (*iii*) new orders are sent.

Latency. Latency is the time between when an order is submitted by a trader and when it is processed by the exchange. It is drawn from a distribution with two-point support $\{\varepsilon, 3\varepsilon\}$. Thus, an order sent at $t \in \mathcal{T}$ will be processed at either $t + \varepsilon$ or $t + 3\varepsilon$. We allow (but do not require) the HFTs and the investor to differ in their probabilities of obtaining the shorter latency, letting p_H and p_I denote those probabilities, respectively.[7] We do require $p_H \geq 0.5$ and $p_I \geq 0.5$. On that domain, an increase in the parameter monotonically reduces both the expected value and the variance of latency, and thus can be unambiguously interpreted as an improvement in technology. The same latency applies to all orders sent by a given trader

[4] Market orders can be thought of as immediate-or-cancel orders that specify infinite limit prices. Because there will be latency in our model, quotes can change while orders are en route. For that reason, using a market order is a weakly dominated strategy: an immediate-or-cancel order specifying an appropriately chosen limit price achieves the same end while also guarding against execution at unattractive prices.

[5] If the incoming order is to buy (sell) at a price at or above (below) the ask (bid), then a trade occurs at the ask (bid). The incoming order is referred to as aggressive and the matching order as passive.

[6] An *infinitesimal* ε is a number for which $|\varepsilon| < 1/n$ for all $n \in \mathbb{N}$. The idea is formalized by a branch of mathematics known as nonstandard analysis (Robinson (1966)). Nonstandard analysis is based on the hyperreal numbers ${}^*\mathbb{R}$, an ordered field extension of the real numbers, that contains both infinites and nonzero infinitesimals. Similarly, the hypernatural numbers ${}^*\mathbb{N}$ contain not only the standard natural numbers but also infinites. Formally, our approach is as follows. Fix $N \in {}^*\mathbb{N} \setminus \mathbb{N}$ to be some infinite hypernatural, and define $\varepsilon = 1/N$. This is tantamount to dividing the unit interval into $N \in \mathbb{N}$ discrete time periods, and then letting N diverge to infinity. But rather than work with the sequence, we work directly in the limit. Remark 1 explains the reasons for this approach.

[7] Though not required for our results, it might be natural to assume $p_H > p_I$. Indeed, different traders face different economic problems when determining which speed technologies to adopt (e.g., due to economies of scale or complementarities with other trading activity), which may lead to differing levels of speed.

to a given exchange in a given time period, but latencies are otherwise drawn independently.

Remark 1. Latency is measured in multiples of the infinitesimal ε. This construction approximates the reality of incredibly fast speeds in modern markets wherein latencies—often on the order of microseconds—are negligible relative to the rate at which real economic activity takes place (e.g., the rate at which investors arrive to trade). This is also useful for tractability, in two ways. First, it allows us to deliver an equilibrium that is stationary in the sense that the spread remains constant until a trade occurs, which would not be possible with conventional constructions of time. (We remark on this further in the penultimate paragraph of Section III.B.) Second, it allows us to deliver a clean formalization of certain alternative trading mechanisms by providing a simple language for lengths of time that are both long relative to latency and short in an overall sense. For instance, if one interprets ε as a microsecond, then one might interpret $\sqrt{\varepsilon}$ as a millisecond and use it to model a one-millisecond order processing delay.

Security. A single security has a fundamental value per share v, which is either -1 or 1 (each with equal probability), is chosen by nature prior to time zero, and is initially unknown by all traders. After trading ends, positions are liquidated at v per share.

Investor. An investor arrives at a time drawn from the uniform distribution on $\mathcal{T}$.[8] He is an information investor with probability $\lambda \in (0,1)$ and a liquidity investor otherwise.

If he is an information investor, then immediately upon arrival he chooses a research intensity $r \in [0,1]$ at the cost $c(r)$. This cost function is assumed to be continuously differentiable, weakly increasing, and weakly convex. Research intensity determines his success in obtaining information: he learns v with probability r and fails to learn it otherwise.

If he is a liquidity investor, then he has either a buying motive or a selling motive (each with equal probability), which is modeled as a utility bonus of size β that is earned if he is, respectively, long or short exactly one share of the security when trading ends. We assume that $\beta \geq \frac{\lambda X}{1-\lambda+\lambda X}$, where recall that X denotes the number of exchanges in the economy. As derivations below reveal, this assumption is sufficient to ensure that adverse selection is not so severe that the equilibrium spread exceeds 2β, which would crowd out trade from liquidity investors and lead to market breakdown.

[8] This distribution can be constructed as in Loeb (1975). Moreover, our assumption that the distribution is uniform can be significantly weakened, although we omit the details here.

HFTs. There are two types of HFTs: liquidity providers and snipers. There are at least two liquidity providers and an infinite number of snipers.[9]

Actions. Aside from an information investor's choice of research intensity, the available actions are orders that can be sent. Information investors, liquidity investors, and snipers are restricted to immediate-or-cancel orders. Liquidity providers are restricted to post-only orders (and cancellations thereof). The investor is restricted to sending orders only in the period of his arrival. HFTs are prohibited from sending orders in any two periods that are infinitely close.[10] Information investors and HFTs can send orders to any exchange. A liquidity investor can send orders only to his "home exchange," which is selected uniformly at random from the set of exchanges.

Payoffs. From acquiring a portfolio consisting of y dollars and z shares, an HFT receives utility $y + zv$, a liquidity investor with a buying motive $y + zv + \beta\mathbb{1}\{z = 1\}$, and a liquidity investor with a selling motive $y + zv + \beta\mathbb{1}\{z = -1\}$. An information investor who acquires this same portfolio and chooses research intensity r receives utility $y + zv - c(r)$.

III. Limit Order Book

In this section, we study the baseline model in which each exchange operates a separate LOB. We first describe equilibrium behavior. We then discuss comparative statics.

A. *Equilibrium Description*

Our first result characterizes the LOB equilibrium in terms of two outcome variables: the bid-ask spread and research intensity. The solution concept is weak perfect Bayesian

[9] To formalize and clarify what we mean by "infinite": one should think of our model as the limit of a sequence of models each with a finite number of HFTs. The precise method for taking this limit does not matter for our main analysis, and hence, we do not detail it here. But it does matter for supplemental analysis that we conduct in Internet Appendix Section VII.B, where we discuss some of the most natural possibilities. (As an aside: essentially equivalent to this sequence-based approach, one could again leverage nonstandard analysis so as to feature a hypernatural number of HFTs.) Indeed, there are many HFTs in practice, on the order of hundreds in U.S. markets.

[10] Two times $t, t' \in \mathcal{T}$ are *infinitely close* if $|t - t'|$ is an infinitesimal. Without this restriction, an HFT that sent an order at t might wish to send a redundant order at $t + \varepsilon$ because, given the randomness of latency, the latter order might arrive before the former. Allowing for such behavior would complicate the subsequent mathematics but not change the main forces of the model.

equilibrium (WPBE), where the relevant beliefs are about the value of the security.[11]

PROPOSITION 1: *Under the LOB, there exists a WPBE in which the spread* s^*_{LOB} *and research intensity* r^*_{LOB} *are the unique solution to*

$$s^*_{LOB} = \frac{2\lambda r^*_{LOB}(X_I + X_S)}{1 - \lambda + \lambda r^*_{LOB}(X_I + X_S)} \tag{1}$$

$$r^*_{LOB} \in \arg\max_{r\in[0,1]} \left\{ r X_I \left(1 - \frac{s^*_{LOB}}{2}\right) - c(r) \right\}, \tag{2}$$

where

$$X_I = X p_I + X(1-p_I)^X + (1-p_H)(X-1)X p_I (1-p_I)^{X-1}$$
$$X_S = X(1-p_I) - X(1-p_I)^X - (X-1)X p_I (1-p_I)^{X-1}$$

represent the expected number of trades made by an information investor and snipers, respectively, conditional on an information investor learning v*.*

In terms of (s^*_{LOB}, r^*_{LOB}) as characterized by Proposition 1, equilibrium strategies are:

- *Investor.* If he is a liquidity investor with a buying (selling) motive, he sends to his home exchange an immediate-or-cancel order to buy (sell) one share at the price β $(-\beta)$.

  If he is an information investor, he conducts research with intensity r^*_{LOB}. If he learns that the value of the security is $v = 1$ $(v = -1)$, he sends to each exchange an immediate-or-cancel order to buy (sell) one share at the price 1 (-1). He sends no orders if he does not learn v.

- *Liquidity providers.* One liquidity provider is active on the equilibrium path and is referred to as "the liquidity provider" in what follows. At time zero, she sends to each exchange a post-only order to buy one share at the bid $-s^*_{LOB}/2$ and another to sell one share at the ask $s^*_{LOB}/2$. If at any time t one or more trades occur, she sends cancellations for all of her remaining orders, doing so in the same period.

  A second liquidity provider who is inactive on path but may be active off path is referred to as "the enforcer." If at some time $t \geq 3\varepsilon$ prior to which no trade has occurred the LOB at some exchange consists of anything other than a post-only order to buy one share at $-s^*_{LOB}/2$ and a post-only order to sell one share at $s^*_{LOB}/2$, then she sends such orders to that exchange, doing so in the same period.

[11] A WPBE consists of strategies and beliefs such that (*i*) strategies are optimal given beliefs and (*ii*) beliefs are consistent with Bayesian updating whenever possible.

The remaining liquidity providers remain completely inactive both on and off path.

- *Snipers.* If at any time t trades occur at the ask (bid) at two or more exchanges, each sniper sends to all other exchanges an immediate-or-cancel order to buy (sell) one share at the price 1 (-1), doing so in the same period.

In the paragraphs that follow we sketch why these strategies constitute an equilibrium; we defer a formal treatment to the proof of Proposition 1 in Internet Appendix Section I. While HFTs cannot directly observe whether the investor is information-motivated or liquidity-motivated, a signal of their motivation can be extracted from the pattern of his trades. HFTs monitor these trades, update their beliefs about the investor's motives using Bayes' rule, and react in a way that depends on their type. The liquidity provider attempts to cancel her quotes whenever the expected value of future trades against them is negative. Snipers attempt to trade whenever there is an opportunity to arbitrage quotes against their beliefs.

To describe behavior on path, we consider two cases. First, suppose that the first trades to occur are two or more that take place simultaneously. Because a liquidity investor sends just a single order, this implies that the investor is an information investor, which allows the HFTs to infer the value of the security. Snipers, now knowing that the remaining quotes present an arbitrage opportunity, attempt to trade against them. For the same reason, the liquidity provider sends cancellations for all remaining quotes. In this case, HFTs react to endogenously generated order flow in a way that is analogous to how they react to exogenous public news in Budish, Cramton, and Shim (2015). Second, suppose that the first trade takes place in isolation. This can occur whether the investor is information-motivated or liquidity-motivated, and it does not permit inference about the value of the security that is sufficiently strong to indicate an arbitrage opportunity. As a result, snipers do not react. The liquidity provider, however, does react, sending cancellations for all remaining quotes. This is because a liquidity investor sends just a single order, which implies that any future trades would be information-motivated. Figure 1 illustrates these two cases.

Bertrand competition among liquidity providers leads us to focus on equilibria in which the liquidity provider earns zero profits in expectation.[12] This is ensured by requiring that the equilibrium spread balance the revenue from trades with liquidity investors against the

[12] Although considerable profits may have accrued in the early days of HFT, they were short-lived (see, for example, "High-Frequency Traders Fall on Hard Times," *Wall Street Journal*, March 21, 2017). Furthermore, much of the rest of the literature also assumes competitive liquidity provision (e.g., Glosten and Milgrom (1985), Kyle (1985)).

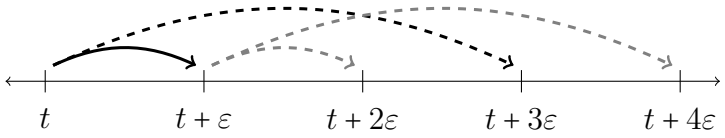


Figure 1. Illustration of equilibrium (LOB). One equilibrium scenario: (*i*) the investor sends orders at $t \in \mathcal{T}$, either to either one or all exchanges depending on his type, and (*ii*) at least one order receives the shorter latency and is processed at $t+\varepsilon$ (solid black arrow), while any remaining orders receive the longer latency and are processed at $t+3\varepsilon$ (dashed black arrow). HFTs react to the trade(s) at $t+\varepsilon$ by sending orders that may be processed at $t+2\varepsilon$ or $t+4\varepsilon$ (dashed gray arrows). With a single trade at $t+\varepsilon$, only the liquidity provider reacts. With two or more trades at $t+\varepsilon$, snipers react as well.

costs of adverse selection exerted by information investors and snipers. When a liquidity investor is present, which occurs with probability $1-\lambda$, the liquidity provider earns the half-spread $s^*_{LOB}/2$ on the one trade that takes place. In contrast, when an information investor is present and learns the value of the security, which occurs with probability λr^*_{LOB}, the liquidity provider loses $1-s^*_{LOB}/2$ on each of the X_I+X_S trades that take place. The zero-profit condition is thus

$$(1-\lambda)\frac{s^*_{LOB}}{2} = \lambda r^*_{LOB}(X_I+X_S)\left(1-\frac{s^*_{LOB}}{2}\right).$$

Solving for the spread yields equation (1) in the proposition. The liquidity provider cannot profitably deviate from this spread: narrower quotes would yield negative expected profits, while wider quotes would be undercut by the enforcer.

Equation (2) in the proposition ensures that information investors conduct research with an intensity that optimally balances costs and benefits. The cost of implementing research intensity r is $c(r)$. The corresponding benefit is the product of (*i*) r, the probability of learning the value of the security, (*ii*) X_I, the number of trades an information investor can expect to complete, and (*iii*) $1-s^*_{LOB}/2$, his profit per trade, that is, the magnitude of his informational advantage minus the half spread.

B. Remarks

Although each liquidity investor desires to trade just a single share, aggregate quoted depth exceeds that amount in equilibrium. The reason relates to our assumption that each

liquidity investor trades only at his home exchange, which we interpret as an extreme way of capturing frictions that affect order routing decisions. In practice, there may be many sources of such frictions, for instance, broker-client conflicts of interest or technological barriers to observing prices in real time. Such frictions fragment the collective order flow of liquidity investors. In response, liquidity providers must quote more aggregate depth. In the model, depth in fact scales linearly in the number of exchanges because the liquidity provider optimally offers one share at both the bid and the ask at each exchange to serve a liquidity investor who might attempt to trade there.[13] Note that fractional shares could be quoted—although liquidity investor demand is not divisible, shares are.

In contrast to liquidity investors, information investors do wish to trade against the entire aggregate depth. For this reason, a larger number of exchanges being targeted is a signal of informationally-motivated trade.[14] This is what allows HFTs to extract information from order flow. What allows them to *use* that information is random latency. While an information investor sends all orders in a wave simultaneously at a time t, random latency may create dispersion in their processing times at the different exchanges, with some orders being processed at $t + \varepsilon$ and others at $t + 3\varepsilon$. It is this dispersion that permits order anticipation. Having observed trades at some exchanges at $t + \varepsilon$, HFTs react by sending orders to the remaining exchanges, which may be processed at $t + 2\varepsilon$, before the information investor's remaining orders are processed at $t + 3\varepsilon$. Passive-side order anticipation occurs when the liquidity provider, reacting in this way, cancels a quote before the information investor can trade against it. Aggressive-side order anticipation occurs when a sniper removes such a quote by trading against it herself.

Smart order routers, with RBC's THOR (Aisen et al. (2015)) as a prominent example, attempt to mitigate the effects of order anticipation by releasing orders in a wave at slightly different times, giving those sent to high-latency exchanges a head start. In our model, however, information investors face the same latency distribution at each exchange, as if

[13] That adding a trading venue deepens the aggregate book is corroborated by empirical evidence (e.g., Boehmer and Boehmer (2003), Fink, Fink, and Weston (2006), Foucault and Menkveld (2008), He, Jarnecic, and Liu (2015), Aitken, Chen, and Foley (2017)). On the theoretical side, Dennert (1993) also features this force.

[14] A similar property is present in many other models (Kyle (1985), Easley and O'Hara (1987), Glosten (1989, 1994), Biais, Martimort, and Rochet (2000)), where it either emerges or is assumed that larger volumes signal more information about the value of the security. The aforementioned property of our model differs only in that ours is a multi-exchange setting in which volume is measured in effect by the number of exchanges targeted. Consistent with this property of our model, empirical evidence shows that large trades tend to be more informed than their smaller counterparts (e.g., Hasbrouck (1991), Lin, Sanger, and Booth (1995), Easley, Kiefer, and O'Hara (1997), Chakravarty et al. (2012)).

they were already employing such an algorithm.

Equilibrium is stationary in the sense that the spread remains constant until a trade occurs, at which point a burst of activity occurs whose nature does not depend on calendar time. Important for delivering this type of tractability is the fact that the distribution governing the investor's arrival time is infinitely more diffuse than the distribution governing latency. This is one way nonstandard analysis helps deliver a tractable model. In analogous models with more conventional timing, the scales of these two distributions necessarily differ by only a finite multiple. In consequence, such models are instead plagued by technicalities that do not correspond to important economic forces. See Internet Appendix Section III.G for details.[15]

Proposition 1 does not assert the uniqueness of equilibrium (only existence), but we do not view this as problematic for two reasons. First, the selected equilibrium has many features that one might reasonably expect (e.g., symmetry across exchanges, zero profits for liquidity providers, etc.). Moreover, we argue in Internet Appendix Section III.J that all equilibria with these features give rise to the same outcome (s^*_{LOB}, r^*_{LOB}). Second, when we compare the LOB to alternative mechanisms in Section V, we ensure that we do so in a consistent way by making the same equilibrium selection. That is, equilibrium strategies remain fixed in all respects except the spread and research intensity, and the different mechanisms have an effect only by influencing how those fixed strategies map into trading outcomes.

C. *Effect of Faster Speed*

Our first main insight concerns the effect of faster HFT technologies. To derive it, we study how the equilibrium of Proposition 1 varies with the parameter p_H.[16] In examining the comparative statics, we again focus on two outcome variables, namely, liquidity, as measured by the bid-ask spread, and information production, as measured by research intensity.

COROLLARY 1: *Both the spread s^*_{LOB} and research intensity r^*_{LOB} are weakly decreasing in p_H.*

[15] In short, the issue is the following. Because a liquidity investor sends a single order while an information investor sends multiple orders at once, an information investor's minimum latency (the minimum taken across the orders he sends) is first-order stochastically dominated by a liquidity investor's latency. If the scale of the latency distribution is not infinitesimal, then calendar time becomes informative about the next order to arrive (its likelihood of being liquidity-motivated relative to information-motivated) and as a result the spread varies with time.

[16] See Internet Appendix Section III.A for comparative statics with respect to the remaining parameters (i.e., p_I, λ, X).

According to Corollary 1, an increase in p_H (i.e., faster speed) reduces equilibrium research intensity. Intuitively, faster HFTs are more adept at order anticipation, which reduces the fills that information investors receive. This disincentivizes research and leads to less of it in equilibrium.[17] This empirical implication—that faster HFT technologies lead to less fundamental research and hence less informative prices—is in line with the findings of Weller (2018), Lee and Watts (2018), and Gider, Schmickler, and Westheide (2019). In contrast, several other papers argue that HFT makes prices *more* informative, but this finding can be explained by use of an incomplete notion of informativeness. For example, Hendershott, Jones, and Menkveld (2011) study an upgrade that resulted in faster HFT and find price changes to be less correlated with trades thereafter. This result is consistent with faster speeds leading to more passive-side order anticipation, and it suggests that information is incorporated into prices faster *conditional on becoming available*. But this evidence says little about the extent of fundamental research or the probability of information becoming available in the first place.[18]

An increase in p_H also reduces the equilibrium spread. The liquidity provider quotes a smaller spread when HFTs are faster because she faces less adverse selection. This occurs for two reasons. First, as described above, faster speeds reduce research intensity and in turn the amount of information that is ultimately traded upon. Second, faster speeds facilitate passive-side order anticipation, which enables the liquidity provider to cancel more mispriced quotes before they are exploited. This empirical implication—that faster HFT technologies lead to smaller spreads—is in line with the majority of the evidence on this topic.[19]

This comparative static also represents an interesting contrast to Budish, Cramton, and Shim (2015), who find that HFT speed has no effect on the spread. In their model, HFTs react to exogenous signals. The resulting race to react is only among themselves, and hence its nature is not affected by their absolute speed. In our model, however, HFTs react to endogenous patterns in order flow. The resulting race to react also includes the information

[17] That information leakage disincentivizes research is also observed by the literature on dual trading (Fishman and Longstaff (1992), Röell (1990)). More broadly, that incentives to acquire information are related to the rents that can be derived from it is illustrated by many models (e.g., Grossman and Stiglitz (1980)).

[18] Other empirical work similarly argues that HFT improves price informativeness (Brogaard (2010), Carrion (2013), Brogaard, Hendershott, and Riordan (2014), Chaboud et al. (2014), Conrad, Wahal, and Xiang (2015), Boehmer, Fong, and Wu (2018)).

[19] Empirical evidence that enhanced HFT reduces the spread (or generally improves liquidity) is provided by Brogaard (2010), Hendershott, Jones, and Menkveld (2011), Hasbrouck and Saar (2013), Menkveld (2013), Frino, Mollica, and Webb (2014), Brogaard et al. (2015), Conrad, Wahal, and Xiang (2015), Boehmer, Fong, and Wu (2018), and Malinova, Park, and Riordan (2018).

investor whose orders triggered the race. Thus, faster HFT speeds do affect the nature of the race, reducing adverse selection by intensifying the speed disadvantage of the information investor.

IV. Feasible Outcomes

Liquidity and information production—or in our context, a small spread and high research intensity—are desirable properties of financial markets. But these two objectives often conflict: to incentivize information production, agents must be able to trade on the information they produce, which then typically exacerbates adverse selection and worsens liquidity. The comparative statics of the previous section illustrate this conflict. According to Corollary 1, when HFTs become faster, liquidity improves but information production declines.

Given the conflict between these dual objectives of liquidity and information production, it is useful to consider what is feasible in their two dimensions. Importantly, reaching the frontier of this feasible set requires that all profits from informed trading flow to the agents who actually produce the information. Our next main insight is that this necessary condition does not hold in the LOB equilibrium for the reason that aggressive-side order anticipation enables snipers to profit from information they did not produce. In this section, we follow an approach that is in the spirit of mechanism design to formalize this insight and to provide a framework for evaluating alternatives to the LOB.

A. Framework

We use two criteria to evaluate a trading mechanism—liquidity and information production. As before, information production is measured by research intensity.[20] In previous sections, liquidity is measured by the bid-ask spread. Here, we wish to consider general mechanisms, including some that do not feature a spread. We therefore employ what amounts to a change of variables and take liquidity investor welfare (or more precisely, the investor's expected utility conditional on being a liquidity investor) to be our measure. For our model of the LOB, liquidity investor welfare is equivalent to the spread through $w^*_{LOB} = \beta - s^*_{LOB}/2$.[21]

[20] Although informative prices are not valuable to the traders in our model, the literature has identified several channels through which they may be a positive externality for nontraders (see Internet Appendix Section V).

[21] Note that with a richer model of liquidity investors, for example, if they were risk averse and traded to hedge their endowment shocks, a reduction in liquidity would crowd out their trading and reduce allocative

(Analogous equalities hold for the mechanisms that we analyze in Section V.)

Of course, other potential criteria could be used, for example, those that consider the welfare of HFTs or information investors. We focus on the two criteria above, however, because of their central position in the literature and because they are the features of our model that relate most closely to the stated objectives of most regulatory bodies.[22]

We next characterize what is feasible in terms of these two criteria. To do so, we imagine the problem of a social planner who recommends a research intensity to the investor and also allocates resources (dollars and shares). The planner does not observe research directly, and thus to incentivize it must make the allocation a function of the state, subject to several constraints that we define formally in the next section. This framework nests optimization over a wide range of trading mechanisms, including the LOB and the alternatives considered in Section V.

B. Formalities

Potential investor characteristics define five states: (*i*) a liquidity investor with a buying motive, (*ii*) a liquidity investor with a selling motive, (*iii*) an information investor who learns $v = 1$, (*iv*) an information investor who learns $v = -1$, or (*v*) an information investor who fails to learn v. We denote these five states $\Theta = \{B, S, 1, -1, 0\}$, and we use θ for a typical element. The probability distribution over Θ is affected by the choice of r: $\mathbb{P}(B) = \mathbb{P}(S) = (1-\lambda)/2$, $\mathbb{P}(1) = \mathbb{P}(-1) = \lambda r/2$, and $\mathbb{P}(0) = \lambda(1-r)$.

Letting $v(\theta)$ denote the expected value of the security conditional on available information, we have $v(B) = v(S) = v(0) = 0$, $v(1) = 1$, and $v(-1) = -1$. Let $\mathcal{H}$ denote the set of HFTs (including both liquidity providers and snipers). The expected utility that HFT $h \in \mathcal{H}$ receives from a portfolio consisting of y dollars and z shares in state θ is $u_h(y, z|\theta) = y + zv(\theta)$. Similarly, the expected utility, gross of research costs, that the investor receives from a portfolio consisting of y dollars and z shares in state θ is

$$u(y, z|\theta) = \begin{cases} y + zv(\theta) + \beta \mathbb{1}\{z = 1\} & \text{if } \theta = B \\ y + zv(\theta) + \beta \mathbb{1}\{z = -1\} & \text{if } \theta = S \\ y + zv(\theta) & \text{if } \theta \in \{1, -1, 0\}. \end{cases}$$

efficiency. This may provide an additional reason to prioritize liquidity.

[22] Nevertheless, in Internet Appendix Section VII we argue that our key results would extend if the liquidity investor welfare criterion were replaced by either (*i*) total investor welfare or (*ii*) total trader welfare.

We consider contracts among the planner, the investor, and HFTs. A contract specifies a research intensity as well as payments to the traders as functions of the state, which may be in the form of dollars, shares, or both. Let $y(\theta)$ and $z(\theta)$ denote, respectively, the number of dollars and shares paid to the investor in state θ under such a contract. Let $y_h(\theta)$ and $z_h(\theta)$ denote the corresponding quantities for HFT $h \in \mathcal{H}$. The following analysis treats these quantities as deterministic, but allowing for randomization would not change any of the results.

The *outcome* of a contract is determined by two criteria: (*i*) research intensity, denoted by r, and (*ii*) liquidity investor welfare, denoted by w. The *feasible set*, $\mathcal{F}$, consists of outcomes that can be implemented by contracts satisfying certain constraints:

$$\mathcal{F} = \left\{ (r,w) \;\middle|\; \begin{array}{l} \exists y(\theta), \exists z(\theta), \exists \{y_h(\theta)\}_{h\in\mathcal{H}}, \exists \{z_h(\theta)\}_{h\in\mathcal{H}} \text{ such that} \\ \text{(W), (BB–1), (BB–2), (IR–I), (IR–H), (O)} \end{array} \right\},$$

where

$$\begin{aligned}
&\text{(W)} && w = \tfrac{1}{2}u(y(B),z(B)|B) + \tfrac{1}{2}u(y(S),z(S)|S) \\
&\text{(BB–1)} && (\forall \theta \in \Theta) : y(\theta) + \textstyle\sum_{h\in\mathcal{H}} y_h(\theta) = 0 \\
&\text{(BB–2)} && (\forall \theta \in \Theta) : z(\theta) + \textstyle\sum_{h\in\mathcal{H}} z_h(\theta) = 0 \\
&\text{(IR–I)} && (\forall \theta \in \Theta) : u(y(\theta),z(\theta)|\theta) \geq 0 \\
&\text{(IR–H)} && (\forall h \in \mathcal{H}) : \mathbb{E}_r[u_h(y_h(\theta),z_h(\theta)|\theta)] \geq 0 \\
&\text{(O)} && r \in \arg\max_{\hat{r}\in[0,1]} \left[\tfrac{\hat{r}}{2}u(y(1),z(1)|1) + \tfrac{\hat{r}}{2}u(y(-1),z(-1)|-1) + (1-\hat{r})u(y(0),z(0)|0) - c(\hat{r}) \right].
\end{aligned}$$

The constraint (W) is definitional: it requires that w represent the investor's expected utility conditional on being a liquidity investor. Constraints (BB–1) and (BB–2) require budget balance with respect to dollars and shares, respectively. Constraint (IR–I) requires individual rationality for the investor at the interim stage, that is, after learning his type but before the resolution of any residual uncertainty about v. Constraint (IR–H) requires individual rationality for HFTs at the ex ante stage. Finally, constraint (O) requires that the investor's choice of research intensity be optimal.[23]

[23]Note that this formulation does not impose a truthfulness constraint to require that the investor find it optimal to report θ. In that sense, we are analyzing a relaxed problem that involves fewer constraints than there "should" be. Nevertheless, because the alternative trading mechanisms that we discuss in Section V achieve the frontier of even this relaxed problem, the absence of this constraint is not relevant to our key results about the frontier of $\mathcal{F}$ (see Corollaries 2, 5, and 6).

C. *The LOB Does Not Achieve the Frontier of $\mathcal{F}$*

The set of feasible outcomes is characterized by the following proposition.

PROPOSITION 2: *The feasible set is*

$$\mathcal{F} = \left\{ (r, w) \;\middle|\; r \in [0,1], w \in \left[0, \beta - \frac{\lambda}{1-\lambda} r c'(r) \right] \right\}.$$

In particular, an outcome $(r, w) \geq (0, 0)$ is on the frontier of $\mathcal{F}$ if and only if

$$(1-\lambda)(\beta - w) = \lambda r c'(r). \tag{3}$$

Intuitively, the constraints of the problem can be combined in such a way that they come down to a single tradeoff: information investors can be incentivized to conduct research, but only if they are paid with funds raised through a tax on liquidity investors. The left-hand side of (3) is the share of liquidity investors multiplied by the per-capita liquidity investor tax. The right-hand side is the share of information investors multiplied by $rc'(r)$, where the latter represents the minimum expected payment required to incentivize an information investor to research with intensity r.[24] To be on the frontier, it must be the case that liquidity investors are not taxed beyond the minimum necessary to incentivize the desired level of research.

While Proposition 2 characterizes the feasible set, it does not speak to the particular point in the set that a social planner would optimally implement. Nevertheless, a necessary condition for optimality is that the outcome be on the frontier of the set. A corollary of Propositions 1 and 2 is that the LOB equilibrium generally fails this requirement.

COROLLARY 2: *If $X > 2$, $p_I < 1$, and $c'(0) < X_I$, then the LOB outcome is not on the frontier of $\mathcal{F}$.*

To be on the frontier, all information rents from the trading stage must accrue to those who actually produce information. Importantly, this is not the case when aggressive-side order anticipation occurs—instead, snipers divert a portion of the rents. In doing so, snipers contribute to adverse selection, raising the spread but without producing any information. In short, aggressive-side order anticipation is a wedge that can push the outcome inside the frontier of the tradeoff between liquidity and information production.

[24] An incentive scheme that achieves this minimum is to pay $c'(r)$ conditional on information being produced. Because information is produced with probability r, the expected payment is $rc'(r)$.

Under the conditions of Corollary 2, the LOB equilibrium involves aggressive-side order anticipation, and as a result its outcome is off the frontier. To understand those conditions, note that aggressive-side order anticipation does not take place if there are two or fewer exchanges or if the investor's latency is deterministic. It also does not take place if research is too costly to occur in equilibrium, which $c'(0) < X_I$ rules out.

Although aggressive-side and passive-side order anticipation have much in common, the analysis above highlights an important difference between the two HFT strategies. Specifically, while aggressive-side order anticipation moves the outcome *off* the frontier of the feasible set, passive-side order anticipation moves the outcome *along* the frontier.[25] As a result, passive-side order anticipation may actually be beneficial, at least if sufficient emphasis is placed on liquidity. In contrast, our findings highlight a sense in which aggressive-side order anticipation is an unambiguously inefficient form of HFT.

D. Robustness

We now turn to the question of robustness. We argue that our main results are not driven by our occasionally stylized modeling choices. As the previous analysis shows, the key economic insight is that the LOB is not on the frontier of the tradeoff between liquidity and information production because it permits aggressive-side order anticipation, through which some information rents are obtained by traders who do not produce information. Accordingly, in the following discussion we focus on the extent to which aggressive-side order anticipation persists in less stylized settings. We conclude that aggressive-side order anticipation is a robust feature of LOB trading.

One limitation of the model is that it does not allow us to consider how informed traders might split orders over time. In the model, at most a single liquidity investor is present. Because an information investor cannot therefore pretend to be a stream of liquidity investors, he is effectively limited to trading at just a single point in time. He therefore trades as intensely as possible at that moment and optimally synchronizes his trading in one large "wave." For private information that is short-lived, this modeling approach seems realistic. But for private information that is long-lived, which is the case in which information acquisition has the greatest social benefit, it may be less than fully realistic. A more realistic model would be one in which a stream of liquidity investors arrive. In such a model, there would indeed be scope for dynamic order splitting and, potentially, less order anticipation

[25] Similarly, Lyle and Naughton (2016) empirically distinguish between "liquidity provider monitoring" by HFTs, which lowers the spread, and residual HFT, which raises it.

would occur. Nevertheless, in Internet Appendix Section III.F where we consider such a model—a multi-exchange version of Back and Baruch (2004)—we show that if the market is sufficiently fragmented, then equilibrium requires that the informed trader sometimes trade in a large wave even though doing so would give himself away and lead to order anticipation. Because modern markets are indeed quite fragmented, this seems likely to be the relevant case.

It is also stylized to assume that liquidity investors do not compare prices when selecting an exchange. As we discuss in Section III.B, this assumption captures a realistic economic force—frictions that affect order routing decisions—albeit in an extreme way. This assumption is not uncommon in the literature on trading in fragmented markets (e.g., Mendelson (1987), Chowdhry and Nanda (1991), Biais et al. (2015), Foucault et al. (2017)). Indeed, as mentioned above, this assumption is important because it delivers the key property that a larger number of exchanges targeted is a signal of informationally motivated trade. However, this property would arise even if frictions that affect order routing were captured in a less extreme way by allowing for a limited amount of cross-exchange elasticity (as in, for example, Baldauf and Mollner (forthcoming)). Depending on modeling specifics, this might weaken the extent to which trading motives are signaled by the number of exchanges targeted. Nevertheless, some signal, and therefore some order anticipation, would remain so long as information investors continue to target more exchanges than liquidity investors in the sense of first-order stochastic dominance. Similar conclusions would apply if we relaxed the assumption that liquidity investors are homogeneous in the quantity they demand (see Internet Appendix Section III.H).

Our assumption of risk neutrality is also not without loss. For example, if information investors were risk averse, then they might trade less intensely on information, sending orders to only a subset of exchanges. Nevertheless, as long as risk aversion is not too extreme, the trading demands of information investors would continue to dominate those of liquidity investors. In turn, some signal would remain in order flow, and order anticipation would continue to occur (see Internet Appendix Section III.I).

Another limitation of the model is that it imposes many restrictions on order types. Information investors, liquidity investors, and snipers are restricted to immediate-or-cancel orders, which limits them to aggressive trading. In contrast, liquidity providers are restricted to post-only orders, which limits them to passive trading. Many of these restrictions are fairly standard in the literature (e.g., Glosten and Milgrom (1985), Kyle (1985), Budish,

Cramton, and Shim (2015)).[26] Moreover, they allow our analysis to be clean and tractable by ensuring that no trader has an opportunity to alternate between aggressive and passive trading.[27] Nevertheless, they are unlikely to be driving our conclusions.[28] In particular, given an information investor's need to synchronize execution across venues, aggressive behavior would likely emerge endogenously, which would again open the door for order anticipation.

It is quite stylized to model latency as draws from a two-point distribution, but this also does not drive our results. For order anticipation to occur in the LOB equilibrium, the necessary property is that the maximum investor latency exceeds the sum of his minimum latency and the minimum HFT latency. We use a simple and convenient distribution with that property. We also assume that latencies are drawn independently. While some correlation likely exists in practice, the necessary condition for order anticipation to occur is only that such correlation is not perfect across either traders or exchanges.

Another limitation is that latencies are exogenous. Endogenizing them (as in, for example, Budish, Cramton, and Shim (2015), Foucault, Kozhan, and Tham (2017)) might capture additional forces. Depending on the specifics, these forces might attenuate or even overturn the comparative statics of Corollary 1, as a change in HFT speed would alter the marginal return on speed investments for investors. But so long as investors do not achieve perfect synchronization (i.e., $p_I < 1$), which seems to be the realistic case (see Internet Appendix Section II.B), order anticipation occurs and the LOB outcome lies inside the frontier (see Corollary 2). Our conclusion that the alternative mechanisms studied in Section V implement outcomes on the frontier also continues to hold.

V. Alternative Trading Mechanisms

As the previous section highlights, the key to improving upon the LOB is to eliminate aggressive-side order anticipation. Motivated by this observation, we consider two realistic alternative trading mechanisms that achieve this. (Conversely, we do not consider other

[26] Some exceptions are Parlour (1998), Foucault (1999), Foucault, Kadan, and Kandel (2005), and Roşu (2009), who develop models of LOB trading in which agents endogenously choose whether to trade aggressively or passively. However, none of these models features asymmetric information.

[27] This is why traders are not given access to "plain vanilla" limit orders in our model, that is, those can be used to trade either passively or aggressively.

[28] Note further that neither type of HFT earns positive profits in any of the equilibria we identify. Accordingly, it would make no difference if the model were augmented with an initial stage in which HFTs were to choose their type endogenously. What is important for our analysis is that no HFT may use a strategy that could lead to an outcome in which she makes both aggressive and passive trades.

familiar alternatives, such as minimum resting times, that do not have this effect.) We find that these alternatives are optimal in the model in the sense that they implement outcomes on the frontier of the tradeoff between liquidity and information production.

A. Noncancellation Delay Mechanisms

We first consider a family of mechanisms in which exchanges process cancellations upon receipt but other order types only after a small (possibly random) delay. As a result, a cancellation received slightly after an order of a different type might be processed first (whereas LOBs process orders strictly in the order received). The effect is to eliminate aggressive-side order anticipation, but not its passive-side counterpart.

In the formulation below, all limit orders (including both immediate-or-cancel and post-only orders) receive a delay. In practice, exchanges provide an even wider variety of order types, and accordingly there are many degrees of freedom concerning exactly which orders to delay. In recent years, versions of this same basic economic mechanism have been proposed or implemented by several venues across multiple asset classes (see Internet Appendix Section VI.A), but these developments have been controversial. We hope to contribute some clarity to this issue by providing a formal model-based analysis of these mechanisms.

A.1. Definition

We focus our formal analysis on a specific family of ND mechanisms, parametrized by $q \in [0,1]$, where the length of the delay is chosen as follows. All noncancellation orders receive a delay of constant length δ_{ND}. With probability q, a noncancellation order receives an additional delay, which is a random variable drawn from some distribution F_{ND}.[29] Aside from these delays, all order processing is as in the LOB.[30] In particular, orders are observed immediately upon processing (but remain unobserved until they are processed, even if they have already been received by the exchange). In addition, we retain all previous restrictions on orders that the various traders can submit.

To have the desired effect, δ_{ND} should be small but should exceed the maximum difference in reaction time that may occur between two HFTs responding to the same event. In the language of the paper, this requirement corresponds to an infinitesimal delay $\delta_{ND} > 2\varepsilon$, which

[29]This is simply one way to parametrize the addition of randomness drawn from F_{ND} to the fixed delay δ_{ND}. Analogues of the results derived here would exist for any other smooth parametrization.

[30]Note that with such a delay, it becomes possible for a cancellation order to be processed *before* the order it was meant to cancel. In that case, the latter order would be cancelled immediately upon processing.

we assume hereafter. Furthermore, we assume that the distribution F_{ND} (i) puts positive probability only on infinitesimal values $t \in \mathcal{T}$ and (ii) puts infinitesimal probability on any particular t. Roughly speaking, the variance of F_{ND} should be "one order of magnitude larger" than latency. We also impose the following correlation structure on the random component of the delay: draws are identical for orders sent by the same trader to the same exchange at the same time and otherwise are independent.

A.2. Equilibrium

We next characterize the ND equilibria in terms of the spread and research intensity.

PROPOSITION 3: *For all $q \in [0,1]$, under qND, there exists a WPBE in which the spread s^*_{qND} and research intensity r^*_{qND} are the unique solution to*

$$s^*_{qND} = \frac{2\lambda r^*_{qND} X_{qND}}{1 - \lambda + \lambda r^*_{qND} X_{qND}} \tag{4}$$

$$r^*_{qND} \in \arg\max_{r \in [0,1]} \left\{ r X_{qND} \left(1 - \frac{s^*_{qND}}{2} \right) - c(r) \right\}, \tag{5}$$

where $X_{qND} = q^X + \sum_{x=1}^{X} \binom{X}{x} (1-q)^x q^{X-x} \left(x p_H (1-p_I)^x + x \left[1 - p_H (1-p_I) \right] \right)$ represents the expected number of trades made by an information investor conditional on learning v.

The strategies that support this outcome in equilibrium are precisely as before (see Section III.A). As mentioned, this provides us with some confidence that our analysis constitutes a consistent comparison of the different trading mechanisms.

As we explain below, trading outcomes change in two ways relative to the LOB baseline: (i) in terms of the number of trades made by information investors, with X_I now mapping into X_{qND}, and (ii) in terms of the number of trades made by snipers, with X_S now mapping into zero. Given this, the equilibrium spread and research intensity are pinned down precisely as before. First, the liquidity provider earns zero profits in expectation. Revenue from liquidity investors must be balanced by the costs of adverse selection, which comes now only from information investors and not from snipers. The zero-profit condition is therefore $(1-\lambda)\frac{s^*_{qND}}{2} = \lambda r^*_{qND} X_{qND} \left(1 - \frac{s^*_{qND}}{2} \right)$, which yields equation (4). Second, information investors must choose research intensity optimally, which yields equation (5).

A key difference relative to the baseline is that now snipers do not trade. Because delays are applied to the orders sent by snipers but not to the cancellations sent by liquidity providers, all mispriced quotes are cancelled before snipers can trade against them. Thus,

NDs eliminate aggressive-side order anticipation, which, as we have argued, is the key to rooting out the main friction in the model.

The deterministic component of the delay, δ_{ND}, by itself eliminates aggressive-side order anticipation. However, adding randomness to the length of the delay has an additional effect: more passive-side order anticipation occurs. To illustrate, consider X_{qND}, the expected number of fills obtained by an informed information investor. If $q = 1$, so that the delay is always random in length, then the information investor's orders are processed with such dispersion that he obtains just a single fill and $X_{1ND} = 1$. After observing this first trade, the liquidity provider sends cancellations for the remaining mispriced quotes, which are processed almost surely before the information investor's remaining orders. If $q = 0$, so that the delay is always fixed in length, then the information investor expects to obtain $X_{0ND} = Xp_H(1-p_I)^X + X[1-p_H(1-p_I)]$ fills. To see this, the second term of the expression accounts for the fact that he fails to obtain a fill for an order only if it has latency 3ε (which occurs with probability $1-p_I$) and the liquidity provider's corresponding cancellation has latency ε (which occurs with probability p_H). The first term corrects for the fact that all orders are filled if all have latency 3ε. See Figures 2 and 3 for depictions of the $q = 0$ and $q = 1$ cases, respectively. Intermediate values of q monotonically bridge these two extremes.

Thus, higher values of q reduce an information investor's expected number of trades. As Corollary 3 states, this leads to a lower research intensity. However, it also reduces adverse selection against the liquidity provider, which leads to a smaller spread.

COROLLARY 3: *Both the spread s^*_{qND} and research intensity r^*_{qND} are weakly decreasing in q.*

A.3. Comparison to Limit Order Book Outcomes

How does the qND equilibrium compare to that of the LOB? In general, the answer depends on q. However, an attractive feature of this family of mechanisms is that for an appropriately chosen value of q, qND dominates the LOB along both dimensions: it implements both a smaller spread and higher research intensity. In this sense, our analysis shows that meaningful improvements can be derived from only minor modifications of the prevailing LOB design.

COROLLARY 4: *There exists a $\hat{q} \in [0,1]$ such that $s^*_{\hat{q}ND} \leq s^*_{LOB}$ and $r^*_{\hat{q}ND} \geq r^*_{LOB}$.*

Because $X_{0ND} \geq X_I$ and $X_{1ND} \leq X_I$, the intermediate value theorem guarantees a $\hat{q}$ for

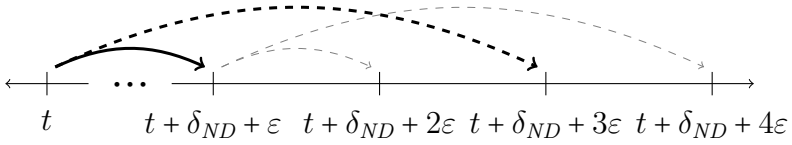


Figure 2. Illustration of equilibrium (0-ND). One equilibrium scenario: (*i*) the investor sends orders at $t \in \mathcal{T}$, to either one or all exchanges depending on his type, and (*ii*) at least one order receives the shorter latency, being received at $t+\varepsilon$ but not processed until $t + \delta_{ND} + \varepsilon$ (solid black arrow), while any remaining orders receive the longer latency and are processed at $t + \delta_{ND} + 3\varepsilon$ (dashed black arrow). The liquidity provider reacts to the trade(s) at $t + \delta_{ND} + \varepsilon$ by sending cancellations that will be processed at $t + \delta_{ND} + 2\varepsilon$ or $t + \delta_{ND} + 4\varepsilon$ (dashed gray arrows). Sniper orders are not processed until $t + 2\delta_{ND} + 2\varepsilon$ at the earliest, by which time all quotes have been cancelled.

which $X_{\hat{q}ND} = X_I$. For this $\hat{q}$, an information investor's optimal research intensity is the same weakly decreasing function of the spread under both $\hat{q}$ND and the LOB (see equations (2) and (5)). To prove the result, it suffices to show that for a fixed research intensity, the spread is weakly lower under $\hat{q}$ND than under the LOB. But this follows from the fact that NDs eliminate sniping, so that $X_{\hat{q}ND} \leq X_I + X_S$ (see equations (1) and (4)).

B. Frequent Batch Auctions

We next consider the FBA mechanism, whereby exchanges conduct uniform-price sealed-bid double auctions at repeated intervals. This differs from the LOB in several ways, most notably in breaking the sequential nature of order processing. The effect is to eliminate not only aggressive-side but also passive-side order anticipation.

This proposal has received a great deal of recent attention from academics—most notably, Budish, Cramton, and Shim (2015)—as well as policymakers and industry participants. Although our formulation below closely follows Budish, Cramton, and Shim (2015), we reverse their conclusion that FBAs implement a smaller spread than the LOB.

B.1. Definition

We consider an FBA design that is motivated by the Budish, Cramton, and Shim (2015) proposal.[31] They advocate a batch length that is "long" relative to latency, yet still relatively

[31]Two subtle differences exist between our model of batch auctions and that in Budish, Cramton, and Shim (2015): (*i*) their model has one order type, whereas ours has two, and (*ii*) their model has one double

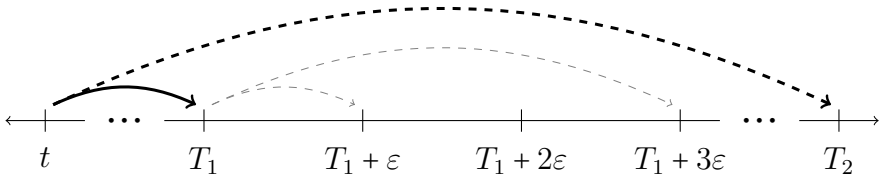


Figure 3. Illustration of equilibrium (1-ND). One equilibrium scenario: the investor sends orders at $t \in \mathcal{T}$, to either one or all exchanges depending on his type. Let T_1 denote the time at which the first of these orders is processed (solid black arrow). The liquidity provider reacts to the trade at T_1 by sending cancellations that will be processed at $T_1+\varepsilon$ or $T_1+3\varepsilon$ (dashed gray arrows). Sniper orders are not processed until $T_1+\delta_{ND}+\varepsilon$ at the earliest, by which time all quotes have been cancelled. Any subsequent orders that the investor had sent (e.g., dashed black arrow) will be processed almost surely after all quotes have been cancelled.

short. The natural analogue in the language of this paper is that the batch length be an element of $\mathcal{T}$ that is infinitesimal yet infinitely larger than ε, which we assume hereafter. In addition, we focus on auctions that are synchronized across exchanges, which those authors also identify as an attractive property (Budish, Cramton, and Shim (2014)).[32] For instance, if the batch length is $\sqrt{\varepsilon}$, then the exchanges would hold separate auctions at each time $\sqrt{\varepsilon}$, $2\sqrt{\varepsilon}$, $3\sqrt{\varepsilon}$, and so on.

In what follows, we describe a specific model of the individual batch auctions that seeks to remain as close as possible to the LOB model of the baseline. The model allows for two order types, *competitive* and *noncompetitive*. Liquidity providers are restricted to using competitive orders, while investors and snipers are restricted to noncompetitive orders. Both order types specify quantities and prices, but they are treated differently within the auction as described in the next paragraph. The language "competitive" and "noncompetitive" comes from Treasury auctions (Federal Reserve Bank of New York (2017)), where, as in our model, competitive orders may set the price while noncompetitive orders may not. Thus, competitive orders are analogous to post-only orders, and noncompetitive orders are

auction, whereas bid and ask sides clear separately in ours. These modifications are primarily for tractability (an additional challenge for us given the presence of asymmetric information). However, Budish, Cramton, and Shim (2015) would obtain the same findings with or without these modifications and thus our findings can still be compared with theirs.

[32] While perfect synchronization might be difficult to achieve and enforce in practice (especially across competing exchanges), small deviations from it do not affect our results, so long as they are bounded by the minimum HFT latency. Internet Appendix Section III.D treats the case of large deviations.

analogous to immediate-or-cancel orders. In terms of these analogies, we retain all previous restrictions on orders that the various traders can submit. We also adopt the conventions that unfilled competitive orders are carried over to the next batch auction, remaining active until cancelled, while unfilled noncompetitive orders expire.

At the end of the interval, each exchange computes four aggregate schedules based on competitive buy, competitive sell, noncompetitive buy, and noncompetitive sell orders. Then, in a bid-side cross, competitive buy orders match with noncompetitive sell orders, and in an ask-side cross, competitive sell orders match with noncompetitive buy orders. For each cross, there are four cases. In the first, there is no market-clearing price (i.e., the schedules do not intersect), in which case no trade occurs. In the second, there is a unique market-clearing price but a range of market-clearing quantities (i.e., the schedules intersect horizontally), in which case the maximum quantity is chosen, and if there is a long side of the market, such orders are rationed pro rata. In the third, there is a unique market-clearing quantity but a range of market-clearing prices (i.e., the schedules intersect vertically), in which case the maximum (minimum) of the range is used for the bid-side (ask-side) auction.[33] The fourth case, in which the market-clearing price and quantity are both unique (i.e., the two schedules intersect at a point), can be treated as a special instance of either of the previous two cases. At the end of the interval, the aggregate schedules are announced, along with the clearing prices and quantities. However, each auction is "sealed bid" in that no information is released until the end of the interval. Finally, for a batch auction, we define the *spread* to be the following analogue of its LOB counterpart: the difference between the lowest price at which there exists a competitive order to sell (the ask) and the highest price at which there exists a competitive order to buy (the bid).

B.2. Equilibrium

We next characterize the FBA equilibrium in terms of the spread and research intensity.

PROPOSITION 4: *Under FBAs, there exists a WPBE in which the spread s^*_{FBA} and research intensity r^*_{FBA} are the unique solution to*

$$s^*_{FBA} = \frac{2\lambda r^*_{FBA} X}{1 - \lambda + \lambda r^*_{FBA} X} \tag{6}$$

$$r^*_{FBA} \in \arg\max_{r \in [0,1]} \left\{ rX \left(1 - \frac{s^*_{FBA}}{2} \right) - c(r) \right\}. \tag{7}$$

[33] Hence, noncompetitive orders are price-taking while competitive orders are price-setting.

The strategies that support this outcome in equilibrium are also essentially as before (see Section III.A). The primary differences are that the liquidity provider uses competitive orders to set her quotes, and investors use noncompetitive orders to trade against them. We defer a formal statement of equilibrium strategies to the proof of Proposition 4 in Internet Appendix Section I.

Trading outcomes again change in two ways relative to the LOB baseline: (*i*) in terms of the number of trades made by information investors, with X_I now mapping into X, and (*ii*) in terms of the number of trades made by snipers, with X_S again mapping into zero. Given this, the equilibrium spread and research intensity are again pinned down as before. Equation (6) ensures that the liquidity provider earns zero profits, and equation (7) ensures that information investors choose research intensity optimally.

The key difference relative to the baseline is that FBAs allow information investors to obtain fills at all X exchanges. This is because the first of an information investor's orders to arrive does not result in an immediate trade. Rather, trade is delayed until the end of the batch interval, by which time all his orders have arrived. Because the auctions clear simultaneously, HFTs have no opportunity to react, and all of these orders are converted into fills (see Figure 4). Thus, like NDs, FBAs prevent aggressive-side order anticipation and as a result root out the main friction in the model. However, they have an additional effect: they also prevent passive-side order anticipation.[34]

B.3. Comparison to Limit Order Book Outcomes

FBAs implement higher research intensity and a larger spread than either the LOB or any ND mechanism. Information investors choose the highest research intensity because FBAs enable them to convert all orders into fills. And the liquidity provider quotes the largest spread because FBAs maximize adverse selection: not only is research intensity at its highest, but also the liquidity provider never succeeds in cancelling a mispriced quote.

PROPOSITION 5: *For any* $q \in [0,1]$, $s^*_{FBA} \geq s^*_{qND}, s^*_{LOB}$ *and* $r^*_{FBA} \geq r^*_{qND}, r^*_{LOB}$.

That FBAs induce the largest spread in our model is in sharp contrast to Budish, Cramton, and Shim (2015), who find that FBAs reduce the spread and indeed eliminate it entirely. The crucial difference between the two models is the source of adverse selection. In both models, batching eliminates adverse selection from snipers—it allows the liquidity provider

[34] As we discuss in Internet Appendix Section III.E, the same effect is obtained by a processing delay applied to *all* orders.

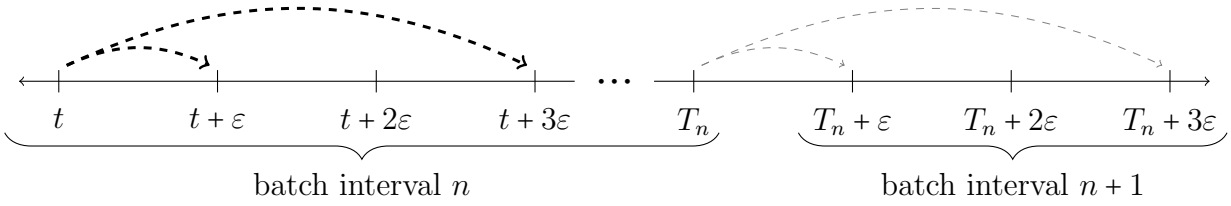


Figure 4. Illustration of equilibrium (FBAs). One equilibrium scenario: the investor sends orders at $t \in \mathcal{T}$ (dashed black arrows), to either one or all exchanges depending on his type. With probability one, these are processed during the same batch interval. Let T_n denote the time at which the auctions for that interval take place and corresponding trades are announced. The liquidity provider reacts to the trade(s) at T_n by sending cancellations that may be processed at $T_n+\varepsilon$ or $T_n+3\varepsilon$ (dashed gray arrows). Sniper orders are processed in the same batch interval as the liquidity provider's cancellations and thus result in no trades.

to cancel her mispriced quotes before snipers can trade against them. However, in Budish, Cramton, and Shim (2015), snipers are the only source of adverse selection, as information is taken to be exogenous and public, whereas our model also features a second source, namely, a trader who endogenously acquires private information. Batching actually *increases* this latter source of adverse selection—it prevents the liquidity provider from cancelling her mispriced quotes before an information investor can trade against them. Moreover, this effect dominates in our model.[35]

C. NDs and FBAs Achieve the Frontier of $\mathcal{F}$

As we see above, both NDs and FBAs eliminate aggressive-side order anticipation. As a result, these mechanisms implement outcomes on the frontier of the tradeoff between liquidity and information production.[36]

COROLLARY 5: *If* $c'(1) \geq \dfrac{(1-\lambda)X}{1-\lambda+\lambda X}$, *then*

(i) for all $q \in [0,1]$, *the qND outcome is on the frontier of* $\mathcal{F}$*;*

[35] In a hybrid model with both private information acquisition and public news, the comparison between s^*_{FBA} and s^*_{LOB} would be theoretically ambiguous. Nevertheless, that $s^*_{FBA} \geq s^*_{qND}$ (as in Proposition 5) would be unambiguous because both mechanisms eliminate sniping on public news.

[36] Risk neutrality is important for this result. With risk aversion, achieving the frontier requires not only that information investors receive all of the profits from informed trading, but also that those profits are deterministic. NDs satisfy the first criterion, but for $q < 1$ they generally fail the second.

(ii) the FBA outcome is on the frontier of $\mathcal{F}$; and
(iii) if in addition either $X \leq 2$ or $p_I = 1$, then the the LOB outcome is on the frontier of $\mathcal{F}$.

We sketch the proof for NDs. In that case, by Proposition 3, the equilibrium is characterized by equations (4) and (5). The latter equation implies $r^*_{qND}c'(r^*_{qND}) = \frac{(1-\lambda)r^*_{qND}X_{qND}}{1-\lambda+\lambda r^*_{qND}X_{qND}}$.[37] Because $w^*_{qND} = \beta - s^*_{qND}/2$, the former equation implies $w^*_{qND} = \beta - \frac{\lambda r^*_{qND}X_{qND}}{1-\lambda+\lambda r^*_{qND}X_{qND}}$. It then follows that $(1-\lambda)(\beta - w^*_{qND}) = \lambda r^*_{qND}c'(r^*_{qND})$, so that by Proposition 2, (r^*_{qND}, w^*_{qND}) is indeed on the frontier of $\mathcal{F}$. Analogous arguments establish the corollary's other claims.

The last part of the corollary states that if investor order synchronization is perfect ($p_I = 1$) or if there are very few exchanges ($X \leq 2$), then the LOB also implements an outcome on the frontier, for the same reason: there is no aggressive-side order anticipation. Indeed, if either condition holds, then $X_S = 0$. But as Corollary 2 above states, the LOB fails to achieve the frontier under more general conditions.

The preceding results emphasize that influence over the trading mechanism may be a useful lever for policy. The next result highlights a complementarity between this lever and another that a policymaker may possess: influence over the number of exchanges. Indeed, the Exchange Act grants the Securities and Exchange Commission (SEC) authority to approve exchange applications, and the recent proliferation of trading venues in the U.S. was driven in large part by SEC Regulations National Market System (NMS) and Alternative Trading Systems (ATS). Corollary 6(i) states that by adjusting the number of exchanges, NDs can implement all points on the frontier except those with very low research intensities. This can be interpreted as a partial converse to Corollary 5(i). FBAs can also implement a number of points on the frontier. In fact, the analogous partial converse would obtain if it were possible to drop the integer constraint on the number of exchanges.

COROLLARY 6: *Assume $c'(1) \geq 1-\lambda$, and let $r_{\min} = \min\left\{r \in [0,1] : c'(r) \geq \frac{1-\lambda}{1-\lambda+\lambda r}\right\}$. If (r,w) is on the frontier of $\mathcal{F}$ and $r \geq r_{\min}$, then*

[37] The proof of this step requires also assuming that the marginal cost of research be sufficiently high at $r = 1$. This eliminates the possibility that information investors are at a corner solution, choosing to conduct research with maximal intensity $r = 1$. Indeed, if an information investor were at such a corner solution, then the outcome would be inside the frontier of $\mathcal{F}$ because his rents could be reduced without affecting his choice of research intensity. Because it seems prohibitively difficult to acquire all knowable information about a security, we view this assumption as mild.

(i) there exist $X' \in \mathbb{N}$ *and* $q \in [0,1]$ *such that* (r,w) *is implemented by qND with* X' *exchanges;*

(ii) there exists $X'' \in \mathbb{R}_{\geq 1}$ *such that* (r,w) *is implemented by FBAs with* X'' *exchanges (where this refers to what would obtain by extending Proposition 4 to the domain* $X \in \mathbb{R}_{\geq 1}$*).*

The driving force behind Corollary 6 is the observation of Section III.B that aggregate depth increases in the number of exchanges. More aggregate depth means more trading opportunities for information investors, which incentivizes more research. However, it also exacerbates adverse selection, increasing the spread, and so the outcome moves along the frontier. The meaning of $r_{\min}$ is the research intensity that prevails under a single exchange. Because research intensity increases in the number of exchanges, neither mechanism can implement an intensity below $r_{\min}$. In other words, there is a minimum amount of adverse selection that none of these mechanisms can eliminate.

The advantages of NDs and FBAs are clear: they achieve the frontier of the tradeoff between liquidity and information production (which LOBs generally do not). However, certain unmodeled disadvantages may also exist. For instance, these alternatives depend fundamentally on delaying certain trades, and such delays might be costly. On the other hand, such delays would be very short (FBAs might require delays on the order of one second, and NDs would be even shorter, on the order of milliseconds or hundreds of microseconds depending on the amount of randomness) in which case these costs are unlikely to be economically significant. In practice, FBAs might require delays on the order of one second, and NDs would be even shorter, on the order of milliseconds or hundreds of microseconds (depending on the amount of randomness). On the other hand, delay costs could substantially detract from the performance of other types of trading mechanisms, such as "infrequent" batch auctions. Therefore, incorporating delay costs into the model would likely strengthen our results on the desirability of the specific alternative trading mechanisms that we consider.

Another commonly voiced concern about NDs is that they lead to passive-side order anticipation and hence contribute to so-called "phantom liquidity." Our model captures this, but it also highlights a countervailing force: NDs eliminate aggressive-side order anticipation. The net effect depends on the parametrization of the delay. If the delay is deterministic (i.e., $q = 0$), our analysis suggests that investors would experience *less* phantom liquidity (in the model, $X_{0ND} \geq X_I$). But if the delay is random (i.e., $q = 1$), investors would likely experience more phantom liquidity (in the model, $X_{1ND} \leq X_I$). Nevertheless, the latter should not necessarily be viewed as problematic—it is precisely the mechanism through which the outcome moves along the frontier, as research intensity is sacrificed for smaller

spreads.

VI. Conclusion

Trillions of dollars are traded on financial platforms each day. As a result, even small changes in their technology or in market design may have considerable consequences and as such deserve careful analysis. This paper provides a framework for evaluating both the consequences of recent acceleration in the speed of HFT and the effects of replacing the LOB with certain alternatives that are currently in debate.

The model highlights two main consequences of faster speeds: smaller spreads and less intensive research. These effects stem from improving the success of two HFT strategies, namely, aggressive-side and passive-side order anticipation. Yet our results also highlight an important distinction between the two strategies. In particular, aggressive-side order anticipation has unambiguously detrimental implications, both increasing the spread and reducing information acquisition. In contrast, our results are more ambiguous about the normative implications of passive-side order anticipation, which, despite also reducing information acquisition, *reduces* the spread.

The analysis of alternative trading mechanisms focuses on two specific proposals: NDs and FBAs. Both alternatives eliminate aggressive-side order anticipation and in so doing implement equilibria on the frontier of the tradeoff between small spreads and intensive research. Collectively, the various NDs implement a segment of the frontier. FBAs implement a point on the frontier characterized by relatively more intensive research but a larger spread. The specific mechanism recommended by our analysis would therefore depend on how the regulator weighs liquidity against information production, which might vary substantially across securities and asset classes. Nevertheless, our analysis unequivocally shows that the LOB—despite being current industry practice—is suboptimal in the sense that it generally does not achieve the frontier.

Initial submission: June 20, 2018; Accepted: July 12, 2019
Editor: Philip Bond

REFERENCES

Aisen, Daniel, Bradley Katsuyama, Robert Park, John Schwall, Richard Steiner, Allen Zhang, and Thomas L. Popejoy, 2015, Synchronized processing of data by networked computing resources, US Patent 8,984,137.

Aït-Sahalia, Yacine, and Mehmet Sağlam, 2017a, High frequency market making: Implications for liquidity, Working paper, Princeton University.

Aït-Sahalia, Yacine, and Mehmet Sağlam, 2017b, High frequency market making: Optimal quoting, Working paper, Princeton University.

Aitken, Michael, Haoming Chen, and Sean Foley, 2017, The impact of fragmentation, exchange fees and liquidity provision on market quality, *Journal of Empirical Finance* 41, 140–160.

Aldrich, Eric M., and Daniel Friedman, 2019, Order protection through delayed messaging, Working paper, University of California, Santa Cruz.

Back, Kerry, and Shmuel Baruch, 2004, Information in securities markets: Kyle meets Glosten and Milgrom, *Econometrica* 72, 433–465.

Baldauf, Markus, and Joshua Mollner, forthcoming, Trading in fragmented markets, *Journal of Financial and Quantitative Analysis*.

BATS Global Markets, Inc. (BATS), 2009, BATS exchange announces further latency reductions, press release, http://cdn.batstrading.com/resources/press_releases/BATS_Latency_Upgrade_FINAL.pdf.

BATS Global Markets, Inc. (BATS), 2018, Cboe system performance: World-class, sustained low latency, http://cdn.batstrading.com/resources/features/bats_exchange_Latency.pdf Accessed: April 13, 2019.

Bernales, Alejandro, 2019, Make-take decisions under high-frequency trading competition, *Journal of Financial Markets* 45, 1–18.

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

Biais, Bruno, David Martimort, and Jean-Charles Rochet, 2000, Competing mechanisms in a common value environment, *Econometrica* 68, 799–837.

Boehmer, Beatrice, and Ekkehart Boehmer, 2003, Trading your neighbor's ETFs: Competition or fragmentation?, *Journal of Banking & Finance* 27, 1667–1703.

Boehmer, Ekkehart, Kingsley Y.L. Fong, and Juan (Julie) Wu, 2018, International evidence on algorithmic trading, Working paper, Singapore Management University.

Bongaerts, Dion, and Mark Van Achter, 2016, High-frequency trading and market stability, Working paper, Erasmus University Rotterdam.

Brogaard, Jonathan, 2010, High Frequency Trading and Its Impact on Market Quality, Working paper, University of Utah.

Brogaard, Jonathan, Björn Hagströmer, Lars Nordén, and Ryan Riordan, 2015, Trading fast and slow: Colocation and market quality, *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.

Brolley, Michael, and David Cimon, 2018, Order flow segmentation, liquidity and price discovery: The role of latency delays, Working paper, Wilfrid Laurier University.

Budish, Eric, Peter Cramton, and John Shim, 2014, Implementation details for frequent batch auctions: Slowing down markets to the blink of an eye, *American Economic Review: Papers & Proceedings* 104, 418–424.

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

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

Chaboud, Alain, Ben Chiquoine, Erik Hjalmarsson, and Clara Vega, 2014, Rise of the machines: Algorithmic trading in the foreign exchange market, *Journal of Finance* 69, 2045–2084.

Chakravarty, Sugato, Pankaj Jain, James Upson, and Robert Wood, 2012, Clean sweep: Informed trading through intermarket sweep orders, *Journal of Financial and Quantitative Analysis* 47, 415–435.

Chowdhry, Bhagwan, and Vikram Nanda, 1991, Multimarket trading and market liquidity, *Review of Financial Studies* 4, 483–511.

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

Dennert, Jürgen, 1993, Price competition between market makers, *Review of Economic Studies* 60, 735–751.

Du, Songzi, and Haoxiang Zhu, 2017, What is the optimal trading frequency in financial markets?, *Review of Economic Studies* 84, 1606–1651.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, 1997, The information content of the trading process, *Journal of Empirical Finance* 4, 159–186.

Easley, David, and Maureen O'Hara, 1987, Price, trade size, and information in securities markets, *Journal of Financial Economics* 19, 69–90.

Federal Reserve Bank of New York, 2017, Treasury auctions.

Fink, Jason, Kristin E. Fink, and James P. Weston, 2006, Competition on the Nasdaq and the growth of electronic communication networks, *Journal of Banking & Finance* 30, 2537–2559.

Fishman, Michael J., and Francis A. Longstaff, 1992, Dual trading in futures markets, *Journal of Finance* 47, 643–671.

Foucault, Thierry, 1999, Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial Markets* 2, 99–134.

Foucault, Thierry, Johan Hombert, and Ioanid Roşu, 2016, News trading and speed, *Journal of Finance* 71, 335–382.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171–1217.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2013, Liquidity cycles and make/take fees in electronic markets, *Journal of Finance* 68, 299–341.

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham, 2017, Toxic arbitrage, *Review of Financial Studies* 30, 1053–1094.

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

Frino, Alex, Vito Mollica, and Robert I. Webb, 2014, The impact of co-location of securities exchanges' and traders' computer servers on market liquidity, *Journal of Futures Markets* 34, 20–33.

Gider, Jasmin, Simon N. M. Schmickler, and Christian Westheide, 2019, High-frequency trading and price informativeness, Working paper, Tilburg University.

Glosten, Lawrence R., 1989, Insider trading, liquidity, and the role of the monopolist specialist, *Journal of Business* 62, 211–235.

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable?, *Journal of Finance* 49, 1127–1161.

Glosten, Lawrence R., 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.

Goettler, Ronald L., Christine A. Parlour, and Uday Rajan, 2009, Informed traders and limit order markets, *Journal of Financial Economics* 93, 67–87.

Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.

Hasbrouck, Joel, 1991, Measuring the information content of stock trades, *Journal of Finance* 46, 179–207.

Hasbrouck, Joel, and Gideon Saar, 2013, Low-latency trading, *Journal of Financial Markets* 16, 646–679.

He, Peng William, Elvis Jarnecic, and Yubo Liu, 2015, The determinants of alternative trading venue market share: Global evidence from the introduction of Chi-X, *Journal of Financial Markets* 22, 27–49.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1–33.

Jovanovic, Boyan, and Albert J. Menkveld, 2016, Middlemen in limit order markets, Working paper, New York University.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Lee, Charles, and Edward M. Watts, 2018, Tick size tolls: Can a trading slowdown improve earnings news discovery?, Working paper, Stanford University.

Lin, Ji-Chai, Gary C. Sanger, and G. Geoffrey Booth, 1995, Trade size and components of the bid-ask spread, *Review of Financial Studies* 8, 1153–1183.

Loeb, Peter A., 1975, Conversion from nonstandard to standard measure spaces and applications in probability theory, *Transactions of the American Mathematical Society* 211, 113–122.

Lyle, Matthew R., and James P. Naughton, 2016, How does algorithmic trading improve market quality?, Working paper, Northwestern University.

Malinova, Katya, Andreas Park, and Ryan Riordan, 2018, Do retail investors suffer from high frequency traders?, Working paper, McMaster University.

Mendelson, Haim, 1987, Consolidation, fragmentation, and market performance, *Journal of Financial and Quantitative Analysis* 22, 189–207.

Menkveld, Albert J., 2013, High frequency trading and the *new market* makers, *Journal of Financial Markets* 16, 712–740.

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

Parlour, Christine A., 1998, Price dynamics in limit order markets, *Review of Financial Studies* 11, 789–816.

Quincy Data, 2019, Quincy extreme data latencies, Technical report.

Robinson, Abraham, 1966, *Non-Standard Analysis* (Princeton University Press).

Röell, Ailsa, 1990, Dual-capacity trading and the quality of the market, *Journal of Financial Intermediation* 1, 105–124.

Rojček, Jakub, and Alexandre Ziegler, 2016, High-frequency trading in limit order markets: Equilibrium impact and regulation, Working paper, University of Zurich.

Roşu, Ioanid, 2009, A dynamic model of the limit order book, *Review of Financial Studies* 22, 4601–4641.

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* (ACM).

Weller, Brian M., 2018, Does algorithmic trading reduce information acquisition?, *Review of Financial Studies* 31, 2184–2226.

Yang, Liyan, and Haoxiang Zhu, forthcoming, Back-running: Seeking and hiding fundamental information in order flows, *Review of Financial Studies*.

Internet Appendix for "High-Frequency Trading and Market Performance"

MARKUS BALDAUF and JOSHUA MOLLNER[1]

ABSTRACT

This Internet Appendix provides additional derivations and analyses that supplement the main text.

[1]Citation format: Baldauf, Markus, and Joshua Mollner, Internet Appendix for "High-Frequency Trading and Market Performance," *Journal of Finance* [DOI String]. Please note: Wiley-Blackwell is not responsible for the content or functionality of any additional information provided by the authors. Any queries (other than missing material) should be directed to the authors of the article.

I. Proofs

LEMMA 1: *The function*

$$F_{LOB}(y) = yp_I + y(1-p_I)^y + (1-p_H)(y-1)yp_I(1-p_I)^{y-1}$$

is increasing on the domain of positive integers.

Proof of Lemma 1: We will instead prove the stronger statement that $F_{LOB}(\cdot)$ is increasing on the domain $[1,\infty)$. Taking the derivative,

$$\begin{aligned} F'_{LOB}(y) = {} & p_I && \text{(IA1)} \\ & + (1-p_I)^y \left[y\ln(1-p_I)+1\right] && \text{(IA2)} \\ & + (1-p_H)p_I(1-p_I)^{y-1}\left[2y + y(y-1)\ln(1-p_I) - 1\right]. && \text{(IA3)} \end{aligned}$$

We proceed by deriving lower bounds on each of (IA1), (IA2), and (IA3).

- For (IA1), we have a lower bound of $\frac{1}{2}$.
- For (IA2), we begin by making the change of variables $z = (1-p_I)^y$ to rewrite it as $z\ln(z)+z$. This expression is minimized when $z = e^{-2}$. Evaluating at that value yields a lower bound of $-e^{-2} \approx -0.135$.
- For (IA3), we begin by instead deriving a lower bound for the expression

  $$(1-p_I)^{y-1}\left[2y + y(y-1)\ln(1-p_I) - 1\right].$$

  We make the change of variables $z = (1-p_I)^{y-1}$ to rewrite that expression as $2zy + zy\ln(z) - z$. Note that $p_I \geq \frac{1}{2}$ requires that $y \leq 1 - \frac{\ln(z)}{\ln(2)}$. The expression is linear in y, so it suffices to derive a lower bound that applies at the endpoints $y \in \left\{1, 1-\frac{\ln(z)}{\ln(2)}\right\}$.
  - Evaluating the expression at $y = 1$, it becomes $z\ln(z)+z$. This expression is minimized when $z = e^{-2}$. Evaluating at that value yields a lower bound of $-e^{-2} \approx -0.135$.
  - Evaluating the expression at $y = 1 - \frac{\ln(z)}{\ln(2)}$, it becomes

    $$z\ln(z)\left[1 - \frac{\ln(z)+2}{\ln(2)}\right] + z,$$

which is minimized at $z \approx 0.045$. Evaluating at that value yields a lower bound of ≈ -0.316.

Thus, ≈ -0.316 is a lower bound for the earlier expression. Using $p_H \geq \frac{1}{2}$ and $p_I \leq 1$, we can transform this into a lower bound for (IA3) of ≈ -0.158.

Summing the three lower bounds, we obtain ≈ 0.207, and so we conclude that the derivative is indeed positive on the domain of interest. □

LEMMA 2: *The function*

$$F_{qND}(y) = q^y + \sum_{x=1}^{y} \binom{y}{x} (1-q)^x q^{y-x} \left(x p_H (1-p_I)^x + x\left[1 - p_H(1-p_I)\right]\right)$$

is weakly increasing on the domain of positive integers.

Proof of Lemma 2: Let

$$f(x) = \begin{cases} \left(x p_H (1-p_I)^x + x\left[1 - p_H(1-p_I)\right]\right) & \text{if } x \geq 1 \\ 1 & \text{if } x = 0. \end{cases}$$

We first establish that $f(\cdot)$ is weakly increasing on the domain of nonnegative integers. Notice that $f(0) = f(1) = 1$. Given this, it suffices to prove that $f(\cdot)$ is increasing on the domain $[1, \infty)$. Taking the derivative,

$$f'(x) = p_H (1-p_I)^x \left[x \ln(1-p_I) + 1\right] \tag{IA4}$$

$$+ 1 - p_H(1-p_I). \tag{IA5}$$

We proceed by deriving lower bounds on both (IA4) and (IA5).

- For (IA4), we begin by making the change of variables $z = (1-p_I)^y$ to rewrite it as $p_H[z \ln(z) + z]$. This expression is minimized when $z = e^{-2}$ and $p_H = 1$. Evaluating at those values yields a lower bound of $-e^{-2} \approx -0.135$.
- For (IA5), the expression is minimized when $p_I = 0.5$ and $p_H = 1$. Evaluating at those values yields a lower bound of $\frac{1}{2}$.

Summing the two lower bounds, we obtain ≈ 0.365, and so we conclude that the derivative is indeed positive on the domain of interest.

Next, notice that $F_{qND}(y)$ is simply a weighted average of $f(x)$ over various values of x:

$$F_{qND}(y) = \sum_{x=0}^{y} \binom{y}{x} (1-q)^x q^{y-x} f(x).$$

Moreover, the effect of an increase in y is to shift the distribution over x upward in the sense of first-order stochastic dominance. Because $f(x)$ is weakly increasing in x, we conclude that X_{qND} is weakly increasing in y. □

Proof of Proposition 1: The proof consists of four parts. First, we argue that there exists a unique solution to (1) and (2). Second, we describe the beliefs that are part of the equilibrium. (The strategies are specified in Section III.A of the main article.) Third, we verify that these beliefs are consistent with Bayes' rule. Fourth, we verify that strategies are sequentially rational given beliefs.

Part One (Existence and Uniqueness): Plugging (1) into (2), r^*_{LOB} is characterized as a fixed point of the correspondence

$$R_{LOB}(\hat{r}) = \arg\max_{r \in [0,1]} \left\{ \frac{(1-\lambda) r X_I}{1 - \lambda + \lambda \hat{r}(X_I + X_S)} - c(r) \right\}.$$

By continuity of $c(\cdot)$, the maximum theorem (Berge (1963)) implies that $R_{LOB}(\cdot)$ is nonempty and upper semicontinuous in $\hat{r}$. Moreover, by convexity of $c(\cdot)$, $R_{LOB}(\cdot)$ is convex valued. Therefore, Kakutani's fixed point theorem (Kakutani (1941)) implies the existence of a fixed point.

We now argue that the fixed point is unique. Suppose to the contrary that there exist $r_1 \neq r_2$, with $r_1 \in R_{LOB}(r_1)$ and $r_2 \in R_{LOB}(r_2)$. By Topkis' Theorem (Topkis (1978)), $R_{LOB}(\cdot)$ is weakly decreasing in the strong set order. Therefore, $r_2 \in R_{LOB}(r_1)$ and $r_1 \in R_{LOB}(r_2)$. Because $c(\cdot)$ is convex and continuously differentiable, this can be the case only if

$$\frac{(1-\lambda) X_I}{1 - \lambda + \lambda r_1 (X_I + X_S)} = c'(r_1) = c'(r_2) = \frac{(1-\lambda) X_I}{1 - \lambda + \lambda r_2 (X_I + X_S)}.$$

However, because $r_1 \neq r_2$, this can hold only if $X_I = 0$, $\lambda = 1$, $\lambda = 0$, or $X_I + X_S = 0$, none of which is the case. We therefore have a contradiction, and a unique fixed point, r^*_{LOB}, exists. Finally, s^*_{LOB} can be uniquely derived from r^*_{LOB} and (1).

Part Two (Description): See Section III.A of the main article for a description of the equilibrium strategies. As a technical point, note that because $\mathcal{T}$, like the real line, is not a

well-ordered set, it is not possible to determine the outcome of this strategy profile via induction on the set of times at which agents can move. Nevertheless, because there is only a finite set of times at which agents choose to move on path, it is possible to uniquely determine the outcome of this strategy profile via induction on the set of times at which agents choose to move, as in equation (4.1) of Simon and Stinchcombe (1989).

If traders use the strategies defined in Section III.A of the main article, then X_I indeed represents the expected number of trades made by an information investor conditional on learning the value of the security. To see this, suppose that the information investor submits orders to all X exchanges at time t. With probability $(1-p_I)^X$, none of these orders are processed at $t+\varepsilon$, in which case all are filled at $t+3\varepsilon$. With probability $Xp_I(1-p_I)^{X-1}$, exactly one of these orders is processed at $t+\varepsilon$, in which case the snipers do not react and the information investor needs only outrace the liquidity provider at the remaining $X-1$ exchanges, so he expects to receive an additional $(1-p_H)(X-1)$ fills at $t+3\varepsilon$. In all other cases, the snipers do react; since there are an infinite number of snipers and latencies are drawn independently across traders, with probability one at least one HFT achieves the minimum latency of ε on each exchange, which precludes the information investor from receiving any further fills at $t+3\varepsilon$. Summing over these cases yields the expression for X_I asserted in the proposition.

Likewise, X_S does indeed represent the expected number of trades made by snipers (in aggregate) conditional on an information investor arriving and learning the value of the security. If the information investor sends orders to all exchanges at time t, then the liquidity provider succeeds in canceling a mispriced quote only if exactly one of the information investor's orders is processed at $t+\varepsilon$. This occurs with probability $Xp_I(1-p_I)^{X-1}$, in which case the liquidity provider expects to cancel $p_H(X-1)$ mispriced quotes before they are exploited by the information investor. Snipers capture all mispriced quotes that are neither cancelled by the liquidity provider nor captured by the information investor. We must then have $X_S = X - X_I - p_H(X-1)Xp_I(1-p_I)^{X-1}$, which yields the expression asserted in the proposition.

To complete the description of the WPBE, it remains to specify beliefs. The relevant beliefs are what the HFTs believe in the time period in which the first trade occurs about who initiated those trades. We consider the cases $X=1$ and $X\geq 2$ separately.

- First, suppose $X=1$. If one trade occurs at the ask (bid) in that time period, then HFTs believe it to have been initiated either by a liquidity investor with a buying (selling)

motive, with probability

$$\frac{1-\lambda}{1-\lambda+\lambda r^*_{LOB}},$$

or by an information investor who learned $v = 1$ ($v = -1$), with probability

$$\frac{\lambda r^*_{LOB}}{1-\lambda+\lambda r^*_{LOB}}.$$

- Second, suppose $X \geq 2$. If exactly one trade occurs at the ask (bid) in that time period, then HFTs believe it to have been initiated either by a liquidity investor with a buying (selling) motive, with probability

$$\frac{1-\lambda}{1-\lambda+\lambda r^*_{LOB} X p_I (1-p_I)^{X-1}},$$

or by an information investor who learned $v = 1$ ($v = -1$), with probability

$$\frac{\lambda r^*_{LOB} X p_I (1-p_I)^{X-1}}{1-\lambda+\lambda r^*_{LOB} X p_I (1-p_I)^{X-1}}.$$

If two or more trades occur at the ask (bid) in that time period, then HFTs believe them to have been initiated by an information investor who learned $v = 1$ ($v = -1$), with probability one.

Part Three (Consistency): Given equilibrium strategies, we have the following. One the one hand, the investor is a liquidity investor with a buying (selling) motive with probability $(1-\lambda)/2$. Conditional on there being such an investor, in the time period in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability one. On the other hand, the investor is an information investor who learns $v = 1$ ($v = -1$) with probability $\lambda r^*_{LOB}/2$. If $X = 1$ and conditional on there being such an investor, then in the time period in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability one. If $X \geq 2$ and conditional on there being such an investor, then in the time period in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability $X p_I (1-p_I)^{X-1}$, and two or more such trades occur with probability $1 - X p_I (1-p_I)^{X-1}$. Applying Bayes' rule, we obtain the beliefs described above.

Part Four (Sequential Rationality): We now argue that a liquidity investor does not have

a profitable deviation. Because he is restricted to sending immediate-or-cancel orders, he is limited to buying at the ask $s^*_{LOB}/2$ or selling at the bid $-s^*_{LOB}/2$. Given all information available to him, the expected value of a share is zero. Thus, because he derives a utility bonus of β from buying (selling) exactly one share, he maximizes his expected profits by sending an order to buy (sell) one share with limit price β ($-\beta$), which is precisely what he does in equilibrium. In addition, because we assume

$$\beta \geq \frac{\lambda X}{1-\lambda+\lambda X} \geq \frac{s^*_{LOB}}{2},$$

this order results in a trade.

We now argue that an information investor does not have a profitable deviation. As above, he is limited to buying at the ask $s^*_{LOB}/2$ or selling at the bid $-s^*_{LOB}/2$. First, suppose that he fails to learn v. Then, given all information available to him, the expected value of a share is zero. As a result, he loses $s^*_{LOB}/2$ in expectation for every share he trades, so it is indeed optimal to submit no orders. Second, suppose that he learns $v = 1$ ($v = -1$). Then he earns the profit $1 - s^*_{LOB}/2 > 0$ for each share he buys (sells), and it is therefore in his interest to buy (sell) as many shares as possible. If he sends orders to y exchanges, then he expects to receive the following number of fills:

$$F_{LOB}(y) = yp_I + y(1-p_I)^y + (1-p_H)(y-1)yp_I(1-p_I)^{y-1}.$$

By Lemma 1, this function is weakly increasing on the domain of positive integers, so it is indeed optimal to submit orders to buy (sell) to each exchange. He then obtains $F(X) = X_I$ fills in expectation. Finally, given the above, his expected profits conditional on a choice of r are

$$rX_I\left(1 - \frac{s^*_{LOB}}{2}\right) - c(r).$$

Therefore, (2) implies the optimality of choosing research intensity r^*_{LOB}.

We now argue that the liquidity provider does not have a profitable deviation. First, we argue that she behaves optimally after one or more trades occur. Given the behavior of the other traders, exactly one trade occurs when the investor is liquidity-motivated. However, multiple trades may occur when the investor is information-motivated. Therefore, conditional on one or more trades having occurred at time t, the liquidity provider expects to lose $1 - s^*_{LOB}/2 > 0$ on any subsequent trades against her quotes. It is therefore indeed optimal for the liquidity provider to send cancellations after one or more trades occur. And for similar

reasons, the liquidity provider cannot profitably deviate by replenishing the LOB after a trade has occurred.

Second, we argue that the liquidity provider cannot profitably deviate before a trade occurs. These arguments are similar to those in Budish, Cramton, and Shim (2015, Proof of Proposition 1). As argued in Section III.A of the main article, equation (1) implies that she earns zero expected profits in the equilibrium. It therefore remains to show that she does not possess a deviation that would yield positive expected profits. If the liquidity provider deviates, then the enforcer will immediately enter post-only orders to buy (sell) one share at $-s^*_{LOB}/2$ ($s^*_{LOB}/2$). Thus, any shares (or fractions of a share) that are quoted at a spread wider than s^*_{LOB} will receive none of the benefits (from liquidity investor orders), but might receive adverse selection costs (from information investor and sniper orders), relative to what transpires on path. Any shares (or fractions of a share) that are quoted at a spread narrower than s^*_{LOB} will receive smaller benefits and larger adverse selection costs. Any shares (or fractions of a share) that are quoted at s^*_{LOB} will receive at most a prorated fraction of the same benefits but at least a prorated fraction of the same adverse selection costs. It is also not profitable to deviate by canceling her post-only orders because that would also lead to zero profits. Finally, it is not possible to deviate by initiating any trades because the model restricts her to post-only orders.

The remaining liquidity providers (including the enforcer) also do not have profitable deviations. They earn zero profits in equilibrium, and none of them possesses a deviation that would yield positive expected profits for reasons similar to those discussed in the previous paragraph. Any shares (or fractions of a share) that are quoted at a spread wider than s^*_{LOB} will receive none of the benefits (from liquidity investor orders), but might receive adverse selection costs (from information investor and sniper orders). Any shares (or fractions of a share) that are quoted at a spread narrower than s^*_{LOB} will receive smaller benefits and larger adverse selection costs. Any shares (or fractions of a share) that are quoted at s^*_{LOB} will receive at most a prorated fraction of the same benefits but at least a prorated fraction of the same adverse selection costs. Finally, it is not possible to deviate by initiating any trades because the model restricts them to post-only orders.

We now argue that snipers do not have profitable deviations. Given their restriction to immediate-or-cancel orders, snipers are limited to buying at the ask $s^*_{LOB}/2$ or selling at the bid $-s^*_{LOB}/2$, so long as those quotes are in place. First, we argue that they behave optimally in a time period in which two or more trades occur. Given their beliefs, the expected value of the security, conditional on trades occurring at the ask (bid) at two or more exchanges,

is 1 (−1). Thus, conditional on this event, a sniper earns the profit $1 - s^*_{LOB}/2 > 0$ for each share she buys (sells). It is therefore optimal for snipers to respond by sending orders to buy (sell).

Second, we argue that snipers behave optimally in a time period in which exactly one trade occurs. Suppose that $X = 1$. Given their beliefs, the expected value of the security, conditional on a trade occurring at the ask (bid) at exactly one exchange, is (the additive inverse of)

$$0 < \frac{\lambda r^*_{LOB}}{1 - \lambda + \lambda r^*_{LOB}} = \frac{s^*_{LOB}}{2}.$$

Next, suppose that $X \geq 2$. Given their beliefs, the expected value of the security, conditional on a trade occurring at the ask (bid) at exactly one exchange, is (the additive inverse of)

$$0 < \frac{\lambda r^*_{LOB} X p_I (1 - p_I)^{X-1}}{1 - \lambda + \lambda r^*_{LOB} X p_I (1 - p_I)^{X-1}} < \frac{s^*_{LOB}}{2}.$$

In either case, conditional on this event, a sniper expects to lose or break even on each trade she makes, and so has no profitable deviations.

Third, we argue that snipers cannot profitably deviate before a trade occurs. Given their restriction to immediate-or-cancel orders, it is not possible for them to deviate by attempting to provide liquidity. It is also not profitable for them to deviate by initiating trades at such times. Any such trades would result in an expected loss of $s^*_{LOB}/2$ per trade, and would lead the liquidity provider to cancel her remaining quotes, thereby also preventing the deviating sniper from completing any future trades. □

Proof of Corollary 1: As before, r^*_{LOB} is characterized as the fixed point of the correspondence

$$R_{LOB}(\hat{r}) \in \arg\max_{r \in [0,1]} \left\{ \frac{(1 - \lambda) r X_I}{1 - \lambda + \lambda \hat{r}(X_I + X_S)} - c(r) \right\}.$$

Other things equal, the argmax is weakly increasing in X_I and weakly decreasing in X_S. Differentiating the expressions given in Proposition 1, we find that X_I is weakly decreasing in p_H and X_S is constant in p_H. These observations imply that, other things equal, the argmax is weakly decreasing in p_H. It is also weakly decreasing in $\hat{r}$. By combining the previous two observations, we conclude that the fixed point, r^*_{LOB}, is weakly decreasing in p_H.

Differentiating the expression for the spread given in (1), we find that, other things equal, it is weakly increasing in X_I, weakly increasing in X_S, and weakly increasing in r^*_{LOB}. As

above, X_I is weakly decreasing in p_H and X_S is constant in p_H. Moreover, as established above, r^*_{LOB} is weakly decreasing in p_H. By combining these observations, we conclude that s^*_{LOB} is weakly decreasing in p_H. □

Proof of Proposition 2: An initial observation is that the combination of (BB–1), (BB–2), and (IR–H) is equivalent to the following single constraint:[2]

$$\frac{1-\lambda}{2}y(B)+\frac{1-\lambda}{2}y(S)+\frac{\lambda r}{2}\left[y(1)+z(1)\right]+\frac{\lambda r}{2}\left[y(-1)-z(-1)\right]+\lambda(1-r)y(0)\leq 0. \tag{BB}$$

The remainder of the proof consists of two parts. First, we show that the set defined in the proposition constitutes an outer bound for $\mathcal{F}$. Second, we show that it constitutes an inner bound for $\mathcal{F}$.

Part One (Outer Bound). We begin by rewriting (O) as

$$r\in\arg\max_{\hat{r}\in[0,1]}\left\{\frac{\hat{r}}{2}\left[y(1)+z(1)\right]+\frac{\hat{r}}{2}\left[y(-1)-z(-1)\right]+(1-\hat{r})y(0)-c(\hat{r})\right\},$$

which implies[3]

$$r\left(\frac{y(1)+z(1)}{2}+\frac{y(-1)-z(-1)}{2}-y(0)\right)\geq rc'(r). \tag{IA6}$$

Next, we rewrite (W) to obtain

$$\begin{aligned} w&=\frac{1}{2}\left[y(B)+\beta\mathbb{1}\{z(B)=1\}\right]+\frac{1}{2}\left[y(S)+\beta\mathbb{1}\{z(S)=-1\}\right]\\ &\leq\beta+\frac{1}{2}y(B)+\frac{1}{2}y(S).\end{aligned}$$

[2]To elaborate: it is straightforward to verify that (BB) is implied by (BB–1), (BB–2), and (IR–H). Moreover, if $y(\theta)$ and $z(\theta)$ satisfy (BB), then (BB–1), (BB–2), and (IR–H) are satisfied by choosing $\{y_h\}_{h\in\mathcal{H}}$ and $\{z_h\}_{h\in\mathcal{H}}$ so that for all $\theta\in\Theta$, $\sum_{h\in\mathcal{H}}y_h(\theta)=-y(\theta)$ and $\sum_{h\in\mathcal{H}}z_h(\theta)=-z(\theta)$.

[3]Define $\Xi=\left[\frac{y(1)+z(1)}{2}+\frac{y(-1)-z(-1)}{2}-y(0)\right]$. By assumption, $c(\cdot)$ is C^1. Therefore, any solution to the maximization problem in (O) must satisfy one of the three conditions (*i*) $r=0$ and $c'(0)\geq\Xi$, (*ii*) $c'(r)=\Xi$, or (*iii*) $r=1$ and $c'(1)\leq\Xi$. In any of these three cases, the claimed inequality holds.

Applying (BB), this becomes

$$\begin{aligned} w &\leq \beta - \frac{\lambda r}{2(1-\lambda)}\left[y(1)+z(1)\right] - \frac{\lambda r}{2(1-\lambda)}\left[y(-1)-z(-1)\right] - \frac{\lambda(1-r)}{1-\lambda}y(0) \\ &= \beta - \frac{\lambda r}{1-\lambda}\left(\frac{y(1)+z(1)}{2} + \frac{y(-1)-z(-1)}{2} - y(0) + 2y(0)\right). \end{aligned}$$

Applying (IA6) yields

$$w \leq \beta - \frac{\lambda}{1-\lambda} r\left[c'(r) + 2y(0)\right].$$

Applying also (IR–I) for $\theta = 0$, which requires $y(0) \geq 0$, this becomes

$$w \leq \beta - \frac{\lambda}{1-\lambda} r c'(r). \tag{IA7}$$

This establishes the desired upper bound on w. The lower bound $w \geq 0$ follows immediately from (W) and (IR–I) for $\theta \in \{B, S\}$.

Part Two (Inner Bound). For this part of the proof, we argue that any element of the set defined in the proposition can be implemented by a contract satisfying all of the constraints of $\mathcal{F}$. In what follows, we use $r_{\max}$ to denote the largest r such that there exists a w for which $(r, w) \in \mathcal{F}$. It is defined implicitly by

$$(1-\lambda)\beta = \lambda r_{\max} c'(r_{\max}).$$

Suppose $r \in [0, r_{\max}]$ and suppose $w \in \left[0, \beta - \frac{\lambda}{1-\lambda} r c'(r)\right]$. Let

$$\begin{aligned} y(B) = y(S) &= w - \beta \\ z(B) = -z(S) &= 1 \\ y(1) = -y(-1) &= 0 \\ z(1) = -z(-1) &= c'(r) \\ y(0) &= 0 \\ z(0) &= 0. \end{aligned}$$

We now argue that these contracts satisfy the constraints (W), (BB), (IR–I), and (O):

(W) Plugging in, $\frac{1}{2}u(y(B), z(B)|B) + \frac{1}{2}u(y(S), z(S)|S) = w$.

(BB) Plugging in, $\frac{1-\lambda}{2}y(B) + \frac{1-\lambda}{2}y(S) + \frac{\lambda r}{2}[y(1)+z(1)] + \frac{\lambda r}{2}[y(-1)-z(-1)] + \lambda(1-r)y(0) = (1-\lambda)(w-\beta) + \lambda r c'(r)$, which is nonpositive by assumption.

(IR–I) First, $u(y(B), z(B)|B) = u(y(S), z(S)|S) = w$, which is nonnegative by assumption. Second, $u(y(1), z(1)|1) = u(y(-1), z(-1)|-1) = c'(r) \geq 0$. Third, $u(y(0), z(0)|0) = 0$.

(O) Plugging in, (O) becomes $r \in \arg\max_{\hat{r}\in[0,1]} \{\hat{r}c'(r) - c(\hat{r})\}$. By convexity of $c(\cdot)$, the optimality of conducting research with intensity r follows from checking the first-order condition. □

Proof of Corollary 2: In the LOB equilibrium, liquidity investor welfare is determined by the spread through $w^*_{LOB} = \beta - s^*_{LOB}/2$. Therefore, by Proposition 2, the LOB equilibrium outcome lies on the frontier of the feasible set if and only if

$$\lambda r^*_{LOB} c'(r^*_{LOB}) = (1-\lambda)\frac{s^*_{LOB}}{2}. \tag{IA8}$$

As before, r^*_{LOB} is characterized by the fixed point of the correspondence

$$R_{LOB}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda)rX_I}{1-\lambda+\lambda\hat{r}(X_I+X_S)} - c(r) \right\},$$

where X_I and X_S are as defined in the statement of Proposition 1. The assumption that $c'(0) < X_I$ ensures that at $\hat{r} = 0$, the maximization problem on the right-hand side of the expression above for $R_{LOB}(\hat{r})$ does not have a solution at zero. Consequently, zero is not a fixed point of that correspondence and so $r^*_{LOB} > 0$. Thus, at the value of $\hat{r}$ that is the fixed point of the correspondence, all solutions to the maximization problem on the right-hand side of the expression above are either a corner solution at one or an interior solution. In either case,

$$c'(r^*_{LOB}) \leq \frac{(1-\lambda)X_I}{1-\lambda+\lambda r^*_{LOB}(X_I+X_S)}.$$

In addition, from equation (1) we have

$$s^*_{LOB} = \frac{2\lambda r^*_{LOB}(X_I+X_S)}{1-\lambda+\lambda r^*_{LOB}(X_I+X_S)}.$$

We then compare these expressions to the condition for being on the frontier of the feasible set, equation (IA8). Given that $\lambda \in (0,1)$ and $r^*_{LOB} > 0$, the LOB equilibrium outcome is on

the frontier only if $X_S = 0$. Restating the expression for X_S, we obtain

$$\begin{aligned} X_S &= X(1-p_I) - X(1-p_I)^X - (X-1)Xp_I(1-p_I)^{X-1} \\ &= X(1-p_I)\left[\sum_{x=2}^{X-1}\binom{X-1}{x}p_I^x(1-p_I)^{X-1-x}\right]. \end{aligned}$$

Given that $p_I \geq 0.5$, $X_S = 0$ only if either $p_I = 1$ or $X \leq 2$ (or both). □

Proof of Proposition 3: The proof consists of four parts. First, we argue that for any value of q, there exists a unique solution to (4) and (5). Second, we describe the beliefs that are part of the equilibrium. (The strategies are analogous to those specified in Section III.A of the main article.) Third, we verify that these beliefs are consistent with Bayes' rule. Fourth, we verify that strategies are sequentially rational given beliefs.

Part One (Existence and Uniqueness): Plugging (4) into (5), r^*_{qND} is characterized as a fixed point of the correspondence

$$R_{qND}(\hat{r}) = \arg\max_{r\in[0,1]}\left\{\frac{(1-\lambda)rX_{qND}}{1-\lambda+\lambda\hat{r}X_{qND}} - c(r)\right\}.$$

As in the proof of Proposition 1, r^*_{qND} uniquely exists, and s^*_{qND} can be uniquely derived from r^*_{qND} and (4).

Part Two (Description): Equilibrium strategies are analogous to those described in Section III.A of the main article. The only differences are that s^*_{qND} and r^*_{qND} assume the roles played by s^*_{LOB} and r^*_{LOB}.

If traders use these strategies, then X_{qND} does indeed represent the expected number of trades made by an information investor conditional on learning the value of the security. To see this, suppose that the information investor submits orders to all X exchanges, and suppose that x of these orders receive only a delay of δ_{ND}, while the other $X-x$ also receive an additional delay drawn from F_{ND}. We consider two cases: $x = 0$ and $x > 0$. First, suppose $x = 0$, so that all orders receive a delay drawn from F_{ND}. The information investor always obtains a fill for his first order to be processed. However, F_{ND} is so diffuse that the probability of obtaining another fill after the first is infinitesimally small. Therefore, conditional on being in this case, the expected number of fills is one. Second, suppose $x > 0$. We begin by determining the expected number of fills for the x orders that do not receive a delay drawn from F_{ND}. The information investor receives a fill for any order that achieves the minimum latency of ε, which occurs with probability p_I. In the event that all

of his orders have latency 3ε, then he obtains fills for all orders. Otherwise, he receives a fill for an order with latency 3ε only if the liquidity provider's corresponding cancellation also has latency 3ε, which occurs with probability $1-p_H$. Therefore, the number of these x orders that the information investor expects to have filled is $xp_H(1-p_I)^x + x\left[1-p_H(1-p_I)\right]$. Furthermore, the information investor does not expect any more fills: F_{ND} is so diffuse that the probability of obtaining a fill for any of the $X-x$ orders receiving a delay drawn from F_{ND} is infinitesimally small. Collecting all of these observations, and considering the fact that $x \sim \text{Binom}(X,q)$, an information investor's expected number of fills, conditional only on learning the value of the security, is

$$X_{qND} = q^X + \sum_{x=1}^{X} \binom{X}{x} (1-q)^x q^{X-x} \left(xp_H(1-p_I)^x + x\left[1-p_H(1-p_I)\right]\right),$$

as claimed.

To complete the description of the WPBE, it remains to specify beliefs. The relevant beliefs are what the HFTs believe in the time period in which the first trade occurs about who initiated those trades. We consider the cases $X = 1$ and $X \geq 2$ separately.

- First, suppose $X = 1$. If one trade occurs at the ask (bid) in that time period, then HFTs believe it to have been initiated either by a liquidity investor with a buying (selling) motive, with probability

  $$\frac{1-\lambda}{1-\lambda+\lambda r^*_{qND}},$$

  or by an information investor who learned $v = 1$ ($v = -1$), with probability

  $$\frac{\lambda r^*_{qND}}{1-\lambda+\lambda r^*_{qND}}.$$

- Second, suppose $X \geq 2$. If exactly one trade occurs at the ask (bid) in that time period, then HFTs believe it to have been initiated either by a liquidity investor with a buying (selling) motive, with probability

  $$\frac{1-\lambda}{1-\lambda+\lambda r^*_{qND}\left(q^X + \sum_{x=1}^{X}\binom{X}{x}(1-q)^x q^{X-x}\left[xp_I(1-p_I)^{x-1}\right]\right)},$$

or by an information investor who learned $v = 1$ ($v = -1$), with probability

$$\frac{\lambda r^*_{qND}\left(q^X + \sum_{x=1}^{X}\binom{X}{x}(1-q)^x q^{X-x}\left[x p_I (1-p_I)^{x-1}\right]\right)}{1-\lambda+\lambda r^*_{qND}\left(q^X + \sum_{x=1}^{X}\binom{X}{x}(1-q)^x q^{X-x}\left[x p_I (1-p_I)^{x-1}\right]\right)}.$$

If two or more trades occur at the ask (bid) in that time period, then HFTs believe them to have been initiated by an information investor who learned $v = 1$ ($v = -1$), with probability one.

Part Three (Consistency): Given equilibrium strategies, we have the following. On the one hand, the investor is a liquidity investor with a buying (selling) motive with probability $(1-\lambda)/2$. Conditional on there being such an investor, in the time period in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability one. On the other hand, the investor is an information investor who learns $v = 1$ ($v = -1$) with probability $\lambda r^*_{qND}/2$. If $X = 1$ and conditional on there being such an investor, then in the time period in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability one. If $X \geq 2$ and conditional on there being such an investor, then in the time period in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability $q^X + \sum_{x=1}^{X}\binom{X}{x}(1-q)^x q^{X-x}\left[x p_I (1-p_I)^{x-1}\right]$, and two or more such trades occur with probability $1 - \left(q^X + \sum_{x=1}^{X}\binom{X}{x}(1-q)^x q^{X-x}\left[x p_I (1-p_I)^{x-1}\right]\right)$. When Bayes' rule can be applied, we obtain the beliefs described above.

Part Four (Sequential Rationality): That liquidity investors, information investors, the liquidity provider, and the remaining liquidity providers (including the enforcer) do not have profitable deviations is as in the proof of Proposition 1.[4]

Snipers also have no deviations that would yield positive expected profits. As in the proof of Proposition 1, it is not profitable to deviate by triggering any trades before a trade occurs. In addition, it is not possible to obtain fills after a trade occurs because the delay $\delta_{ND} > 2\varepsilon$ applied to noncancellations means that the liquidity provider always cancels her quotes before any sniper can react to trade against them: even if the liquidity provider's cancellation obtains the maximum latency of 3ε, it is still processed before any sniper orders that obtain the minimum latency of ε. Therefore, no aggressive-side order anticipation can occur in equilibrium. □

[4]The primary difference lies in establishing that it is optimal for information investors to send orders to all X exchanges. The argument uses Lemma 2 in the same way that Proposition 1 uses Lemma 1.

Proof of Corollary 3: As before, define

$$X_{qND} = q^X + \sum_{x=1}^{X} \binom{X}{x} (1-q)^x q^{X-x} \left(x p_H (1-p_I)^x + x \left[1 - p_H (1-p_I) \right] \right),$$

where as before, the index x is interpreted as the number of exchanges at which the information investor's orders receive only the fixed delay δ_{ND}. The number of fills that the information investor expects to receive conditional on a realization of x is weakly increasing in x. The effect of an increase in q is to shift the distribution over x downward in the sense of first-order stochastic dominance. Therefore, X_{qND} is weakly decreasing in q.

As before, r^*_{qND} is characterized as the fixed point of the correspondence

$$R_{qND}(\hat{r}) = \arg\max_{r \in [0,1]} \left\{ \frac{(1-\lambda) r X_{qND}}{1 - \lambda + \lambda \hat{r} X_{qND}} - c(r) \right\}.$$

Applying Topkis' Theorem, $R_{qND}(\hat{r})$ is weakly increasing in X_{qND}, and therefore is weakly decreasing in q. Because the correspondence is also weakly decreasing in $\hat{r}$, we obtain that r^*_{qND} is weakly decreasing in q.

Differentiating the expression for the spread given in (4), we find that, other things equal, it is weakly increasing in r^*_{qND} and weakly decreasing in q. Moreover, as established above, r^*_{qND} is weakly decreasing in q. By combining these observations, we conclude that s^*_{qND} is weakly decreasing in q. □

Proof of Corollary 4: As before, define

$$X_{qND} = q^X + \sum_{x=1}^{X} \binom{X}{x} (1-q)^x q^{X-x} \left(x p_H (1-p_I)^x + x \left[1 - p_H (1-p_I) \right] \right).$$

Also, recall that

$$\begin{aligned}
X_{0ND} &= Xp_H(1-p_I)^X + X[1-p_H(1-p_I)] \\
&= X(1-p_I)^X - X(1-p_H)(1-p_I)^X + X[1-p_H(1-p_I)] \\
&= X(1-p_I)^X + \sum_{x=1}^{X}[x+(X-x)(1-p_H)]\binom{X}{x}p_I^x(1-p_I)^{X-x} \\
&\geq X(1-p_I)^X + [1+(X-1)(1-p_H)]Xp_I(1-p_I)^{X-1} + \sum_{x=2}^{X} x\binom{X}{x}p_I^x(1-p_I)^{X-x} \\
&= Xp_I + X(1-p_I)^X + (1-p_H)(X-1)Xp_I(1-p_I)^{X-1} \\
&= X_I.
\end{aligned}$$

Moreover, $X_{1ND} = 1 \leq X_I$. Lastly, X_{qND} is continuous as a function of q. Thus, by the intermediate value theorem, there exists $\hat{q} \in [0,1]$ for which $X_{\hat{q}ND} = X_I$. Next, define $s^*(\Omega)$ and $r^*(\Omega)$ as the solution to the system

$$s^* = \frac{2\lambda r^*(X_I+\Omega)}{1-\lambda+\lambda r^*(X_I+\Omega)} \tag{IA9}$$

$$r^* \in \arg\max_{r\in[0,1]} \left\{ rX_I\left(1-\frac{s^*}{2}\right) - c(r) \right\}. \tag{IA10}$$

Notice that $s^*_{\hat{q}ND}$ and $r^*_{\hat{q}ND}$ correspond to $s^*(\Omega)$ and $r^*(\Omega)$ evaluated at $\Omega = 0$. Similarly, s^*_{LOB} and r^*_{LOB} correspond to $s^*(\Omega)$ and $r^*(\Omega)$ evaluated at $\Omega = X_S \geq 0$. From equation (IA9), s^* is, other things equal, (i) weakly increasing in r^* and (ii) weakly increasing in Ω. Furthermore, from equation (IA10), r^* is, other things equal, (i) weakly decreasing in s^* and (ii) unaffected by Ω. By combining these observations, we establish that $s^*(\Omega)$ is weakly increasing and that $r^*(\Omega)$ is weakly decreasing. Therefore, we conclude that $s^*_{\hat{q}ND} \leq s^*_{LOB}$ and $r^*_{\hat{q}ND} \geq r^*_{LOB}$, as desired. □

Proof of Proposition 4: The proof consists of four parts. First, we argue that there exists a unique solution to (6) and (7). Second, we describe the strategies and beliefs of the equilibrium. Third, we verify that these beliefs are consistent with Bayes' rule. Fourth, we verify that strategies are sequentially rational given beliefs.

Part One (Existence and Uniqueness): From (6) and (7), r^*_{FBA} is characterized as a fixed

point of the correspondence

$$R_{FBA}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda)rX}{1-\lambda+\lambda\hat{r}X} - c(r) \right\}.$$

As in the proof of Proposition 1, r^*_{FBA} uniquely exists, and s^*_{FBA} can be uniquely derived from r^*_{FBA} and (6).

Part Two (Description): In terms of the quantities s^*_{FBA} and r^*_{FBA} characterized by Proposition 4, the equilibrium strategies are as follows:

- *Investor.* If he is a liquidity investor with a buying (selling) motive, then he sends to his home exchange a noncompetitive order to buy (sell) one share at the price β ($-\beta$).

  If he is an information investor, then he conducts research with intensity r^*_{FBA}. If he learns that the value of the security is $v = 1$ ($v = -1$), then he sends to each exchange a noncompetitive order to buy (sell) one share at the price 1 (-1). He sends no orders if he does not learn v.

- *Liquidity providers.* One liquidity provider is active on the equilibrium path and is referred to as "the liquidity provider" in what follows. At time zero, she sends to each exchange a competitive order to buy one share at the bid $-s^*_{FBA}/2$ and another to sell one share at the ask $s^*_{FBA}/2$. If in any batch interval one or more trades occur, then she sends cancellations for all of her remaining orders, doing so in the next batch interval.

  A second liquidity provider who is inactive on path but may be active off path is referred to as "the enforcer." If in some batch interval prior to which no trade has occurred the competitive schedules at some exchange are reported to have consisted of anything other than a competitive order to buy one share at $-s^*_{FBA}/2$ and a competitive order to sell one share at $s^*_{FBA}/2$, then she sends such orders to that exchange, doing so in the next batch interval.

  The remaining liquidity providers remain completely inactive both on and off path.

- *Snipers.* If in any batch interval trades occur at the ask (bid) at two or more exchanges, then each sniper sends to all other exchanges a noncompetitive order to buy (sell) one share at the price 1 (-1), doing so in the next batch interval.

If traders use these strategies, then an information investor makes X trades conditional on learning the value of the security. To see this, note that the processing times of his orders

are separated by at most 2ε. Because the batch length is assumed to be infinitely longer than ε, it is with only infinitesimal probability that his orders are processed in different batch intervals. And if his orders are processed in the same batch interval, then all X are filled before any of the other traders can react.

To complete the description of the WPBE, it remains to specify beliefs. The relevant beliefs are what the HFTs believe after the batch interval in which the first trade occurs about who initiated those trades (i.e., who submitted the noncompetitive orders involved in those trades). We consider the cases $X = 1$ and $X \geq 2$ separately.

- First, suppose $X = 1$. If one trade occurs at the ask (bid) in that batch interval, then HFTs believe it to have been initiated either by a liquidity investor with a buying (selling) motive, with probability
$$\frac{1-\lambda}{1-\lambda+\lambda r^*_{FBA}},$$
or by an information investor who learned $v = 1$ ($v = -1$), with probability
$$\frac{\lambda r^*_{FBA}}{1-\lambda+\lambda r^*_{FBA}}.$$

- Second, suppose $X = 2$. If exactly one trade occurs at the ask (bid) in that batch interval, then HFTs believe it to have been initiated by a liquidity investor with a buying (selling) motive, with probability one. If $X \geq 2$ and two or more trades occur at the ask (bid) in that batch interval, then HFTs believe them to have been initiated by an information investor who learned $v = 1$ ($v = -1$), with probability one.

Part Three (Consistency): Given equilibrium strategies, we have the following. On the one hand, the investor is a liquidity investor with a buying (selling) motive with probability $(1-\lambda)/2$. Conditional on there being such an investor, in the batch interval in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability one. On the other hand, the investor is an information investor who learns $v = 1$ ($v = -1$) with probability $\lambda r^*_{FBA}/2$. If $X = 1$ and conditional on there being such an investor, then in the batch interval in which the first trade occurs, there occurs exactly one trade initiated by him at the ask (bid) with probability one. If $X \geq 2$ and conditional on there being such an investor, then in the batch interval in which the first trade occurs, there occur two or more trades initiated by him at the ask (bid) with probability one. Applying Bayes' rule, we obtain the beliefs described above.

Part Four (Sequential Rationality): That liquidity investors, information investors, the liquidity provider, and the remaining liquidity providers (including the enforcer) do not have profitable deviations is as in the proof of Proposition 1.

Snipers also have no deviations that would yield positive expected profits. As in the proof of Proposition 1, it is not profitable to deviate by triggering any trades before a trade occurs. In addition, it is not possible to obtain fills after a trade occurs because batching means that the liquidity provider's cancellations will be processed in the same batch interval as the sniper orders. Therefore, no aggressive-side order anticipation can occur in equilibrium. □

Proof of Proposition 5: The proof consists of two parts. First, we show that for all $q \in [0,1]$, $s^*_{FBA} \geq s^*_{qND}$ and $r^*_{FBA} \geq r^*_{qND}$. Second, we show that $s^*_{FBA} \geq s^*_{LOB}$ and $r^*_{FBA} \geq r^*_{LOB}$.

Part One ($s^*_{FBA} \geq s^*_{qND}$ and $r^*_{FBA} \geq r^*_{qND}$). Define $s^*(\Omega)$ and $r^*(\Omega)$ as the solution to the system

$$s^* = \frac{2\lambda r^* \Omega}{1 - \lambda + \lambda r^* \Omega} \tag{IA11}$$

$$r^* \in \arg\max_{r \in [0,1]} \left\{ r\Omega \left(1 - \frac{s^*}{2}\right) - c(r) \right\}. \tag{IA12}$$

Notice that s^*_{FBA} and r^*_{FBA} correspond to $s^*(\Omega)$ and $r^*(\Omega)$ evaluated at $\Omega = X$. Similarly, s^*_{qND} and r^*_{qND} correspond to $s^*(\Omega)$ and $r^*(\Omega)$ evaluated at $\Omega = X_{qND}$. We define X_{qND} in Proposition 3, and it is not more than X. It therefore suffices to show that $s^*(\Omega)$ and $r^*(\Omega)$ are both weakly increasing in Ω. To do so, notice that $r^*(\Omega)$ is characterized as the fixed point of the correspondence

$$R(\hat{r}, \Omega) = \arg\max_{r \in [0,1]} \left\{ \frac{(1-\lambda) r \Omega}{1 - \lambda + \lambda \hat{r} \Omega} - c(r) \right\}.$$

By Topkis' Theorem, $R(\hat{r}, \Omega)$ is weakly increasing in Ω. Because the correspondence is also weakly decreasing in $\hat{r}$, this fact implies that $r^*(\Omega)$ is weakly increasing. Furthermore, from equation (IA11), s^* is, other things equal, (*i*) weakly increasing in r^* and (*ii*) weakly increasing in Ω. Therefore, we also have that $s^*(\Omega)$ is weakly increasing.

Part Two ($s^*_{FBA} \geq s^*_{LOB}$ and $r^*_{FBA} \geq r^*_{LOB}$). Define $s^*(\Omega)$ and $r^*(\Omega)$ as the solution to the

system

$$s^* = \frac{2\lambda r^*[X\Omega + (X_I + X_S)(1-\Omega)]}{1-\lambda+\lambda r^*[X\Omega + (X_I + X_S)(1-\Omega)]} \tag{IA13}$$

$$r^* \in \arg\max_{r\in[0,1]} \left\{ r[X\Omega + X_I(1-\Omega)]\left(1 - \frac{s^*}{2}\right) - c(r) \right\}. \tag{IA14}$$

Notice that s^*_{FBA} and r^*_{FBA} correspond to $s^*(\Omega)$ and $r^*(\Omega)$ evaluated at $\Omega = 1$. Similarly, s^*_{LOB} and r^*_{LOB} correspond to $s^*(\Omega)$ and $r^*(\Omega)$ evaluated at $\Omega = 0$. It therefore suffices to show that $s^*(\Omega)$ and $r^*(\Omega)$ are both weakly increasing in Ω. To do so, notice that $r^*(\Omega)$ is characterized as the fixed point of the correspondence

$$R(\hat{r}, \Omega) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda) r[X\Omega + X_I(1-\Omega)]}{1-\lambda+\lambda\hat{r}[X\Omega + (X_I + X_S)(1-\Omega)]} - c(r) \right\}.$$

By Topkis' Theorem, $R(\hat{r}, \Omega)$ is weakly increasing in Ω. Because the correspondence is also weakly decreasing in $\hat{r}$, this fact implies that $r^*(\Omega)$ is weakly increasing. Furthermore, from equation (IA13), s^* is, other things equal, (i) weakly increasing in r^* and (ii) weakly increasing in Ω. Therefore, we also have that $s^*(\Omega)$ is weakly increasing. □

Proof of Corollary 5: In the equilibria of the LOB, of NDs, and of FBAs, liquidity investor welfare is determined by the spread through $w^* = \beta - s^*/2$. Therefore, by Proposition 2, an equilibrium outcome of one of these trading mechanisms lies on the frontier of the feasible set if and only if the equilibrium spread, s^*, is related to the equilibrium research intensity, r^*, through the equation

$$\lambda r^* c'(r^*) = (1-\lambda)\frac{s^*}{2}. \tag{IA15}$$

As before, the equilibrium research intensities r^*_{LOB}, r^*_{qND}, and r^*_{FBA} are characterized, respectively, by the fixed points of the correspondences

$$R_{LOB}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda) r X_I}{1-\lambda+\lambda\hat{r}(X_I + X_S)} - c(r) \right\}$$

$$R_{qND}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda) r X_{qND}}{1-\lambda+\lambda\hat{r}X_{qND}} - c(r) \right\}$$

$$R_{FBA}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda) r X}{1-\lambda+\lambda\hat{r}X} - c(r) \right\}.$$

where X_I and X_S are as defined in the statement of Proposition 1 and X_{qND} is as defined in the statement of Proposition 3.

The assumption that $c'(1) \geq \dfrac{(1-\lambda)X}{1-\lambda+\lambda X}$ ensures that at $\hat{r} = 1$, none of the maximization problems on the right-hand sides above has a corner solution at one. Consequently, at the values of $\hat{r}$ that are the respective fixed points of these correspondences, each of the solutions to the maximization problems on the right-hand sides above is either a corner solution at zero or an interior solution. We have the following equations in either case:

$$r^*_{LOB}c'(r^*_{LOB}) = \frac{(1-\lambda)r^*_{LOB}X_I}{1-\lambda+\lambda r^*_{LOB}(X_I+X_S)}$$
$$r^*_{qND}c'(r^*_{qND}) = \frac{(1-\lambda)r^*_{qND}X_{qND}}{1-\lambda+\lambda r^*_{qND}X_{qND}}$$
$$r^*_{FBA}c'(r^*_{FBA}) = \frac{(1-\lambda)r^*_{FBA}X}{1-\lambda+\lambda r^*_{FBA}X}.$$

In addition, from equations (1), (4), and (6), we have

$$s^*_{LOB} = \frac{2\lambda r^*_{LOB}(X_I+X_S)}{1-\lambda+\lambda r^*_{LOB}(X_I+X_S)}$$
$$s^*_{qND} = \frac{2\lambda r^*_{qND}X_{qND}}{1-\lambda+\lambda r^*_{qND}X_{qND}}$$
$$s^*_{FBA} = \frac{2\lambda r^*_{FBA}X}{1-\lambda+\lambda r^*_{FBA}X}.$$

Comparing these expressions to the condition for being on the frontier of the feasible set, equation (IA15), we find that the qND outcome, for all $q \in [0,1]$, is on the frontier. Similarly, the FBA outcome is on the frontier. Finally, the LOB outcome is on the frontier if $X_S = 0$, which is the case if either $X \leq 2$ or $p_I = 1$ (or both). □

Proof of Corollary 6: First, note that, under the corollary's assumption that $c'(1) \geq 1-\lambda$, $r_{\min} = \min\left\{r \in [0,1] : c'(r) \geq \dfrac{1-\lambda}{1-\lambda+\lambda r}\right\}$ is indeed well-defined. Indeed, the assumption implies that the set $\left\{r \in [0,1] : c'(r) \geq \dfrac{1-\lambda}{1-\lambda+\lambda r}\right\}$ is nonempty (as it contains $r = 1$). And because $c(r)$ is assumed to be continuously differentiable, that set is compact and thus does have a minimum value. Suppose now that (r^*, w^*) is an outcome on the frontier of the feasible set where $r^* \geq r_{\min}$. By Proposition 2, $w^* = \beta - \dfrac{\lambda}{1-\lambda} r^* c'(r^*)$ and $w^* \geq 0$.

Proof of Claim (ii). Consider FBAs with $X'' = \dfrac{(1-\lambda)c'(r^*)}{1-\lambda-\lambda r^* c'(r^*)}$ exchanges. Note that the right-hand side is greater than or equal to one when $r^* = r_{\min}$ and is increasing in r^*, so we do indeed have $X'' \in \mathbb{R}_{\geq 1}$, as claimed. We wish to use Proposition 4 to characterize the equilibrium outcomes that prevail in this scenario. That proposition relies on the assumption $\beta \geq \dfrac{\lambda X}{1-\lambda+\lambda X}$ to characterize the equilibrium of FBAs with X exchanges. If $X'' > X$, then an analogous inequality it not guaranteed to hold. However, the only role of the inequality in the proposition's proof is to ensure that the equilibrium spread does not exceed 2β. We therefore suppose for the moment that this is the case in the present scenario, and we verify it later. By Proposition 4, FBAs with X'' exchanges lead to equilibrium research intensity that is characterized by the fixed point of the correspondence

$$\begin{aligned} R_{FBA}(\hat{r}) &= \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda)rX''}{1-\lambda+\lambda\hat{r}X''} - c(r) \right\} \\ &= \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda)rc'(r^*)}{1-\lambda-\lambda(r^*-\hat{r})c'(r^*)} - c(r) \right\}. \end{aligned}$$

The fixed point of this correspondence occurs when $r = \hat{r} = r^*$. Furthermore, the associated spread is

$$s^* = \frac{2\lambda r^* X''}{1-\lambda+\lambda r^* X''} = \frac{2\lambda}{1-\lambda} r^* c'(r^*),$$

which implies that liquidity investor welfare in this equilibrium is $\beta - s^*/2 = w^*$. It only remains to verify that $s^* \leq 2\beta$. However, this follows immediately from the fact that $s^* = 2(\beta - w^*)$ and the fact that $w^* \geq 0$. We conclude that the outcome (r^*, w^*) can be implemented by FBAs with X'' exchanges.

Proof of Claim (i). Next, choose X' to be any natural number for which

$$X' p_H (1-p_I)^{X'} + X'[1 - p_H(1-p_I)] \geq \frac{(1-\lambda)c'(r^*)}{1-\lambda-\lambda r^* c'(r^*)}.$$

Such a selection is possible because the left-hand side of the equation above diverges as $X' \to \infty$. Given that choice of X', choose $q \in [0,1]$ such that

$$q^{X'} + \sum_{x=1}^{X'} \binom{X'}{x} (1-q)^x q^{X'-x} \left(x p_H (1-p_I)^x + x\left[1 - p_H(1-p_I)\right] \right) = \frac{(1-\lambda)c'(r^*)}{1-\lambda-\lambda r^* c'(r^*)}.$$

Such a selection is possible by the intermediate value theorem because (*i*) the left-hand side of the equation above is continuous in q, (*ii*) the left-hand side exceeds the right-hand side

when $q = 0$ (by the definition of X'), (*iii*) the left-hand side evaluates to one when $q = 1$, and (*iv*) the right-hand side is greater than or equal to one, since it is greater than or equal to one when $r^* = r_{\min}$ and is increasing in r^*. It can then be shown through methods similar to those used above that qND with X' exchanges implements the outcome (r^*, w^*). □

II. Empirical Evidence of Random Latency

Order anticipation enabled by random latency lies at the heart of the model. Indeed, if latency were completely deterministic, then no order anticipation would take place,[5] the equilibrium spread would be invariant to changes in the speed of HFTs and the LOB would lead to an outcome identical to that of both FBAs and 0-ND. In this appendix, (*i*) we provide additional details concerning the sources of randomness in latency, (*ii*) we provide some statistics to quantify the extent of randomness, (*iii*) we conduct a simulation of the model to form a rough estimate of the amount of order anticipation allowed by the prevailing latency structure in modern markets, and (*iv*) we discuss some statistics pertaining to the amount of order anticipation arising in practice.

A. Sources of Randomness

A variety of factors contribute to latency in data networks (Bertsekas and Gallager (1992), Kay (2009), Ixia (2012), Corvil (2014), Deutsche Börse Group (2018)). Summarizing Kay (2009): (*i*) network interface delays for serialization occur as data are translated into packets (or vice versa) as they pass from one domain to another, (*ii*) signal propagation delays occur as packets travel across a physical distance, (*iii*) network processing delays occur as gateways, firewalls, routers, or switches determine how to treat a packet, (*iv*) router and switch delays occur as packets travel through them from an input port to a corresponding output port, and (*v*) queuing delays occur when packets from different input ports are queued for sequential transmission through the same output port.

Randomness in latency, known as *jitter*, may be stem from (*i*) variation in packet size, which influences network interface delays, (*ii*) variation in route path, which influences signal propagation delays, and (*iii*) variation in network congestion, which influences queuing

[5] If latency were completely deterministic, then it would be possible to evade order anticipation by coordinating the processing times of orders. In fact, even if exchanges differed in their latencies, such coordination could be achieved by using smart order routers such as Royal Bank of Canada's THOR (Aisen et al. (2015), Lewis (2014)). Rather than releasing orders simultaneously from a trading desk, THOR releases orders at slightly different times so as to coordinate their arrival times. Nevertheless, latency is not deterministic in practice. For example, as Lewis (2014) writes,

> In theory, the fastest travel time, from Katsuyama's desk in Manhattan to the BATS exchange in Weehawken, N.J., was about two milliseconds, and the slowest, from Katsuyama's desk to the Nasdaq exchange in Carteret, N.J., was around four milliseconds. *In practice, the times could vary much more than that, depending on network traffic, static and glitches in the equipment between any two points* [emphasis added].

delays.[6] For the within-exchange leg of communication, exchange architecture can interact with the aforementioned factors to influence the degree of randomness. (See Kirilenko and Lamacie (2015) and Menkveld and Zoican (2017) for a discussion of exchange architecture.) For the trader-to-exchange leg of communication, variation may also be caused by the weather, which affects signal propagation delays for radio communication.[7]

B. Evidence of Randomness

Three types of latency play a role in order anticipation: (*i*) trader-to-exchange latency, which is the amount of time between when a trader sends an order to an exchange and when the order is received by the exchange, (*ii*) within-exchange latency, which is the amount of time between when an order is received by the exchange and when it is processed, and (*iii*) exchange-to-trader latency, which is the amount of time between when an order is processed and when ensuing announcements reach traders.[8] All three components possess some randomness. Below, we report some statistics that quantify the degree of randomness present in two of these components.

Within-exchange. Exchanges require some amount of time to accept, process, and execute orders. Publicly available summary statistics concerning this latency are available for the Bats-Y Exchange (BYX) for the week of May 30, 2016. According to BATS (2016), the order acknowledgement latency for traders using the binary protocol was 75 microseconds on average and 114 microseconds at the 80th percentile. Similar latencies were realized for traders using the FIX protocol: 88 microseconds on average and 114 microseconds at the 80th percentile. The amount of randomness is extremely small in real terms, but it is significant relative to the scale of the distribution. In addition, similar statistics prevail at other exchanges, including NYSE Arca and NYSE American (NYSE (2018)), NASDAQ OMX (NASDAQ OMX (2012)), SIX Swiss Exchange (SIX Swiss Exchange (2016)), Eurex

[6] One contributing factor to variation in network congestion may be "quote stuffing," a practice in which certain traders occasionally submit large numbers of orders in order to slow down certain exchanges (Gai, Yao, and Ye (2013), Egginton, Van Ness, and Van Ness (2016)).

[7] See "This Heat Wave's So Bad It's Even Slowing Down U.S. Stock Trades," *Bloomberg News*, July 2, 2018. Weather may even force traders to abandon radio communication entirely to instead use slower fiber-optic technology (Shkilko and Sokolov (2019)).

[8] Another type of latency, though less relevant to the model, is the time between a quote update or a trade and the corresponding update of the Securities Information Processor (SIP). Bartlett and McCrary (2017) study this form of latency for the time period between August 2015 and June 2016 and find a significant degree of randomness.

Exchange (Deutsche Börse Group (2018)), MIAX Options Exchange (MIAX Miami International Securities Exchange (2018)), Toronto Stock Exchange (TMX Group (2014)), and Brazil's BM&FBOVESPA (Kirilenko and Lamacie (2015)). Furthermore, in a recent industry study, Lehr (2016) finds evidence of considerable variability in the outbound data feeds of Nasdaq, Arca, and NYSE.[9]

These statistics indicate that randomness in within-exchange latency is a meaningful barrier to traders who wish to synchronize the timing of their trades across different exchanges. Thus, although there have been efforts aimed at suppressing the randomness of *trader-to-exchange latency*, including a recent patent application from Renaissance Technology (Mercer and Brown (2016)), within-exchange latency remains a significant constraint.

Exchange-to-trader. We obtain order-level data that allow us to perform a detailed analysis of exchange-to-trader latency. These data, provided by Thesys Technologies, LLC, comprise a historical record of all messages broadcast by all exchanges in the U.S. A unique feature of these data is that each order has two timestamps: the timestamp affixed by the exchange at the point of processing, and the timestamp affixed by the firm at the point of receipt. The difference between these two timestamps is the realized exchange-to-trader latency.

For the analysis below, we restrict our focus to messages from Nasdaq for the week of May 30, 2016. However, roughly similar results would obtain for other exchanges and other sample periods. We record the latency of each message pertaining to SPDR S&P 500 Trust ETF (SPY) sent between 9:30 and 16:00 on the trading days of that week. Figure IA.1 displays the histogram of those latencies, and Table IA.I presents corresponding summary statistics. Two aspects of these data merit mention: latencies are extremely small, on the order of microseconds, and there is a significant amount of randomness. After 99 percent winsorization, the latency is 31 microseconds on average, with a standard deviation of 19 microseconds.

Although these data pertain to just a single message recipient, we have no reason to believe that the magnitude of randomness illustrated by our analysis is driven by this choice. For instance, according to a CME spokesperson, "Every market participant will experience some degree of variability."[10]

[9] That analysis calculates the difference in latencies between two data products offered by the same exchange and finds substantial variability in each case. This requires variability in the latencies of the individual products, although summary statistics pertaining to the individual latencies are not reported separately.

[10] See "High-Speed Traders Profit From Return of Loophole at CME," *Wall Street Journal*, February 12, 2018.

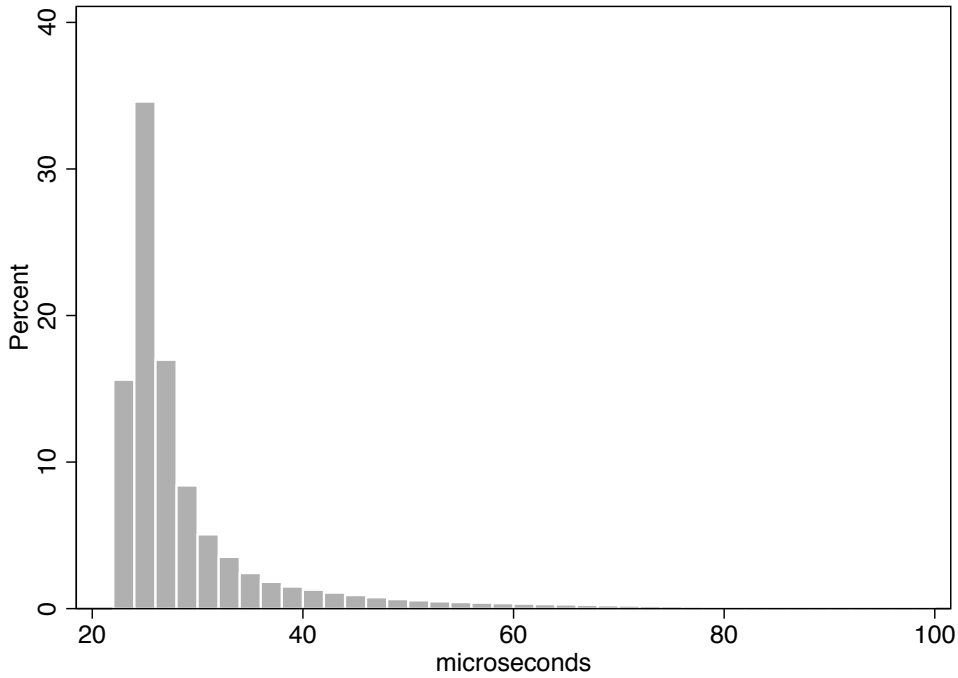


Figure IA.1. Histogram of exchange-to-trader latency. This figure presents a histogram of latencies of all messages pertaining to SPDR S&P 500 Trust ETF (SPY) between Nasdaq and an HFT firm between 9:30 and 16:00 on trading days from May 30, 2016 and June 3, 2016, with 99 percent winsorization. The histogram is truncated at 100 microseconds.

C. Model Simulation

In this section, we simulate the model, using empirical latencies described in the previous section and assuming that traders continue to behave as in equilibrium. We then analyze the fill rates that prevail in this simulation, interpreting them as a back-of-the-envelope estimate of the fill rates predicted by the model. These simulation results suggest that prevailing amounts of randomness in latency are more than sufficient to generate the levels of order anticipation observed in practice, which we summarize in the next section.

While our model ignores the exchange-to-trader component of latency, we reintroduce it for the purposes of this exercise. When simulating a race between an information investor, the liquidity provider, and snipers, we assume that traders behave essentially as in the model: (*i*) the information investor sends immediate-or-cancel orders to each exchange, (*ii*) the liquidity provider sends cancellations to each exchange after observing one trade, and (*iii*) snipers send immediate-or-cancel orders to each exchange after observing two trades. A

Table IA.I

Summary Statistics of Exchange-to-Trader Latency (microseconds)

This table presents summary statistics of latencies of all messages pertaining to SPDR S&P 500 Trust ETF (SPY) between Nasdaq and an HFT firm between 9:30 and 16:00 on trading days from May 30, 2016 and June 3, 2016, with 99 percent winsorization.

mean	st. dev.	p_{20}	p_{40}	p_{60}	p_{80}	N
31.07	18.67	24	25	27	32	7,581,400

trader's order is processed by an exchange only after the corresponding trader-to-exchange and within-exchange latencies, and a trader observes the results of this processing only after the corresponding exchange-to-trader latencies.

To simulate trader-to-exchange and exchange-to-trader latencies, we draw from the empirical distribution described in the previous section and depicted in Figure IA.1. To simulate within-exchange latencies, we draw from a shifted exponential distribution calibrated to match the mean and 80th percentile of the latency distribution discussed in BATS (2016). This calibration yields 11 microseconds for the shift parameter and 64 microseconds for the rate parameter. Throughout, we assume that these three components are distributed independently, not only across traders but also of each other within a trader.

We use this procedure to simulate the outcomes of races between an information investor, the liquidity provider, and snipers, as the number of exchanges, X, ranges from one to 15. Figure IA.2 summarizes the outcomes of 1,000 simulations per choice of X. When $X = 1$, there is no opportunity for order anticipation of any sort, and the investor's fill rate is 100 percent. When $X = 2$, there is no opportunity for snipers to become active and the races are divided between the investor (90.7 percent) and the liquidity provider (9.3 percent). Snipers become active when $X \geq 3$, and win a progressively larger proportion of races as X increases further. When $X = 10$, the simulations predict a fill rate of 56.6 percent for the investor.

D. Evidence of Order Anticipation

If latency is random, as the evidence documented above indicates, then order anticipation should take place in practice. This appendix summarizes a variety of evidence to that effect.

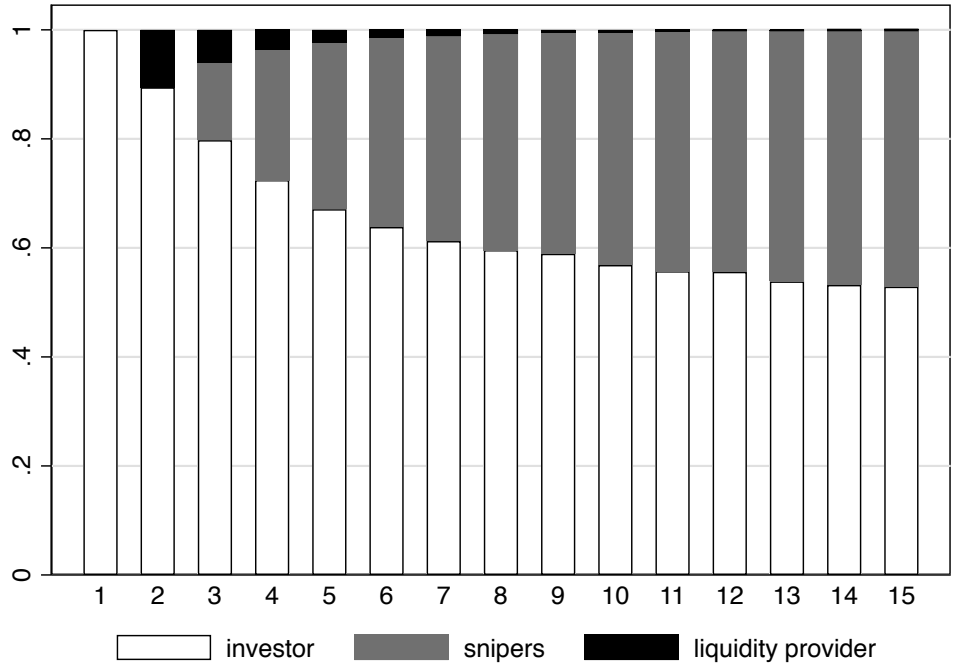


Figure IA.2. Calibrated Fill Rates, by Number of Exchanges. This figure shows the fraction of races between an information investor, the liquidity provider, and snipers won by each party in simulations of the model, for different numbers of exchanges. Each bar is based on 1,000 simulations. Simulations assume that traders behave as in the equilibrium of the model. Simulated exchange-to-trader and trader-to-exchange latencies are drawn iid from the empirical distribution of latencies between Nasdaq and an HFT firm, as described in the text. Simulated within-exchange latencies are drawn iid from a shifted exponential distribution calibrated to match statistics from BATS (2016), as described in the text.

Evidence from industry sources. The summary statistics on fill rates reported by Barclays (2014) hint at the amount of order anticipation they experience. They define the fill rate of a wave as

$$fill\ rate = \frac{\#shares\ filled}{\min\{\#shares\ in\ order, \#shares\ displayed\}}. \tag{IA16}$$

Consistent with the model, they face virtually no order anticipation when attempting to access just a single exchange, achieving a fill rate of 99 percent. However, as they attempt to access quotes at more exchanges, they expose themselves to more order anticipation, and

their fill rate declines.[11] When targeting six or more exchanges, they achieve a fill rate of 89 percent, and a hit ratio (the frequency of complete fills) of just 25 percent.

These fill rates are somewhat higher than those prevailing in the simulations conducted in Internet Appendix II.C. A combination of factors may be responsible for this gap. One set of factors generate upward bias in the Barclays statistics relative to the empirical analogue of the simulation results:

- Only for a subset of the orders it handles does the Barclays router behave like information investors in our model, routing orders for the entire quoted depth. When less than the entire quoted depth is desired, a fill rate of 100 percent could be realized even if some order anticipation occurs.
- The possibility of trading against hidden orders mechanically biases fill rates upward. Executions against hidden orders are added to the numerator of (IA16), but the number of hidden orders quoted is not accounted for in the denominator. It therefore becomes possible that an order for the entire displayed depth would receive a 100 percent fill rate even if order anticipation takes place.

Another set of explanations for the gap pertains to ways in which the simulations may be an imperfect model of real behavior.

- The simulations, although based on real latency data, assume that latencies are drawn independently across traders and across exchanges. Some correlation likely prevails in practice, which would reduce the effectiveness of order anticipation strategies.
- The simulations suppose that traders behave exactly as in the equilibrium of the model—in particular, that liquidity investors target just a single exchange. In practice, there likely exists some size heterogeneity among liquidity investors, so that some target two or more exchanges at once. If the differences between the order routing decisions of liquidity investors and information investors are less stark, then the signals that can be extracted from order flow will be weaker (see Internet Appendix III.H). This would also lessen the extent of order anticipation. Snipers in particular would exercise more caution before acting, since a strong signal is needed to offset the costs of crossing the spread.

Although a coarser form of evidence, a general inability to achieve 100 percent fill rates is

[11] Complicating interpretation of these figures is the fact that the number of exchanges that Barclays targets is most likely an endogenous variable, and therefore we do not claim that this relationship is causal. Nevertheless, these data are consistent with order anticipation, and indeed that is the explanation that Barclays themselves provide.

also consistent with order anticipation. Evidence to that effect comes from KCG (2014). As they report, they achieve a fill rate of approximately 94 percent when attempting to access the entire national best bid and offer (NBBO) depth. Furthermore, their fill rate drops to approximately 91 percent when they widen their sweeps to include dark pools as well.

An additional source of such evidence comes from the exchange IEX. They offer their clients an order routing service, and they publish the fill rates that their router achieves to their website (IEX Group (2018b)). In March 2018, their weighted average fill rate across all Reg NMS "protected" market centers was 94.32 percent. This figure is a single average across the entire set of orders handled by IEX, without distinguishing, as Barclays does, by the number of exchanges targeted. While it is therefore less directly comparable to the simulation results, the fact that IEX does not always achieve a perfect fill rate is nevertheless informative. Moreover, this shortfall is not due to a relative lack of sophistication on the part of IEX. To the contrary, their methods are cutting edge. In particular, they connect to other exchanges through a custom-built fiber optic network that is specifically designed for synchronization.

Related evidence comes from Nanex (2014), a market data provider with an associated research arm. They provide a highly detailed analysis of a small set of trades. The first of these trades—although not implemented using state-of-the-art smart order routing technology—triggered large amounts of order anticipation. The last of these trades was conducted using IEX's more sophisticated smart order router, which reduced order anticipation but does not appear to have eliminated it.

Evidence from academic literature. Order anticipation has also been investigated in the academic literature. For example, Malinova and Park (2017) provide direct empirical evidence of both aggressive-side and passive-side order anticipation for Canadian equities. In the millisecond following a trade on a particular exchange, they detect market-wide increases both in the probability of HFTs canceling their quotes and in the probability of HFTs trading aggressively in the same direction. Also consistent with order anticipation, they report significantly lower fill rates for orders that target two exchanges than for orders that target a single exchange, even after controlling for order size. Similarly, van Kervel (2015) finds that executions on one venue are followed by cancellations elsewhere within the subsequent 100 milliseconds, which is consistent with passive-side order anticipation. Additionally, a number of other studies also find evidence of order anticipation, although not necessarily of the form modeled in this paper (Hirschey (2019), Korajczyk and Murphy (2019), van Kervel

and Menkveld (2019)).

Lastly, Weller (2018) and Gider, Schmickler, and Westheide (2019) find a connection between HFT and less informative prices, which they suggest may be caused by the form of order anticipation that we model in this paper.

III. Additional Results

In this appendix, we present additional results that may be of interest. Internet Appendix III.A supplements Corollary 1 by discussing how the LOB equilibrium is affected by changes in the model parameters beyond p_H. Similarly, Appendices III.B and III.C discuss comparative statics for the equilibria of NDs and FBAs. Internet Appendix III.D studies an asynchronous version of the FBAs mechanism, finding that it generates an equilibrium outcome identical to that prevailing under 1-ND. Internet Appendix III.E studies the consequences of adding a small delay to *all* orders, finding that it generates an equilibrium outcome identical to that prevailing under (synchronized) FBAs. Internet Appendix III.F demonstrates that the economic forces of our model carry over to a setting in which liquidity investors arrive in a stream, which effectively opens the door for informed traders to split orders dynamically. Internet Appendix III.G motivates our use of the hyperreal-based construction of time by describing the difficulties that would arise given a more standard construction. Internet Appendix III.H considers an extension of the model in which liquidity investors may be heterogeneous in the quantities they demand. Internet Appendix III.I considers an extension of the model in which information investors may be risk averse. Internet Appendix III.J presents an argument in support of the equilibrium selection that we have made.

A. *Limit Order Book Comparative Statics*

Corollary 1 in the main text summarizes the comparative statics of the LOB equilibrium with respect to p_H. As the parameter of the model that governs the speed of HFTs, this is likely the most interesting and policy-relevant comparative static exercise. But for completeness, in this appendix we provide and discuss comparative statics with respect to the remaining parameters.

COROLLARY 1: *The spread s^*_{LOB} and the research intensity r^*_{LOB} have the following comparative statics:*

(i) s^*_{LOB} is weakly increasing in X, weakly increasing in p_I, and weakly increasing in λ.
(ii) r^*_{LOB} is weakly decreasing in λ.

Proof of Corollary 1:

With respect to X: Differentiating the expression for the spread given in (1), we find that,

other things equal, it is weakly increasing in $X_I + X_S$. It can be shown that $X_I + X_S$ is weakly increasing in X on the domain of the positive integers.[12] These two observations imply that, other things equal, the spread is weakly increasing in X. It is also weakly increasing in research intensity, other things equal.

Moreover, applying Topkis' Theorem to (2), we find that, other things equal, research intensity is weakly increasing in X_I. By Lemma 1, X_I is weakly increasing in X. These two observations imply that, other things equal, research intensity is weakly increasing in X. It is also weakly decreasing in the spread, other things equal.

By combining these observations, we conclude that $s^{\star}_{LOB}$ is weakly increasing in X.

With respect to p_I: Differentiating the expression for the spread given in (1), we find that, other things equal, it is weakly increasing in $X_I + X_S$. It can be shown that $X_I + X_S$ is weakly increasing in p_I.[13] These two observations imply that, other things equal, the spread is weakly increasing in p_I. It is also weakly increasing in research intensity, other things equal.

Moreover, applying Topkis' Theorem to (2), we find that, other things equal, research intensity is weakly increasing in X_I. It can be shown that X_I is weakly increasing in p_I.[14]

[12] Notice that when $X = 1$, $X_I + X_S = 1$, and when $X = 2$, $X_I + X_S \geq 1$. Thus, it suffices to show that $\frac{\partial(X_I + X_S)}{\partial X} \geq 0$ on the domain where $X \geq 2$. Computing the derivative,

$$\frac{\partial(X_I + X_S)}{\partial X} = 1 - p_H p_I (1 - p_I)^{X-1} \left[2X - 1 + \ln(1 - p_I)(X^2 - X)\right].$$

Because $p_I \geq 0.5$, $p_I \leq 1$, and $p_H \leq 1$, we have

$$p_H p_I (1 - p_I)^{X-1}[2X - 1 + \ln(1 - p_I)(X^2 - X)] \leq 0.5^{X-1}[2X - 1 + \ln(0.5)(X^2 - X)].$$

The right-hand side is maximized on the domain $X \geq 2$ at $X = 2$, where it achieves ≈ 0.807. We conclude that the derivative has the desired sign.

[13] Notice that

$$\frac{\partial(X_I + X_S)}{\partial p_I} = -p_H(X - 1)X(1 - p_I X)(1 - p_I)^{X-2}.$$

When $X = 1$, this derivative evaluates to zero. When $X \geq 2$ (and using the fact that $p_I \geq 0.5$), the derivative is nonnegative.

[14] Notice that

$$\frac{\partial X_I}{\partial p_I} = X - X^2(1 - p_I)^{X-1} + (1 - p_H)(X - 1)X(1 - p_I X)(1 - p_I)^{X-2}.$$

This derivative is minimized on the domain $p_H \geq 0.5$ at $p_H = 0.5$. Plugging that in and taking the first order condition with respect to p_I, we see that when $X \in \mathbb{N}$, there are no critical points $p_I \in (0.5, 1)$, so we conclude that the derivative is minimized either by setting $p_I = 0.5$ or by setting $p_I = 1$. If the latter, then the derivative evaluates to X, which is positive. If the former, then the derivative evaluates to $X - (X^3 - X^2 + 2X)(0.5)^X$, which is weakly positive on $X \in \mathbb{N}$.

These two observations imply that, other things equal, research intensity is weakly increasing in p_I. It is also weakly decreasing in the spread, other things equal.

By combining these observations, we conclude that s^*_{LOB} is weakly increasing in p_I.

With respect to λ: Differentiating the expression for the spread given in (1), we find that, other things equal, it is weakly increasing in λ. Applying Topkis' Theorem to (2), we find that, other things equal, research intensity is weakly decreasing in the spread and constant in λ. By combining these observations, we establish the claimed comparative statics for s^*_{LOB} and r^*_{LOB} with respect to λ. □

Number of exchanges (X). Given the high degree of fragmentation in modern equity markets, another very relevant set of comparative statics are those with respect to the number of exchanges, X. According to the proposition, an increase in X increases the equilibrium spread. The intuition can be seen through the following chain of forces. First, as observed in Section III.B of the main article, aggregate depth increases in the number of exchanges. Indeed, it scales linearly in the number of exchanges because the liquidity provider optimally offers one share at both the bid and the ask at each exchange in order to serve a liquidity investor who might attempt to trade there. Next, this increase in the depth of the aggregate book also increases the number of shares available to informed traders (either directly informed information investors or indirectly informed snipers). The liquidity provider is therefore exposed to more adverse selection, and she must charge a larger spread to compensate.[15]

This represents an interesting contrast to the irrelevance result of Glosten (1994), who demonstrates that in an idealized frictionless setting in which investors can costlessly send simultaneous orders to separate exchanges in order to complete their trades at the best possible price, the liquidity of the aggregate market is invariant to the degree of fragmentation. We do not obtain the same invariance result in our model because investors do not always act to complete their trades at the best possible price. Rather, we make the extreme assumption that liquidity investors are perfectly inelastic in their exchange choices. Nevertheless, similar forces would break the Glosten (1994) invariance result even in more realistic models with some cross-exchange elasticity, at least so long as some friction precludes perfect elasticity.

[15] A countervailing force not captured by this model is the following. If exchanges are strategic players that compete for order flow, then an increase in their number might reduce their market power and therefore the fees that they charge. All else equal, smaller fees might be passed on as smaller spreads. Baldauf and Mollner (forthcoming) propose a model that incorporates both this "competition channel" and the "exposure channel" described above, but empirical analysis indicates that the exposure channel dominates.

On the other hand, the response of research intensity to an increase in X is theoretically ambiguous, due to two competing effects. The direct effect of an increase in X is to create more opportunities for an information investor to trade on any piece of information, which tends to incentivize research. However, an increase in X also creates more opportunities for snipers to trade, which contributes to adverse selection and raises the spread. Larger spreads make each trade less profitable, so the indirect effect of an increase in X tends to disincentivize research.

Investor speed (p_I). An increase in p_I represents an improvement in the order routing technology of the investor, because orders are processed sooner and are less dispersed. With better technology, an information investor obtains more fills. While snipers may obtain either more or fewer fills when p_I increases, the overall effect is that more information-motivated trades take place. To offset the additional adverse selection, the liquidity provider must charge a larger spread. The response of equilibrium research intensity to a change in p_I is theoretically ambiguous for essentially the same reason that research may either increase or decrease in X.

Probability of information investor (λ). Finally, an increase in the probability of an information investor, λ, intensifies the adverse selection faced by the liquidity provider, who then quotes a larger spread. The larger spread reduces the benefits of research and leads to lower research intensity.

B. Noncancellation Delay Comparative Statics

This appendix uses the characterization of the qND equilibrium given in Proposition 3 to study how this outcome varies with the parameters of the model.

COROLLARY 2: *The spread s^*_{qND} and the research intensity r^*_{qND} have the following comparative statics:*

(i) s^*_{qND} is weakly decreasing in p_H, weakly increasing in X, weakly increasing in p_I, and weakly increasing in λ.

(ii) r^*_{qND} is weakly decreasing in p_H, weakly increasing in X, weakly increasing in p_I, and weakly decreasing in λ.

Proof of Corollary 2: As before, define

$$X_{qND} = q^X + \sum_{x=1}^{X} \binom{X}{x} (1-q)^x q^{X-x} \left(x p_H (1-p_I)^x + x \left[1 - p_H (1-p_I) \right] \right).$$

Lemma 2 establishes that X_{qND} is weakly increasing in X. Moreover, by differentiating X_{qND}, we find that it is weakly decreasing in p_H and weakly increasing in p_I.

As is evident from equation (4), other things equal, the spread is (*i*) weakly increasing in research intensity, (*ii*) weakly increasing in λ, and (*iii*) weakly increasing in X_{qND}. Moreover, the latter implies that, other things equal, the spread is also (*iv*) weakly decreasing in p_H, (*v*) weakly increasing in X, and (*vi*) weakly increasing in p_I. Applying Topkis' Theorem (5), we see that, other things equal, research intensity is (*i*) weakly decreasing in the spread, (*ii*) constant in λ, and (*iii*) weakly increasing in X_{qND}. Moreover, the latter implies that, other things equal, research intensity is also (*iv*) weakly decreasing in p_H, (*v*) weakly increasing in X, and (*vi*) weakly increasing in p_I. By combining these observations, we establish all claimed comparative statics except those of r^*_{qND} with respect to p_H, X, and p_I.

Recall that r^*_{qND} is characterized as the fixed point of the correspondence

$$R_{qND}(\hat{r}) = \arg\max_{r \in [0,1]} \left\{ \frac{(1-\lambda) r X_{qND}}{1 - \lambda + \lambda \hat{r} X_{qND}} - c(r) \right\}.$$

Applying Topkis' Theorem, $R_{qND}(\hat{r})$ is weakly increasing in X_{qND} and therefore is (*i*) weakly decreasing in p_H, (*ii*) weakly increasing in X, and (*iii*) weakly increasing in p_I. Because the correspondence is also weakly decreasing in $\hat{r}$, we obtain that r^*_{qND} has the claimed comparative statics with respect to these parameters. □

The intuition for the comparative statics with respect to p_H and λ is analogous to the intuition for the corresponding comparative statics under the LOB. The intuition for the comparative statics with respect to X and p_I is as follows. Adding another exchange (i.e., increasing X) or improving the investor's routing technology (i.e., increasing p_I) increases the number of venues at which an information investor may trade after learning the value of the security, which increases the returns to research and incentivizes a higher research intensity. The higher research intensity increases the adverse selection faced by the liquidity provider, who quotes a larger spread.

These comparative statics under NDs are in contrast to what prevails under the LOB, where it is theoretically ambiguous how research intensity responds to changes in X and p_I.

The crucial difference is that with NDs, adverse selection against the liquidity provider comes only from the information investor, whereas under the LOB, snipers are another source of adverse selection.

C. *Frequent Batch Auctions Comparative Statics*

This appendix uses the characterization of the FBA equilibrium given in Proposition 4 to study how this outcome varies with the parameters of the model.

COROLLARY 3: *The spread s^*_{FBA} and the research intensity r^*_{FBA} have the following comparative statics:*

(i) s^*_{FBA} *is weakly increasing in* X *and weakly increasing in* λ.
(ii) r^*_{FBA} *is weakly increasing in* X *and weakly decreasing in* λ.

Proof of Corollary 3: As is evident from equation (6), other things equal, the spread is (*i*) weakly increasing in research intensity, (*ii*) weakly increasing in λ, and (*iii*) weakly increasing in X. Applying Topkis' Theorem to (7), we see that, other things equal, research intensity is (*i*) weakly decreasing in the spread, (*ii*) constant in λ, and (*iii*) weakly increasing in X. By combining these observations, we establish all claimed comparative statics except those of r^*_{FBA} with respect to X.

Recall that r^*_{FBA} is characterized as the fixed point of the correspondence

$$R_{FBA}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda)rX}{1-\lambda+\lambda\hat{r}X} - c(r) \right\}.$$

By Topkis' Theorem, $R_{FBA}(\hat{r})$ is weakly increasing in X. Because the function is also weakly decreasing in $\hat{r}$, we obtain that r^*_{FBA} has the claimed comparative static with respect to X. □

The intuition for the comparative statics with respect to λ is analogous to the intuition for the corresponding comparative statics under the LOB. The intuition for the comparative statics with respect to X is analogous to that under NDs. However, the comparative statics with respect to p_H and p_I are all zero. Under the LOB or under NDs, these parameters control the number of orders that an information investor expects to convert into fills. In contrast, under FBAs, information investors always convert all orders, and therefore changes in these parameters have no effect.

D. *Asynchronous Frequent Batch Auctions*

While Budish, Cramton, and Shim (2014) advocate the use of FBAs that are synchronized across exchanges, in practice it may be difficult to achieve synchronization among competitors. As a result, a natural question that arises is what would transpire if the batch intervals were not synchronized. Our model allows us to examine what would happen in such a setting.

We suppose that all exchanges use FBAs with "long" batch intervals that are "sufficiently" asynchronous. Formally, in the language of the model, we suppose that for any batch auction conducted by any exchange, no batch auction is conducted by the same or another exchange until at least 3ε has passed. Otherwise, we assume all is as before, and in particular, that the batch length of each exchange is an infinitesimal.

It can be shown that in such a setting, the equilibrium outcome would be the same as that which prevails under 1-ND, characterized in Proposition 3. The intuition is the same: aggressive-side order anticipation is eliminated, and in addition, an information investor's orders are processed with sufficient temporal dispersion to allow the liquidity provider to react after just a single trade. Whereas 1-ND achieves this dispersion by adding randomness to the processing times of noncancellation orders, asynchronous FBAs achieve it by fixing clearing times in a way that is dispersed across exchanges.

E. *Universal Delay*

It is interesting to note that in this model, the outcome implemented by FBAs can also be achieved with a *universal delay* applied to all orders before they are processed by an exchange.[16] In this appendix, we describe the mechanism and then discuss some of its attractive (albeit unmodeled) properties relative to FBAs.

Formally, we define universal delay to be the following proposal. All orders receive a delay of length δ_{UD}. To have the desired effect, δ_{UD} should be small but should exceed the difference between (*i*) the maximum latency that the investor may experience and (*ii*) the sum of the minimum latency that the investor may experience together with the minimum latency that an HFT may experience. In the language of the paper, this requirement corresponds to an infinitesimal delay $\delta_{UD} > \varepsilon$. Aside from this delay, all order processing is as in the LOB. It can

[16] Our analysis considers a constant delay. In contrast, versions of universal delay in which orders are delayed by random lengths of time are advocated by Harris (2013) and were subsequently implemented by the foreign exchange venues EBS and ParFX (under the label "latency floor").

be shown that in such a setting, the equilibrium outcome would be the same as that which prevails under FBAs, characterized in Proposition 4. The intuition is the same: information investors become able to trade against all mispriced quotes before HFTs can react. Whereas batching achieves this by synchronizing the time of trade across exchanges, a universal delay achieves this by delaying the reaction time of HFTs.

There are several reasons to think that a universal delay would be preferable to FBAs. First, by virtue of being so near the status quo, it would be easier to implement and therefore less likely to suffer from glitches, loopholes, or other complications. In particular, a universal delay could be implemented by, for example, forcing all orders to travel through additional lengths of fiber-optic cable. Such a scheme is already used in practice by the exchange IEX, which implements a 350-microsecond delay by placing 38 miles of coiled cable between their matching engine and their point of presence.[17] Second, there are some legal questions pertaining to whether it is possible for FBAs to operate simultaneously on multiple exchanges in a way that satisfies laws as they are currently written, particularly Regulation NMS in the U.S. In contrast, the SEC has already approved IEX's universal delay, characterizing it as de minimis. Third, FBAs require synchronization across exchanges, which in practice might be difficult to achieve since exchanges are competitors. In contrast, order processing delays could be implemented in a decentralized way.

Furthermore, it might be useful to implement a universal delay in conjunction with a qND, for various values of q. Certain points on the frontier of the feasible set can be implemented by NDs but only if the number of exchanges is increased beyond X, as in the spirit of Corollary 6. Some of those points can be implemented by a hybrid mechanism in this class without the need to vary the number of exchanges.

F. Order Synchronization in a Dynamic Version of the Model

In the model presented in Section II of the main article, there can be at most a single liquidity investor present, who trades just a single share. This effectively limits the information investor to trading at a single point in time, and he therefore trades as intensely as possible at that time, synchronizing his trading in one large "wave." However, a more realistic model of trading on long-lived private information would feature multiple liquidity traders who arrive gradually. This would create the opportunity for an informed trader to camouflage himself by making many small trades over time (à la Kyle (1985))—trading less

[17] The NYSE American exchange subsequently adopted essentially the same scheme. A difference is that their delay is implemented via software instead of hardware.

intensely now in order to have less price impact and therefore better terms of trade in the future.

In this appendix we consider such a dynamic model. While we do not solve this model completely, we show that in certain regions of the parameter space—in particular, when trading is fragmented across a sufficiently large number of exchanges—equilibrium requires that the informed trader sometimes trade in a large wave even though that gives himself away. The driving insight is that if he makes many trades at separate times, then each trade affects the price of the next, regardless of whether those trades occur at the same exchange or different exchanges. In contrast, if he makes many trades at the same time by synchronizing across exchanges, then none of those trades can influence the price he receives for another. If many exchanges are available, then such an "ambush" can be large and worthwhile.

We view this analysis as justification and motivation for our current approach, which can be interpreted as a model of a single one of these waves. While the model that we present in Section II of the main article is more simple in the temporal dimension, it is more complex in other dimensions, for instance, by capturing latency. Thus, although it does not allow us to consider how an informed trader might split orders over time, it does allow us to consider in more detail the issues that arise when such a trader splits orders across exchanges. What the results of this appendix indicate is that these issues remain relevant even when orders can be split over time.

One-exchange model. In the case of one exchange (i.e., $X = 1$), the model reduces to the "Glosten-Milgrom model" of Back and Baruch (2004, Section 2). The text in the following paragraphs is copied almost verbatim from that paper.[18] Below, we generalize to multiple exchanges (i.e., $X > 1$) by adapting Back and Baruch (2004) in a natural way.

We consider a continuous-time market for a risky asset and one risk-free asset with interest rate set to zero. A public release of information takes place at a random time τ, distributed as an exponential random variable with parameter r. After the public announcement has been made, the value of the risky asset, denoted by $\tilde{v}$, will be either zero or one, and all positions are liquidated at that price. There is a single informed trader who knows $\tilde{v}$ at date 0. Uninformed (i.e., noise) buy and sell orders arrive as Poisson processes with constant, exogenously given, arrival intensities β. We denote the order size by δ.

We denote the total number of buy orders by noise traders through time t by z_t^+ and the

[18] We also use the notation of Back and Baruch (2004). The model in this appendix is distinct from our baseline model, which we present and analyze in Section II of the main article, and notation should not be assumed to have the same meaning across both models.

total number of sell orders by noise traders through time t by z_t^-, and we set $z_t = z_t^+ - z_t^-$. The net number of shares bought by noise traders is then $z_t\delta$. Similarly, we denote the number of informed buys by x_t^+, the number of informed sells by x_t^-, and the net informed orders by $x_t = x_t^+ - x_t^-$. Finally, we denote the number of net noise and informed orders through time t by y_t. The process y reveals the complete history of anonymous trades. The σ-field generated by $\{y_s | s \leq t\}$ is denoted by $\mathcal{F}_t^y$. As usual, we denote the left limit of y at time t by y_{t-} and set $\Delta y_t = y_t - y_{t-}$. If there is a buy order at date t then $\Delta y_t = 1$, and if there is a sell order then $\Delta y_t = -1$. Competition among the market makers implies that any transaction takes place at price $p_t \equiv \mathbb{E}[\tilde{v}|\mathcal{F}_t^y]$. The posted ask and bid prices at time $t < \tau$ are

$$\begin{aligned} \text{ask}_t &= \mathbb{E}[\tilde{v}|\mathcal{F}_{t-}^y, \Delta y_t = +1] \\ \text{bid}_t &= \mathbb{E}[\tilde{v}|\mathcal{F}_{t-}^y, \Delta y_t = -1]. \end{aligned}$$

Here, $\mathcal{F}_{t-}^y \equiv \bigcup_{s<t} \mathcal{F}_s^y$ denotes the information available to the market makers just before time t. The informed trader chooses a trading strategy x to maximize

$$\mathbb{E}\left[\delta \int_0^\tau [\tilde{v} - \text{ask}_t] dx_t^+ + \delta \int_0^\tau [\text{bid}_t - \tilde{v}] dx_t^- \;\middle|\; \tilde{v} \right].$$

We direct the reader to Back and Baruch (2004) for additional details.

One-exchange equilibrium. Theorem 2 of Back and Baruch (2004) characterizes an equilibrium of this model. In addition, their Figures 1 and 2 illustrate features of that equilibrium, computed numerically for certain parameter values. Table IA.II below extracts key features of the equilibrium from those figures. The outcome variables recorded in the table are the informed trader's equilibrium profits and the initial quotes set by the market maker. The table's last four columns correspond to a different set of parameter values.

Back and Baruch (2004) do not assert that this is the unique equilibrium of the model. Nevertheless, it seems to be a natural selection, and it is the unique equilibrium on which they focus their analysis. We therefore proceed under the assumption that this equilibrium is the appropriate benchmark for our analysis.

Multi-exchange model. We consider the following multi-exchange adaptation of the Back and Baruch (2004) model. Let X denote the number of exchanges. As in our baseline model of Section II of the main article, we suppose that noise traders choose an exchange on which to transact uniformly at random. Thus, uninformed buy and sell orders arrive to each

Table IA.II
Numerical Examples Based on Back and Baruch (2004)

This table presents the outcome variables are approximations based on Figures 1 and 2 in Back and Baruch (2004).

Parameters					
r	discount factor	1	1	1	1
p_0	prior expectation of $\tilde{v}$	0.5	0.5	0.5	0.5
δ	order size	3	1	0.2	0.1
β	noise trader arrival rates	$0.0\bar{5}$	0.5	12.5	50
Outcome variables					
	expected profit	0.06	0.19	0.22	0.25
	bid_0	0.02	0.12	0.36	0.42
	ask_0	0.98	0.88	0.64	0.58
Fragmentation cutoff		1.0	1.6	3.0	6.0

exchange as Poisson processes with arrival intensities β/X. The processes z_t^+, z_t^-, z_t, x_t^+, x_t^-, x_t, and y_t are defined as before, with the exception that they are now vector-valued, where each component of the vector corresponds to trading activity on one of the X exchanges. The posted ask and bid prices at time $t < \tau$ at exchange i are

$$\text{ask}_t^i = \mathbb{E}[\tilde{v}|\mathcal{F}_{t-}^y, \Delta y_t^i = +1]$$
$$\text{bid}_t^i = \mathbb{E}[\tilde{v}|\mathcal{F}_{t-}^y, \Delta y_t^i = -1],$$

where, as before, $\mathcal{F}_{t-}^y \equiv \bigcup_{s<t} \mathcal{F}_s^y$ denotes the information available to market makers just before time t.[19]

Suppose now that the informed trader is prohibited from trading in waves. In other words, suppose it is a constraint that Δx_t can be nonzero in only one dimension. Under this constraint, the model can be solved as in the single-exchange case, and "the benchmark equilibrium" is again as characterized by Back and Baruch (2004). The only difference is that whenever the informed trader decides to trade, he needs to choose an exchange, a choice that he makes uniformly at random.

[19] In summary, market makers are able to update their quotes in response to trades that took place in the past, regardless of the exchange on which those trades took place. However, they cannot respond in the same instant that a trade takes place. Moreover, the informed trader is capable of accessing quotes at multiple exchanges in the same instant and therefore before any of the market makers can respond. This is as if latency were deterministic and positive, though infinitesimally small.

Our approach is to argue that this constraint is binding in certain regions of the parameter space. For this constraint to be binding, it would be sufficient to show that it is profitable for the informed trader to deviate by trading on all X exchanges immediately, and then never trade again. The expected profit of such a deviation is $p_0 \delta X (1 - \text{ask}_0) + (1 - p_0) \delta X (\text{bid}_0 - 0)$. Because this expression diverges as $X \to \infty$, the deviation is profitable if X is sufficiently large. We solve for the smallest value of X that makes this deviation profitable. We refer to this as the *fragmentation cutoff*. The last row of Table IA.II shows the fragmentation cutoffs corresponding to the four sets of parameter values.

In conclusion, if there is a sufficient amount of fragmentation, then the benchmark equilibrium cannot survive when the constraint against waves is lifted. This indicates that in such cases, equilibrium requires that the informed trader use waves.[20,21] In turn, this suggests that the insights delivered by a static model of a single wave may be relevant even though reality is dynamic.

One of these main insights, which our static model delivers, is that informed traders are adversely affected by random latency because it enables order anticipation. It seems intractable to add random latency to this dynamic model. Nevertheless, in the same way that prohibiting the informed trader from using waves is a binding constraint, it would seem that random latency—the effect of which is to make it more difficult for the informed trader to use waves successfully—would adversely affect his profits in a similar way.

Discussion. We conclude that equilibrium behavior in our current, effectively static model—the informed trader attempts to achieve cross-exchange synchronization of his trades—would extend to a dynamic setting. In consequence, the insights generated by our current approach should extend in a similar way. While these conclusions hinge upon fragmentation being sufficiently severe, it seems to us that reality is better approximated by that case, for several reasons:

- Trading is quite fragmented today. Between exchanges and alternative trading systems,

[20] Note that progressively more extreme amounts of fragmentation are required to make this argument as $\delta \to 0$, or as we converge to the "Kyle model" of Back and Baruch (2004, Section 1). This suggests that while fragmentation of this nature might not alter equilibrium behavior in Kyle models materially, it does so in Glosten-Milgrom models.

[21] In addition, several unmodeled forces might further increase the relative attractiveness of waves over intertemporal order splitting, and thereby further intensify this effect. As one example, there might be costs associated with splitting orders over time, which could originate from discounting, cognitive costs, trading algorithm usage fees, etc. As another example, the informed trader might also worry about the possibility of another trader with the same information, in which case prices would move against him more quickly than the theory predicts. An early wave would then become more attractive.

many stocks are traded on over 30 venues in the U.S. alone.

- The effort that has been devoted to improving the synchronization of order arrivals is a smoking gun suggestive of the desirability of conducting at least some trading in large waves. One example of such an effort is the smart order router THOR (Aisen et al. (2015)), whose development by RBC is chronicled by Lewis (2014).
- Finally, there is evidence that traders do in fact use waves in practice. For instance, 15 percent of all orders included in the data set of Malinova and Park (2017) occur in waves. The authors also point to institutional differences between Canada and the U.S., arguing that their findings may understate the pervasiveness of this phenomenon.

In summary, it is true that traders do "work" large orders over time (à la Kyle (1985)), and our current approach—while parsimonious and tractable—prevents us from modeling that behavior. However, both theoretical and empirical evidence suggest that such traders nevertheless trade in waves, so that the forces modeled in this paper remain relevant.

G. The Advantage of a Hyperreal Construction of Time

In this appendix, we explain why the results in the main text hinge on the use of hyperreal numbers to index time. As mentioned earlier, our approach is tantamount to working in the limit, as N goes to infinity, of a sequence of models in which the unit interval is divided into N discrete time periods. An advantage of this approach is that the equilibrium is stationary, and in particular, the bid-ask spread remains constant at least until the first trade occurs.

To demonstrate that this stationarity is true only in the limit, Internet Appendix III.G.1 analyzes a version of the model with a finite number of time periods, demonstrating where things go awry. Intuitively, only *in* the limit will the scale of the latency distribution be infinitely smaller than the scale of the distribution governing investor arrival times. In addition to being a reasonably close approximation of modern trading, this is the crucial feature for delivering a tractable model with a stationary equilibrium. In contrast, when there are only a finite number of time periods, the scales of the two aforementioned distributions necessarily differ by only a finite multiple.

Building on that analysis, Internet Appendix III.G.2 illustrates a sense in which our baseline model, with a hyperfinite number of time periods, truly corresponds to working in the limit of a sequence of finite-period models. This, we believe, highlights that our time construction is neither masking anything suspicious nor introducing anything contrary to

the intuitions that would come out of more conventional time constructions. Rather, our approach merely attempts to formalize those very intuitions in a simple and tractable way.

G.1. The Finite-Period Model

Given a natural number $N \geq 3$, we define a set $\mathcal{T}_N = \{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$. In the baseline model, time periods are indexed by the hyperfinite set $\mathcal{T}$, which can be thought of as the limit of $\mathcal{T}_N$ as N diverges. In this appendix we consider the alternative version of the model in which time periods are indexed by the finite set $\mathcal{T}_N$.

As in the baseline model, suppose that the minimum latency is $\frac{1}{N}$ while the maximum latency is $\frac{3}{N}$. Moreover, suppose that the investor's arrival time is drawn from the uniform distribution on $\{0, \frac{1}{N}, \ldots, \frac{N-3}{N}\}$.[22] Of particular interest may be the simplest version of this finite model, which is the case of $N = 3$. In that special case, the investor always arrives at $t = 0$. Orders he sends at $t = 0$ arrive at either $t = \frac{1}{3}$ or $t = 1$.

To illustrate the challenges that these models would pose, we analyze the special case in which there are only two exchanges ($X = 2$), HFTs always obtain the minimum latency ($p_H = 1$), and research is costless ($c(r) \equiv 0$). Similar challenges would arise in more complex cases.

Suppose that liquidity investors and information investors continue to behave as in the equilibrium of the baseline model. In particular, each information investor conducts research with intensity $r = 1$ and sends orders to both exchanges immediately upon learning the value of the security. Likewise, each liquidity investor sends one order to a randomly chosen exchange immediately upon arrival. Given this behavior, we can derive, for each $t \in \mathcal{T}_N$, the probability that both (*i*) no order arrives at either exchange at all times $s < t$ and (*ii*) an information investor order arrives at a given exchange (say, Exchange 1) at time t. We denote this probability $P^i_{N,t}$. We can also derive the corresponding probability for liquidity investor orders, which we denote $P^l_{N,t}$. As the subsequent analysis will make clear, these probabilities will be the relevant quantities for determining the zero-profit spread.

- For period $t = \frac{1}{N}$ and focusing on the information investor case, the event in question occurs if and only if the investor is an information investor (which occurs with probability λ) who arrives at time zero (which occurs with probability $\frac{1}{N-2}$) and who obtains a short

[22] Under the alternative assumption that the investor's arrival time is drawn from the uniform distribution on $\mathcal{T}_N$, some of his orders would arrive after the market close. We therefore assume instead that investors do not arrive in the last three periods of $\mathcal{T}_N$, to ensure that none of the investor's orders are lost in this way. In the baseline model, we assume that the investor's arrival time is drawn from the uniform distribution on the entirety of $\mathcal{T}$, but this is without consequence because orders are lost with only infinitesimal probability.

latency draw for his order sent to the given exchange (which occurs with probability p_I). Given the independence of these events, $P^i_{N,t} = \frac{\lambda}{N-2} p_I$.

For period $t = \frac{1}{N}$ and focusing on the liquidity investor case, the event in question occurs if and only if the investor is a liquidity investor (which occurs with probability $1-\lambda$) who arrives at time zero (which occurs with probability $\frac{1}{N-2}$) who has the given exchange as his home exchange (which occurs with probability $\frac{1}{2}$), and who obtains a short latency draw for his order (which occurs with probability p_I). Thus, $P^l_{N,t} = \frac{1-\lambda}{2(N-2)} p_I$.

If $N > 3$, then these same probabilities apply to period $t = \frac{2}{N}$ as well. If $N = 3$, then these probabilities are both zero for period $t = \frac{2}{3}$.

- For periods $t \in \{\frac{3}{N}, \frac{4}{N}, \ldots, \frac{N-2}{N}\}$ and focusing on the information investor case, the event in question can occur in either of two ways. First, it occurs if the investor is an information investor (which occurs with probability λ) who arrives at time $t - \frac{3}{N}$ (which occurs with probability $\frac{1}{N-2}$) and who obtains long latency draws for both orders that he sends (which occurs with probability $(1-p_I)^2$). Second, it occurs if the investor is an information investor (which occurs with probability λ) who arrives at time $t - \frac{1}{N}$ (which occurs with probability $\frac{1}{N-2}$) and who obtains a short latency draw for his order sent to the given exchange (which occurs with probability p_I). Thus, $P^i_{N,t} = \frac{\lambda}{N-2}\left[(1-p_I)^2 + p_I\right]$.

  For periods $t \in \{\frac{3}{N}, \frac{4}{N}, \ldots, \frac{N-2}{N}\}$ and focusing on the liquidity investor case, the event in question can occur in either of two ways. First, it occurs if the investor is a liquidity investor (which occurs with probability $1-\lambda$) who arrives at time $t - \frac{3}{N}$ (which occurs with probability $\frac{1}{N-2}$) who has the given exchange as his home exchange (which occurs with probability $\frac{1}{2}$), and who obtains a long latency draw for his order (which occurs with probability $1-p_I$). Second, it occurs if the investor is a liquidity investor (which occurs with probability $1-\lambda$) who arrives at time $t - \frac{1}{N}$ (which occurs with probability $\frac{1}{N-2}$) who has the given exchange as his home exchange (which occurs with probability $\frac{1}{2}$) and who obtains a short latency draw for his order (which occurs with probability p_I). Thus, $P^l_{N,t} = \frac{1-\lambda}{2(N-2)}(1-p_I)$.

- For period $t = 1$ and focusing on the information investor case, the event in question occurs if and only if the investor is an information investor (which occurs with probability λ) who arrives at time $\frac{N-3}{N}$ (which occurs with probability $\frac{1}{N-2}$) and who obtains long latency draws for both orders that he sends (which occurs with probability $(1-p_I)^2$). Thus, $P^i_{N,t} = \frac{\lambda}{N-2}(1-p_I)^2$.

For period $t = 1$ and focusing on the liquidity investor case, the event in question occurs if and only if the investor is a liquidity investor (which occurs with probability $1-\lambda$), who arrives at time $\frac{N-3}{N}$ (which occurs with probability $\frac{1}{N-2}$), who has the given exchange as his home exchange (which occurs with probability $\frac{1}{2}$), and who obtains a long latency draw for his order (which occurs with probability $1-p_I$). Thus, $P^l_{N,t} = \frac{1-\lambda}{2(N-2)}(1-p_I)$.

If $N > 3$, then these same probabilities apply to period $t = \frac{N-1}{N}$ as well. If $N = 3$, then, as above, these probabilities are both zero for period $t = \frac{2}{3}$.

For the purposes of this appendix, suppose that when an exchange receives multiple orders simultaneously, any orders sent by HFTs are processed first. (In the baseline model, such ties between HFTs and investors do not arise with positive probability on path, but they may occur here.) Suppose also that HFTs continue to behave as in the equilibrium of the baseline model. In particular, because there are only two exchanges, snipers send no orders. Given these assumptions and the earlier assumption that $p_H = 1$, when liquidity providers set their quotes for a time t, they need only consider orders that may arrive from liquidity investors or information investors at that time t. They need not consider investor orders that might arrive after time t because they will be able to update their quotes again after the period, and they need not consider sniper orders because no such orders are sent.

As in equilibrium of the baseline model, liquidity providers cancel their quotes after a trade takes place. Thus, the relevant spread for time t is the one that prevails conditional on no trade having taken place in any earlier time period. As in the baseline, this spread, which we denote $s_{N,t}$, is pinned down by a zero-profit condition, which in this case is[23]

$$\frac{s_{N,t}}{2} P^l_{N,t} = \left(1 - \frac{s_{N,t}}{2}\right) P^i_{N,t}.$$

[23]To clarify, the probabilities needed to formulate this zero-profit condition are conditional ones: the probability that an information investor order (or a liquidity investor order) arrives at a given exchange at time t conditional on no order having arrived at either exchange at all times $s < t$. But because the conditioning event is the same in both cases, we can multiply through by its probability, so that $P^i_{N,t}$ and $P^l_{N,t}$ become the relevant probabilities.

For the cases in which $N > 3$, we therefore have

$$s_{N,t} = \begin{cases} \dfrac{4\lambda}{1+\lambda} & \text{if } t \in \{\frac{1}{N}, \frac{2}{N}\} \\ \dfrac{4\lambda[p_I + (1-p_I)^2]}{1-\lambda+2\lambda[p_I+(1-p_I)^2]} & \text{if } t \in \{\frac{3}{N}, \frac{4}{N}, \ldots, \frac{N-2}{N}\} \\ \dfrac{4\lambda(1-p_I)}{(1-\lambda)+2\lambda(1-p_I)} & \text{if } t \in \{\frac{N-1}{N}, 1\}. \end{cases}$$

For the case in which $N = 3$, the expression is the same except that $s_{3,2/3}$ is undefined. Crucially, the breakeven spread $s_{N,t}$ is not constant in t, but rather declines over time (as depicted in Figure IA.3 for the case in which $\lambda = 0.5$ and $p_I = 0.75$).

The intuition for why $s_{N,t}$ fails to be constant is most clearly seen in the case of $N = 3$, in which case the investor sends orders only at $t = 0$. Because a liquidity investor sends a single order, while an information investor sends multiple orders at once, an information investor's minimum latency (where the minimum is taken across the orders he sends) is first-order stochastically dominated by a liquidity investor's latency. Thus, when no order arrives at either exchange at $t = \frac{1}{3}$, that is a signal that the investor is more likely to be liquidity-motivated, which causes the liquidity provider to set a tighter spread at $t = 1$. However, this inference depends fundamentally on the liquidity provider's knowledge that investor orders are sent precisely at the time $t = 0$. In practice, traders would not possess such knowledge and would therefore be unable to make such inferences. Removing the ability of traders to make these unrealistic inferences is exactly what is accomplished in the model by increasing N.

In summary, we have illustrated that when time periods are indexed by the finite set $\mathcal{T}_N$, the spread $s_{N,t}$ will no longer be constant in t. We therefore lose the tractability of a stationary equilibrium. This is true even in the simple parametrization of the model considered above (with $X = 2$, $p_H = 1$, and $c(r) \equiv 0$). Further complications and challenges arise outside of that special case.

G.2. Convergence

As N diverges, $s_{N,t}$ converges to a constant—namely, to the equilibrium spread of the baseline model, s^*_{LOB}—in the pointwise almost everywhere sense depicted in Figure IA.3. At its core, what increasing N does is to make the scale of the latency distribution progressively smaller relative to the scale of the distribution governing investor arrival times. Because latencies are incredibly small in practice—on the order of microseconds—very large values of N are most realistic. At these large values, $s_{N,t}$ is very nearly constant, so that the challenges

stemming from its failure to be perfectly constant would seem to be merely a sideshow that distracts from the economic forces on which we are trying to focus. To sidestep these distractions, our approach is to work directly *in* the limit, where the spread is perfectly constant and these challenges do not arise.

G.3. Alternative Trading Mechanisms

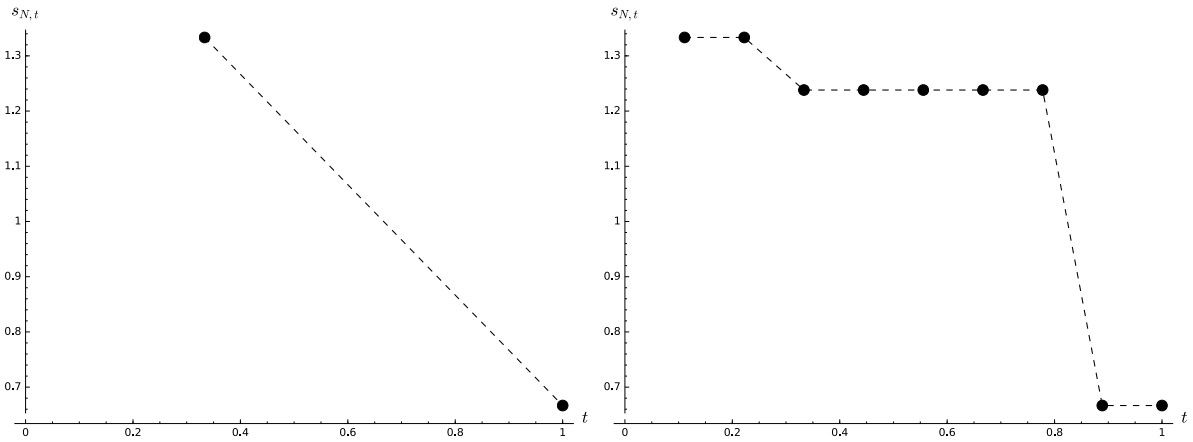
This analysis above focuses on one difficulty that would arise without a hyperreal construction of time. However, a second difficulty is that alternatives to the hyperreal construction do not lend themselves as easily to clean formalizations of the alternative trading mechanisms that we consider. The hyperreal numbers contain infinitesimals that are infinitely larger than ε (e.g., $\sqrt{\varepsilon}$). And in modeling NDs and FBAs, we leverage the existence of such quantities to deliver a clean analysis. Under alternative modeling approaches, analogues of these quantities would be unavailable. We would then be forced to deal with inelegant distractions similar to those highlighted by the analysis above, in Internet Appendix III.G.1.

H. Extension: Large Liquidity Investors

In the baseline model, each liquidity investor seeks to trade just a single share, and each attempts to do so by trading on only a single exchange, his "home exchange." In this appendix, we consider a version of the model in which liquidity investors are heterogeneous in the number of shares that they seek to trade and in which larger liquidity investors split their orders across exchanges. We characterize the limit order book equilibrium in this setting, and we argue that the equilibrium is qualitatively similar to the equilibrium that we derive in our baseline analysis. In particular, both passive-side and aggressive-side order anticipation remain features of that equilibrium.

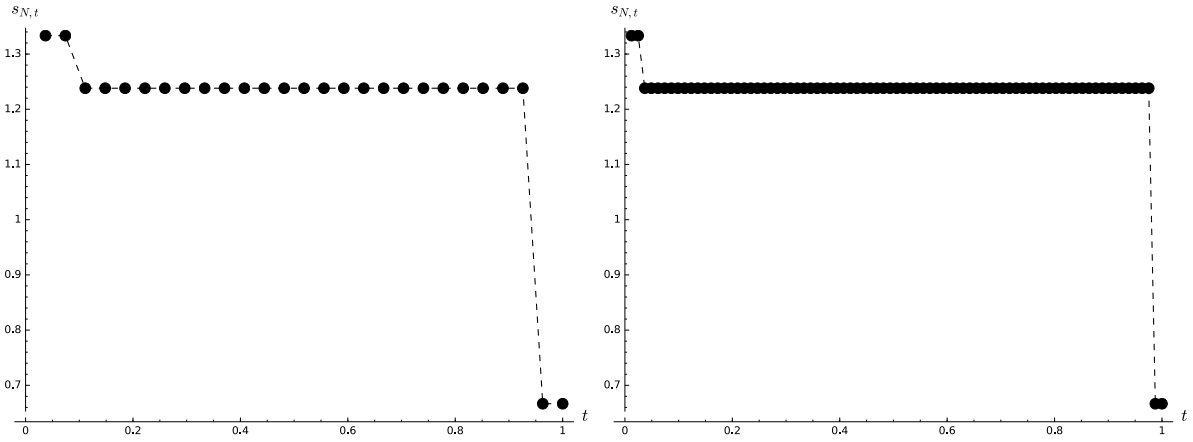
We enrich the model by supposing that the investor may be one of three types (whereas only two types were possible in the baseline model):

- An information investor, with probability λ. This type is identical to what we referred to in the baseline model as the information investor type.
- A small liquidity investor, with probability $(1-\lambda)\alpha$. This type is identical to what we referred to in the baseline model as the liquidity investor type. In particular, the investor would have a liquidity demand that is satisfied by trading a single share. The investor would have a single "home exchange" that is chosen uniformly at random from the set of exchanges, and would be restricted to sending orders only to that exchange.



Panel A. $N = 3$

Panel B. $N = 9$



Panel C. $N = 27$

Panel D. $N = 81$

Figure IA.3. Breakeven spreads for finite time periods. The figure plots the breakeven spread $s_{N,t}$, which is defined in the text, over times $t \in \mathcal{T}_N$ for different finite values of N for the special case in which $X = 2$, $p_H = 1$, $c(r) \equiv 0$, $\lambda = 0.5$, and $p_I = 0.75$.

- A large liquidity investor, with probability $(1-\lambda)(1-\alpha)$. This type is identical to what we referred to in the baseline model as the liquidity investor type, but with the following exceptions. The investor would have a liquidity demand that is satisfied by trading two shares. The investor would have two "home exchanges" that are drawn uniformly at random (without replacement) from the set of exchanges, and would be restricted to sending orders only to those exchanges.

To make our analysis of this extension tractable, we assume the following parametric restrictions. Research is costless (i.e., $c(r) \equiv 0$), HFTs always obtain the lowest latency (i.e., $p_H = 1$), and there are at least three exchanges (i.e., $X \geq 3$). The exact nature of the equilibrium depends on the parameters and falls into one of two cases.

H.1. Case A

In the first case, α is close to one, so that the model is close to a special case of the baseline model. As might then be expected, the equilibrium resembles the equilibrium of the baseline model. In particular, snipers initiate aggressive-side order anticipation after two or more trades take place. In terms of the yet-to-be-specified quantities $(s^*_{A,1}, s^*_{A,2})$, we conjecture that (and subsequently check whether) the following profile of strategies constitutes an equilibrium. As will become clear, $s^*_{A,1}$ should be interpreted as the initial spread, and $s^*_{A,2}$ should be interpreted as the "updated spread" that arises in the event that only one trade takes place in the period in which the first trade occurs.

- *Investor.* If he is a small liquidity investor with a buying (selling) motive, then he sends to his home exchange an immediate-or-cancel order to buy (sell) one share at the price β $(-\beta)$.

  If he is a large liquidity investor with a buying (selling) motive, then he sends to each of his two home exchanges an immediate-or-cancel order to buy (sell) one share at the price β $(-\beta)$.

  If he is an information investor, then he conducts research with intensity $r = 1$. If he learns that the value of the security is $v = 1$ $(v = -1)$, he sends to each exchange an immediate-or-cancel order to buy (sell) one share at the price 1 (-1).

- *Liquidity providers.* One liquidity provider is active on the equilibrium path ("the liquidity provider"). At time $t = 0$, she sends to each exchange a post-only order to buy one share at the bid $-s^*_{A,1}/2$ and another to sell one share at the ask $s^*_{A,1}/2$. If at any time t a

trade occurs at the ask (bid) at exactly one exchange, then she sends replacement orders for all of her remaining orders to sell (buy) at the price $s^*_{A,2}/2$ ($-s^*_{A,2}/2$), and she sends cancellations for all of her orders to buy (sell). If by any time t two or more trades have occurred, then she sends cancellations for all of her remaining orders.

A second liquidity provider who is inactive on path but may be active off path is referred to as "the enforcer." If at some time $t \geq 3\varepsilon$ prior to which no trade has occurred the LOB at some exchange consists of anything other than a post-only order to buy one share at the bid $-s^*_{A,1}/2$ and another to sell one share at the ask $s^*_{A,1}/2$, then she sends such orders to that exchange, doing so in that same period. And if at some time $t \geq 3\varepsilon$ for which at time $t - \varepsilon$ a trade occurred at the ask (bid) at exactly one exchange, the LOB at some exchange consists of anything other than a post-only order to sell (buy) at the price $s^*_{A,2}/2$ ($-s^*_{A,2}/2$), then she sends such orders to that exchange.

The remaining liquidity providers remain completely inactive both on and off path.

- *Snipers.* If, by any time t, trades have occurred at the ask (bid) at two or more exchanges, then each sniper sends to all other exchanges an immediate-or-cancel order to buy (sell) one share at the price 1 (-1), doing so in that same period.

Crucially, given these conjectured equilibrium strategies, both passive-side and aggressive-side order anticipation occur in this alternative model. In what follows, we derive expressions for the quantities $(s^*_{A,1}, s^*_{A,2})$, and we derive restrictions on the parameters—which we express as a lower bound on α—under which the conjectured strategies do indeed constitute an equilibrium.

We begin by using Bayes' rule to derive the beliefs that HFTs must possess if the above conjecture about strategies constitutes a WPBE. In particular, we compute their expectation of the value of the security in the time period in which the first trade occurs. We define V_n as the expectation of the value of the security, under HFT beliefs in the time period in which the first trade occurs, conditional on a trade occurring at the ask at exactly n exchanges in that period. Under these conjectured strategies, Bayes' rule implies that HFTs must have the following beliefs:

$$V_n = \begin{cases} \dfrac{\lambda X p_I (1-p_I)^{X-1}}{\lambda X p_I (1-p_I)^{X-1} + (1-\lambda)\alpha + (1-\lambda)(1-\alpha)2p_I(1-p_I)} & \text{if } n = 1 \\ \dfrac{\lambda \frac{X(X-1)}{2} p_I^2 (1-p_I)^{X-2}}{\lambda \frac{X(X-1)}{2} p_I^2 (1-p_I)^{X-2} + (1-\lambda)(1-\alpha)p_I^2} & \text{if } n = 2 \\ 1 & \text{if } n \geq 3. \end{cases} \tag{IA17}$$

Given that the asset value distribution is symmetric about zero, it follows that for the case of trades at the bid, beliefs are the negative of the aforementioned.

The next step is to derive the zero-profit condition that will pin down the initial spread $s^*_{A,1}$. That condition can be built up from the following three cases:

- First, suppose the investor is a small liquidity investor. Then the investor always completes exactly one trade at the initial quotes. Snipers initiate no additional trades.
- Second, suppose the investor is a large liquidity investor. Then the investor completes two trades at the initial quotes with probability $p_I^2 + (1-p_I)^2$ and completes one trade at the initial quotes with probability $2p_I(1-p_I)$. He therefore completes $2p_I + 2(1-p_I)^2$ such trades in expectation. Snipers initiate trades only in the event that the investor completes two trades, in which case they complete an additional $X-2$ trades at the initial quotes. They therefore complete $(X-2)(p_I^2+(1-p_I)^2)$ such trades in expectation.
- Third, suppose the investor is an information investor. Then just as in the baseline model, the investor completes $Xp_I + X(1-p_I)^X$ trades at the initial quotes in expectation. And just as in the baseline model, snipers complete an additional $X(1-p_I) - X(1-p_I)^X - (X-1)Xp_I(1-p_I)^{X-1}$ such trades in expectation.

Thus, the zero-profit condition for the initial spread $s^*_{A,1}$ is

$$\begin{aligned} &(1-\lambda)\alpha\frac{s^*_{A,1}}{2} + (1-\lambda)(1-\alpha)\left[2p_I + 2(1-p_I)^2 + (X-2)(p_I^2+(1-p_I)^2)\right]\frac{s^*_{A,1}}{2} \\ &\quad = \lambda\left[X - (X-1)Xp_I(1-p_I)^{X-1}\right]\left(1 - \frac{s^*_{A,1}}{2}\right). \end{aligned}$$

Solving that for the spread, we obtain

$$s^*_{A,1} = \frac{2X\left(\lambda - (X-1)\lambda p_I (1-p_I)^{X-1}\right)}{(X-1)\alpha\lambda - 2\left((X-1)\alpha - ((X-1)\alpha - X + 1)\lambda - X + 1\right)p_I^2 - (X-1)\alpha - \left((X^2-X)\lambda(1-p_I)^{X-1} - 2(X-1)\alpha + 2\left((X-1)\alpha - X + 1\right)\lambda + 2X - 2\right)p_I + X}.$$

In addition, optimality of the sniper strategy—in particular, optimality of the response to the situation in which exactly two trades occur in the time period in which the first trades take place—requires that $\frac{s^*_{A,1}}{2} \leq V_2$. Plugging in and rearranging, we obtain a lower bound on the fraction of small liquidity investors, $\alpha \geq \bar{\alpha}$, where

$$\bar{\alpha} = \frac{2(X-1)^2 p_I^4 (1-p_I)^{X-2} + 2\left(2(X-1)(1-p_I)^{X-1} - (X-1)^2(1-p_I)^{X-2}\right)p_I^3 - \left(4(X-1)(1-p_I)^{X-1} - \left(X^2-X\right)(1-p_I)^{X-2} + 4\right)p_I^2 + 2\left((X-1)(1-p_I)^{X-1} + 2\right)p_I - 2}{2(X-1)^2 p_I^4 (1-p_I)^{X-2} + 2\left(2(X-1)(1-p_I)^{X-1} - (X-1)^2(1-p_I)^{X-2}\right)p_I^3 - \left(4(X-1)(1-p_I)^{X-1} - (X-1)^2(1-p_I)^{X-2} + 4\right)p_I^2 + 2\left((X-1)(1-p_I)^{X-1} + 2\right)p_I - 2}. \tag{IA18}$$

It only remains to derive the "updated spread" $s^*_{A,2}$, which arises in the event that only one trade takes place in the period in which the first trade occurs. To build up the zero-profit condition that pins it down, we again consider three cases:

- First, suppose the investor is a small liquidity investor. In that case, he will not trade against the updated quotes.
- Second, suppose the investor is a large liquidity investor. In that case, he will trade one share against the updated quotes in the event that he traded exactly one share against the initial quotes. From above, this event occurs with probability $2p_I(1-p_I)$.
- Third, suppose the investor is an information investor. In that case, he will trade $X-1$ shares against the updated quotes in the event that he traded exactly one share against the initial quotes. This event occurs with probability $Xp_I(1-p_I)^{X-1}$.

And in all cases, snipers will not trade against the updated quotes. Thus, the zero-profit condition for the updated spread $s^*_{A,2}$ is

$$(1-\lambda)(1-\alpha)2p_I(1-p_I)\frac{s^*_{A,2}}{2} = \lambda X p_I (1-p_I)^{X-1}(X-1)\left(1 - \frac{s^*_{A,2}}{2}\right).$$

Solving for the spread yields

$$s^*_{A,2} = \frac{2\left(X^2-X\right)\lambda(1-p_I)^{X-1}}{\left(X^2-X\right)\lambda(1-p_I)^{X-1} + 2(\alpha-1)\lambda - 2\left((\alpha-1)\lambda - \alpha + 1\right)p_I - 2\alpha + 2}. \tag{IA19}$$

H.2. Case B

In the second case, α is further from one, and the equilibrium is somewhat different from above. In particular, snipers wait for a stronger signal of informed trading before initiating aggressive-side order anticipation—they now wait until after *three* or more trades take place. In terms of the yet-to-be-specified quantities $(s^*_{B,1}, s^*_{B,2})$, we conjecture that (and

subsequently check whether) the following profile of strategies constitutes an equilibrium. As before, $s^*_{B,1}$should be interpreted as the initial spread, and $s^*_{B,2}$ should be interpreted as the "updated spread" that arises in the event that only one trade takes place in the period in which the first trade occurs.

The strategies for the investor and the liquidity providers are identical to those in Case A. However, those for snipers differ:

- *Snipers.* If, by any time t, trades have occurred at the ask (bid) at three or more exchanges, then each sniper sends to all other exchanges an immediate-or-cancel order to buy (sell) one share at the price 1 (-1), doing so in that same period.

Crucially, given these conjectured equilibrium strategies, both passive-side and aggressive-side order anticipation occur in this alternative model. There may be less aggressive-side order anticipation than in our baseline analysis because snipers now require stronger signals to react, but some amount of order anticipation would nevertheless continue to take place (at least when there are four or more exchanges). In what follows, we derive expressions for the quantities $(s^*_{B,1}, s^*_{B,2})$, and we derive restrictions on the parameters—which we express as an upper bound on α—under which the conjectured strategies constitute an equilibrium.

The beliefs of HFTs are identical to those specified by equation (IA17) in Case A. The next step is to derive the zero-profit condition that pins down the initial spread $s^*_{B,1}$. That condition can be built up from the following three cases:

- First, suppose the investor is a small liquidity investor. Then the investor always completes exactly one trade at the initial quotes. Snipers initiate no additional trades.
- Second, suppose the investor is a large liquidity investor. Then the investor completes two trades at the initial quotes with probability $p_I^2 + (1-p_I)^2$ and completes one trade at the initial quotes with probability $2p_I(1-p_I)$. He therefore completes $2p_I + 2(1-p_I)^2$ such trades in expectation. Snipers initiate no additional trades.
- Third, suppose the investor is an information investor. Then just as in the baseline model, the investor completes $Xp_I + X(1-p_I)^X$ trades at the initial quotes in expectation. And snipers complete an additional $X - Xp_I - X(1-p_I)^X - (X-1)Xp_I(1-p_I)^{X-1} - (X-2)\frac{X(X-1)}{2}p_I^2(1-p_I)^{X-2}$ such trades in expectation.

Thus, the zero-profit condition for the initial spread $s^*_{B,1}$ is

$$(1-\lambda)\alpha\frac{s^*_{B,1}}{2} + (1-\lambda)(1-\alpha)\left[2p_I + 2(1-p_I)^2\right]\frac{s^*_{B,1}}{2}$$
$$= \lambda\left[X - (X-1)Xp_I(1-p_I)^{X-1} - (X-2)\frac{X(X-1)}{2}p_I^2(1-p_I)^{X-2}\right]\left(1-\frac{s^*_{B,1}}{2}\right).$$

Solving the equation above for the spread, we obtain

$$s^*_{B,1} = \frac{2\left(\left(X^3-3X^2+2X\right)\lambda p_I^2(1-p_I)^{X-2}+2\left(X^2-X\right)\lambda p_I(1-p_I)^{X-1}-2X\lambda\right)}{\left(\left(X^3-3X^2+2X\right)\lambda(1-p_I)^{X-2}-4(\alpha-1)\lambda+4\alpha-4\right)p_I^2-2(X+\alpha-2)\lambda+2\left(\left(X^2-X\right)\lambda(1-p_I)^{X-1}+2(\alpha-1)\lambda-2\alpha+2\right)p_I+2\alpha-4}.$$

In addition, optimality of sniper strategy—in particular, optimality of their response to the situation in which exactly two trades occur in the time period in which the first trades take place—requires that $\frac{s^*_{B,1}}{2} \geq V_2$. Plugging in and rearranging, we obtain an upper bound on the fraction of small liquidity investors, $\alpha \leq \bar{\alpha}$, where $\bar{\alpha}$ is as defined in (IA18). Recall that in Case A, the necessary parametric restriction was just the opposite, that is, $\alpha \geq \bar{\alpha}$. Thus, for any set of parameter values, either the Case A strategies or the Case B strategies constitute a WPBE.

It remains to derive the updated spread $s^*_{B,2}$. Given that snipers do not trade against these updated quotes and that strategies for the remaining traders are identical to those of Case A, the zero-profit condition for the updated spread is as above, that is,

$$(1-\lambda)(1-\alpha)2p_I(1-p_I)\frac{s^*_{B,2}}{2} = \lambda X p_I(1-p_I)^{X-1}(X-1)\left(1-\frac{s^*_{B,2}}{2}\right).$$

Therefore, $s^*_{B,2} = s^*_{A,2}$, as defined in equation (IA19).

I. Extension: Risk-Averse Information Investor

In this appendix we study equilibrium behavior in the case of a risk-averse information investor. Relative to our baseline analysis, the primary difference is that such an investor, conditional on learning the security value, may send orders to a strict subset of the exchanges (in fact, he may randomize over the cardinality of that set). This is in contrast to the baseline of risk neutrality in which information investors send orders to all X exchanges. Aside from this, however, the LOB equilibrium remains similar to our baseline analysis (as described in Section III of the main article), and most of its qualitative features remain intact. In particular, aggressive-side and passive-side order anticipation continue to occur on path.

Finally, numerical experimentation suggests that the main comparative static continues to hold: when p_H increases, so that HFTs become faster, the spread weakly decreases.

I.1. Model and Equilibrium

For the purposes of this appendix, we modify our setting as follows. Let u be the utility function of information investors. In contrast to our baseline analysis, we do not assume that u is linear. Another modification is that we allow for the possibility that successful research results in obtaining only an imperfect signal of the value of the security. Let the security value be $v = \tilde{v} + \varepsilon$. As before, if an information investor conducts research with intensity r, then he learns $\tilde{v} \sim \text{unif}\{-1, 1\}$ with probability r and learns no information otherwise. However, he does not learn ε. This additional price risk has no effect under risk neutrality, but it may become relevant under risk aversion, which is why we allow for it here. Whereas we allow shares to be divisible in our baseline analysis, we require them to be indivisible for the purposes of this appendix. This simplifies some of the derivations without significantly altering the qualitative insights that come out of this analysis.

Let Δ denote the set of probability distributions over $\{0, 1, 2, \ldots, X\}$. A typical element of Δ will be denoted $\xi = (\xi_0, \xi_1, \xi_2, \ldots, \xi_X)$. In terms of the yet-to-be-specified quantities (s^*, r^*, ξ^*)—where $s^* \in [0, \infty)$, $r^* \in [0, 1]$, and $\xi^* \in \Delta$—we conjecture that (and subsequently check whether) the following profile of strategies constitutes an equilibrium.

- *Investor.* If he is a liquidity investor with a buying (selling) motive, then he sends to his home exchange an immediate-or-cancel order to buy (sell) one share at the price β ($-\beta$).

  If he is an information investor, then he conducts research with intensity r^*. If he learns that the value of the security is $\tilde{v} = 1$ ($\tilde{v} = -1$), then he draws a number $y \in \{0, 1, 2, \ldots, X\}$ from the distribution ξ^*. Given this realization, he chooses y exchanges uniformly at random and sends to each of them an immediate-or-cancel order to buy (sell) one share at the price 1 (-1). He sends no orders if he does not learn $\tilde{v}$.

- *Liquidity providers.* One liquidity provider is active on the equilibrium path ("the liquidity provider"). At time $t = 0$, she sends to each exchange a post-only order to buy one share at the bid $-s^*/2$ and another to sell one share at the ask $s^*/2$. If at any time t one or more trades occur, then she sends cancellation orders for all of her remaining orders, doing so in that same period.

  A second liquidity provider who is inactive on path but may be active off path is referred

to as "the enforcer." If at some time $t \geq 3\varepsilon$ prior to which no trade has occurred the LOB at some exchange consists of anything other than a post-only order to buy one share at $-s^*/2$ and a post-only order to sell one share at $s^*/2$, then she sends such orders to that exchange, doing so in that same period.

The remaining liquidity providers remain completely inactive both on and off path.

- *Snipers.* If, at any time t, trades occur at the ask (bid) at two or more exchanges, then each sniper sends to all other exchanges an immediate-or-cancel order to buy (sell) one share at the price 1 (-1), doing so in that same period.

Just as in the baseline of risk neutrality, two conditions must be satisfied for the above to constitute an equilibrium: (*i*) information investor optimization and (*ii*) a zero-profit condition for spread. Below, we characterize what is entailed by these two conditions. Relative to the baseline, the primary difference is that an information investor's choice of the number of exchanges to target now becomes nontrivial. Under risk neutrality as in the baseline model, it is optimal to target all exchanges—in terms of the above notation, to choose $\xi^* = (0, 0, 0, \ldots, 1)$. But under risk aversion, this may no longer be optimal.

To proceed, we begin by defining $p_y(f)$ as the probability that an investor who sends y orders receives f fills, given the aforementioned behavior of the other traders. In the case of $y > 1$, $p_y(f)$ is as follows:

$$p_y(f) = \begin{cases} y p_I (1-p_I)^{y-1} p_H^{y-1} & \text{if } f = 1 \\ \binom{y}{f} p_I^f (1-p_I)^{y-f} + y p_I (1-p_I)^{y-1} \binom{y-1}{f-1} (1-p_H)^{f-1} p_H^{y-f} & \text{if } f < y \\ (1-p_I)^y + p_I^y + y p_I (1-p_I)^{y-1} (1-p_H)^{y-1} & \text{if } f = y \\ 0 & \text{otherwise.} \end{cases}$$

In the trivial cases, we have $p_0(0) = p_1(1) = 1$.

The first requirement for equilibrium derives from a zero-profit condition that the spread s^* must satisfy. The implied restriction takes essentially the same form as in the baseline model, although it is not precisely identical unless $\xi^* = (0, 0, 0, \ldots, 1)$:

$$\sum_{y=0}^{X} \xi_y^* \left[(1-\lambda)\frac{s^*}{2} - \lambda r^* \left(\tilde{X}_S(y) + \tilde{X}_I(y) \right) \left(1 - \frac{s^*}{2} \right) \right] = 0, \tag{IA20}$$

where

$$\tilde{X}_I(y) = \sum_{f=0}^{y} f p_y(f)$$

$$\tilde{X}_L(y) = \begin{cases} X-1 & \text{if } y = 1 \\ y p_I (1-p_I)^{y-1} \left[p_H (X-1) + p_H^{y-1}(1-p_H)(X-y) \right] & \text{otherwise} \end{cases}$$

$$\tilde{X}_S(y) = \begin{cases} 0 & \text{if } y = 0 \\ X - \tilde{X}_I(y) - \tilde{X}_L(y) & \text{otherwise.} \end{cases}$$

Echoing the notation used in the baseline analysis, $\tilde{X}_I(y)$ and $\tilde{X}_S(y)$ represent the expected number of trades made by an information investor and snipers, respectively, conditional on an information investor sending orders to y exchanges. In addition, we use $\tilde{X}_L(y)$ for the expected number of cancellations made by the liquidity provider conditional on an an information investor sending orders to y exchanges. The intuition for the expressions for these three quantities is as follows:

- From the definition of $p_y(f)$, it follows that $\tilde{X}_I(y) = \sum_{f=0}^{y} f p_y(f)$.
- Given the strategy profile delineated above, it follows that $\tilde{X}_L(1) = X - 1$. In all other cases, the liquidity provider is successful in cancelling orders only in the event that exactly one of the information investor's y orders receives a short latency draw, which occurs with probability $y p_I (1-p_I)^{y-1}$. Conditional on this event, the liquidity provider is successful in cancelling at one of the remaining $X-1$ exchanges if she obtains a short latency draw for the cancellation order sent there, resulting in $p_H(X-1)$ cancellations in expectation. Finally, $p_H^{y-1}(1-p_H)(X-y)$ is a correction term to account for the fact that if the liquidity provider obtains short latency draws for all of the remaining $y-1$ exchanges targeted by the information investor, then she also successfully cancels at all of the $X-y$ other exchanges (even those for which she obtains high latency draws).
- Given the strategy profile delineated above, it follows that $\tilde{X}_S(0) = 0$ and that for $y > 0$, $\tilde{X}_I(y) + \tilde{X}_L(y) + \tilde{X}_S(y) = X$.

The second requirement for equilibrium is that the information investor optimally chooses r and ξ, taking the spread s^* as given, that is,

$$(r^*, \xi^*) \in \underset{r \in [0,1], \xi \in \Delta}{\arg\max}\, \mathbb{E} \left\{ r \sum_{y=0}^{X} \xi_y \sum_{f=0}^{y} p_y(f) u \left(f \left[1 + \varepsilon - \frac{s^*}{2} \right] - c(r) \right) + (1-r) u\left(-c(r)\right) \right\}, \quad \text{(IA21)}$$

where the expectation is taken over ε. Given the above, the aforementioned strategy profile constitutes an equilibrium if the tuple (r^*, ξ^*, s^*) is a solution to the system (IA20) and (IA21).

I.2. Numerical Example

For the purposes of what follows, we assume that research is costless, i.e., that $c(r) \equiv 0$. Note that the objective (IA21) then becomes linear in r. Moreover, it evaluates to $u(0)$ not only for a choice of $r = 0$ but also for a choice of $\xi = (1, 0, 0, \ldots, 0)$. For these reasons, it is essentially without loss to focus on solutions in which $r^* = 1$, in which case (IA21) reduces to

$$\xi^* \in \arg\max_{\xi \in \Delta} \mathbb{E}\left\{\sum_{y=0}^{X} \xi_y \sum_{f=0}^{y} p_y(f) u\left(f\left[1 + \varepsilon - \frac{s^*}{2}\right]\right)\right\}. \tag{IA22}$$

In what follows, we describe a procedure that allows one to solve for an equilibrium that is pure (in the sense that ξ^* is degenerate) if one exists. We then illustrate with a parametrized example. Further below, we describe a procedure that allows one to solve for an equilibrium in mixed strategies (in the sense that ξ^* is nondegenerate) when a pure equilibrium fails to exist, and we illustrate with the same parametrized example.

Pure strategy equilibria. We begin by searching for pure strategy equilibria. In these equilibria, the information investor chooses a degenerate distribution ξ^*, which puts all weight on a single $y^* \in \{0, 1, 2, \ldots, X\}$. In that case, the problem of finding an equilibrium reduces to finding a pair (y^*, s^*), where s^* is derived from y^* through (IA20) and where y^* is a fixed point of the equation obtained by plugging (IA20) into (IA22):

$$y^* \in \arg\max_{y \in \{0,1,2,\ldots,X\}} \mathbb{E}\left\{\sum_{f=0}^{y} p_y(f) u\left(f\left[\frac{1-\lambda}{1-\lambda+\lambda r^*(\tilde{X}_I(y^*) + \tilde{X}_S(y^*))} + \varepsilon\right]\right)\right\}. \tag{IA23}$$

The objective of this maximization can be interpreted as an information investor's expected utility derived by deviating to a choice of y from a putative equilibrium involving a choice of y^*. The desired fixed points of (IA23) can be derived using a guess-and-verify approach.

Next, we illustrate this guess-and-verify approach in Figure IA.4, under the following parametric assumptions: (*i*) $u(w) = -\exp(-\gamma w)$, so that γ represents the coefficient of absolute risk aversion, (*ii*) $\varepsilon \sim \mathcal{N}(0, 1)$, (*iii*) $\lambda = 0.3$, (*iv*) $X = 3$, (*v*) $p_H = 1$, and (*vi*) $p_I = 0.5$. But to illustrate the effects of risk aversion, we leave γ unrestricted. The objective in (IA23)

now reduces to

$$U_\gamma(y \mid y^*) = \sum_{f=0}^{y} p_y(f)\left[-\exp\left(\frac{-0.7\gamma f}{0.7 + 0.3(\tilde{X}_I(y^*) + \tilde{X}_S(y^*))} + \frac{\gamma^2}{2} f^2\right)\right]. \tag{IA24}$$

In terms of this notation, (IA23) can be written as $y^* \in \arg\max_y U_\gamma(y \mid y^*)$. The four panels of Figure IA.4 correspond to different guesses of y^*. For example, Panel C is based on the guess $y^* = 2$. In that panel, when $\gamma = 0.5$, utility is highest when $y = 2$, which establishes that we do indeed have an equilibrium with $y^* = 2$ when $\gamma = 0.5$. The figure also reveals that, in this case, y^* depends on γ as follows:

$$y^* = \begin{cases} 3 & \text{if } \gamma \in (0, 0.235] \\ 2 & \text{if } \gamma \in [0.249, 0.359] \\ 1 & \text{if } \gamma \in [0.467, 1.400] \\ 0 & \text{if } \gamma \in [2.000, \infty). \end{cases} \tag{IA25}$$

This approach does not always isolate an equilibrium in pure strategies, as is illustrated by the fact that the domain of γ in equation (IA25) does not span all of $\mathbb{R}_+$. For the remaining values of γ, it is nevertheless possible to construct a mixed strategy equilibrium using the method that we describe next.

Mixed strategy equilibria. For convenience, we define $V(y \mid s^*)$ to be the objective of (IA22) for the case in which ξ is a degenerate distribution that puts all its weight on y:

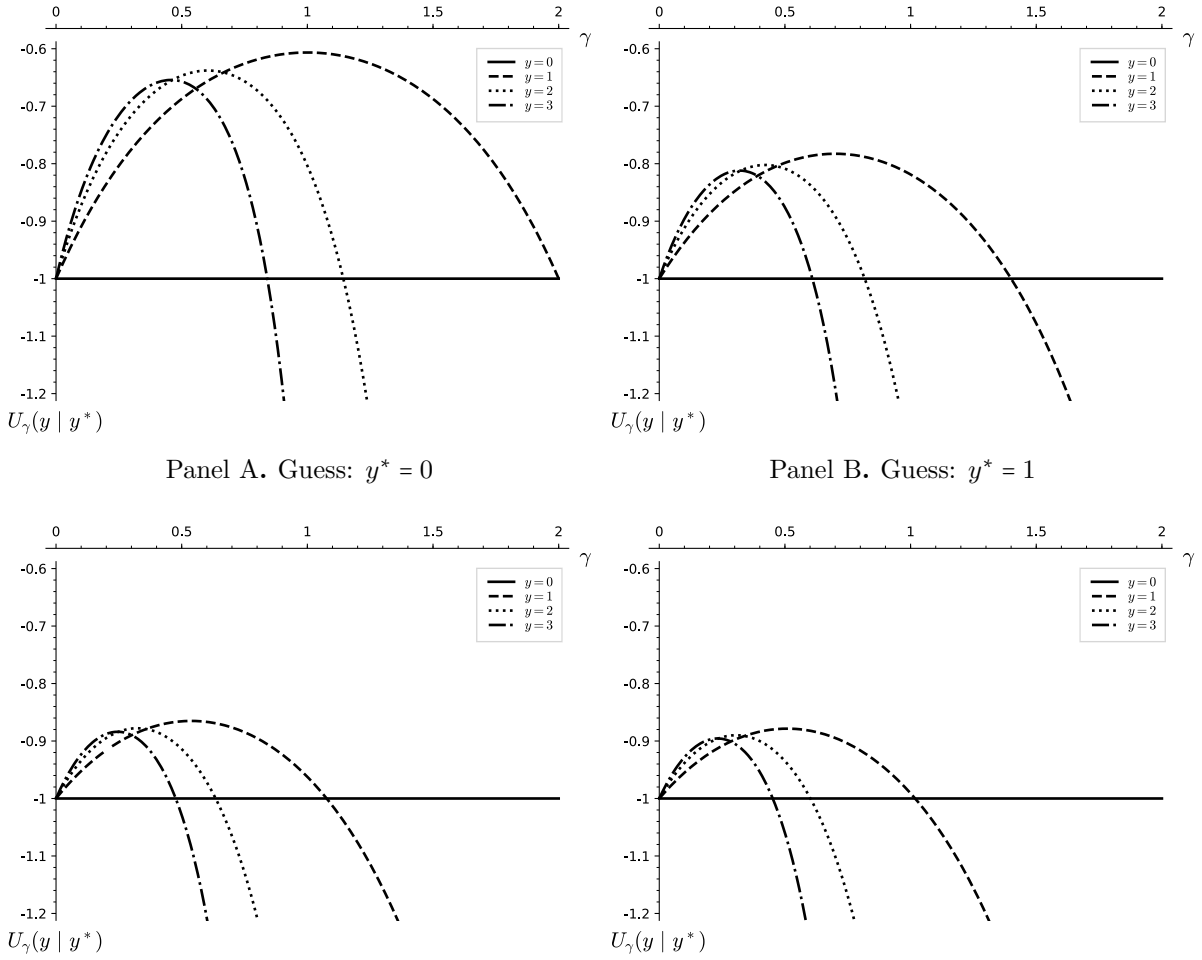
$$V(y \mid s^*) = \mathbb{E}\left\{\sum_{f=0}^{y} p_y(f) u\left(f\left[1 + \varepsilon - \frac{s^*}{2}\right]\right)\right\}. \tag{IA26}$$

In words, V provides the expected utility of an information investor from a choice of y given a postulated equilibrium spread s^*. It will also be convenient to define $\pi(y, s)$ as the profits from providing liquidity at spread s when an information investor sends orders to y exchanges:

$$\pi(y, s) = (1 - \lambda)\frac{s}{2} - \lambda\big(\tilde{X}_S(y) + \tilde{X}_I(y)\big)\left(1 - \frac{s}{2}\right).$$

The following guess-and-verify approach can be used to search for equilibria:

1. For a candidate value $y \in \{0, 1, 2, \ldots, X-1\}$, compute s^* to solve $V(y \mid s) = V(y+1 \mid s)$, provided that such a solution exists.



Panel A. Guess: $y^* = 0$

Panel B. Guess: $y^* = 1$

Panel C. Guess: $y^* = 2$

Panel D. Guess: $y^* = 3$

Figure IA.4. Expected Utility of an Information Investor. The figure illustrates the guess-and-verify approach for deriving pure equilibria that is described in the text under the parametric assumptions described in the text. The panels correspond to different guesses $y^* \in \{0, 1, 2, 3\}$. Within each panel, the lines depict, against the degree of risk aversion γ, the expected utility $U_\gamma(y \mid y^*)$, as defined in equation (IA24), of an information investor for the different choices of y: (i) $y = 0$ (solid line), (ii) $y = 1$ (dashed line), (iii) $y = 2$ (dotted line), and (iv) $y = 3$ (dash-dotted line).

2. Define ξ^* as the probability distribution with support $\{y, y+1\}$ in which $\xi^*_{y+1} = \frac{\pi(y,s^*)}{\pi(y,s^*)-\pi(y+1,s^*)}$ and $\xi^*_y = 1 - \xi^*_{y+1}$, provided that it is a well-defined probability distribution.

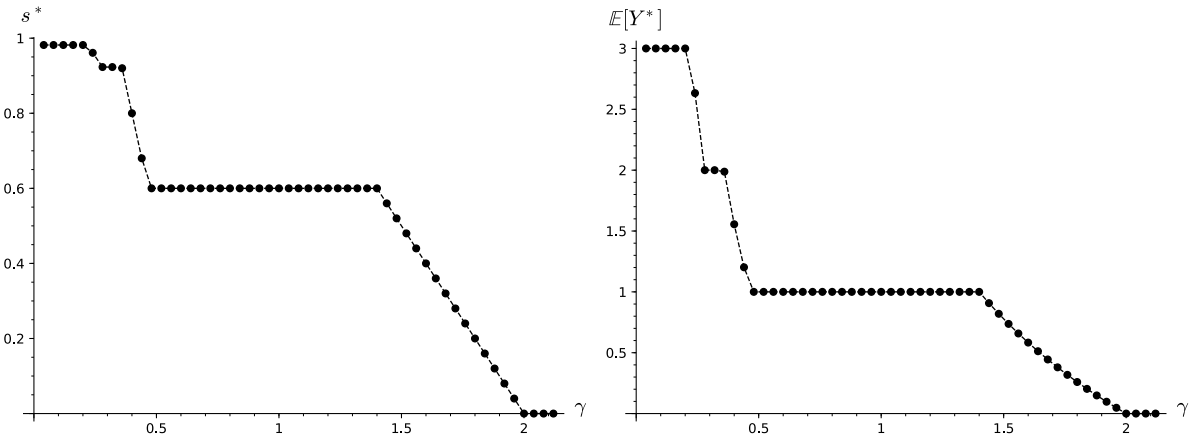
3. Verify that $V(y' \mid s^*) \leq V(y \mid s^*)$ for all $y' \in \{0, 1, 2, \ldots, X\}$. If verification succeeds, then (ξ^*, s^*) corresponds to a mixed strategy equilibrium.

To illustrate, recall the parametric assumptions considered earlier: (*i*) $u(w) = -\exp(-\gamma w)$, (*ii*) $\varepsilon \sim \mathcal{N}(0,1)$, (*iii*) $\lambda = 0.3$, (*iv*) $X = 3$, (*v*) $p_H = 1$, and (*vi*) $p_I = 0.5$. As evidenced by equation (IA25), our approach does not isolate an equilibrium in pure strategies for $\gamma = 1.5$, which lies between the values of γ for which $y^* = 0$ and $y^* = 1$.[24] Nevertheless, it is possible to derive an equilibrium in this case by allowing the information investor to mix over these two adjacent choices for y. Following the guess-and-verify approach described above, we first find the spread s^* that makes an information investor indifferent between $y = 0$ and $y = 1$. The result is $s^* = 0.5$. We next find the probability distribution ξ^* with support $\{0, 1\}$ under which the zero-profit spread is s^*. The result is $\xi^* = (0.22, 0.78, 0, 0)$. It then only remains to verify that the information investor cannot profitably deviate to choices $y \in \{2, 3\}$, which is easily checked. We use Y^* to represent the random variable whose distribution is given by ξ^*. Thus, $\mathbb{E}[Y^*] = \sum_{y=0}^{X} \xi^*_y y$ is the expected number of orders sent by an information investor, and in this case we have $\mathbb{E}[Y^*] = 0.78$.

This guess-and-verify approach can similarly be used to identify a mixed strategy equilibrium for the other values of γ that are outside the domain of (IA25). Figure IA.5 illustrates the result: Panel A plots s^* against γ, and Panel B plots $\mathbb{E}[Y^*]$ against γ. The "flat" parts of the two panels correspond to the pure strategy equilibria characterized above, and in the case of Panel B, they align with (IA25). The "sloped" parts of the panels correspond to mixed strategy equilibria that are computed using the approach described just above.

Comparative static. Next, we investigate how the equilibrium spread s^* changes as HFT speed improves (i.e., as p_H increases). One of the main results that we are able to prove in the baseline of risk neutrality is that s^* is weakly decreasing in p_H. Although we do not prove that this result is robust to the inclusion of risk aversion, numerical experimentation suggests that it may be. To illustrate, we use the approaches outlined in this subsection to compute equilibria (sometimes in pure strategies and sometimes in mixed strategies) for a selection of

[24] Indeed, for a guess of $y^* = 0$, the information investor has a profitable deviation to $y = 1$ because, using (IA24), $U_{1.5}(0 \mid 0) = -1 < -0.69 = U_{1.5}(1 \mid 0)$. And for a guess of $y^* = 1$, the information investor has a profitable deviation to $y = 0$ because $U_{1.5}(0 \mid 1) = -1 > -1.08 = U_{1.5}(1 \mid 1)$.



Panel A. Equilibrium Spread

Panel B. Expected Number of Targeted Exchanges

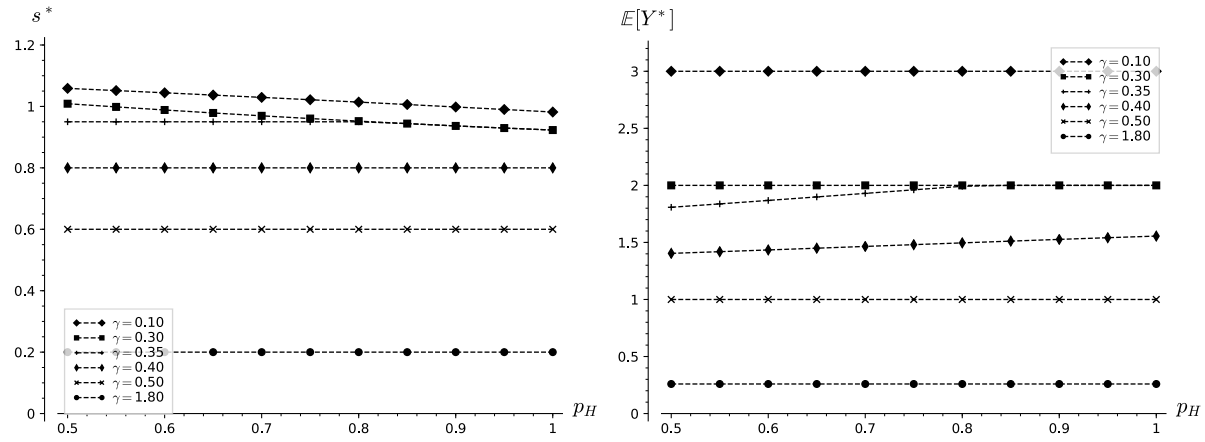
Figure IA.5. How equilibrium quantities depend on risk aversion. The figure plots s^* and $\mathbb{E}[Y^*]$ for different values of γ under the parametric assumptions described in the text. Each point corresponds to an equilibrium (in either pure or mixed strategies) as determined by the approach described in the text.

parameters. The results are depicted in Figure IA.6, with Panel A plotting s^* against p_H for a selection of choices for the risk-aversion parameter γ and with all remaining parameters fixed as in the numerical example analyzed above. Indeed, s^* is weakly decreasing in p_H for all choices of γ. In addition, Figure IA.6 Panel B contains an analogous plot of $\mathbb{E}[Y^*]$, and in so doing it characterizes the information investor routing behavior that is behind the scenes in Panel A.

In conclusion, this appendix illustrates how equilibrium would change in the presence of risk aversion. Specifically, when the information investor is risk averse, equilibrium loses the feature that he always targets all exchanges—in terms of the above notation, choosing $\xi^* = (0, 0, 0, \ldots, 1)$. Nevertheless, equilibrium remains qualitatively similar to the baseline of risk neutrality in the key respect that aggressive-side order anticipation may still take place. Furthermore, it appears that the equilibrium spread remains weakly decreasing in p_H.

J. Equilibrium Uniqueness

Recall that the equilibrium in Section III.A of the main article is not unique. In this appendix, we propose a refinement that we call "plausible equilibrium." The equilibrium described in Section III.A is plausible. Although it is not the unique plausible equilibrium,



Panel A. Equilibrium Spread

Panel B. Expected Number of Targeted Exchanges

Figure IA.6. How equilibrium quantities depend on HFT speed. The figure plots s^* and $\mathbb{E}[Y^*]$ for different values of γ and p_H under the parametric assumptions described in the text. Each point corresponds to an equilibrium (in either pure or mixed strategies) as determined by the approach described in the text.

we go on to argue that every plausible equilibrium is identical to it in terms of the spread and research intensity (s^*_{LOB}, r^*_{LOB}), as characterized by Proposition 1.[25]

To illustrate what plausible equilibrium rules out, consider two of the several alternative equilibria. First, there is an equilibrium in which no orders are ever submitted: because any trade requires two parties, no individual trader would have a unilateral incentive to deviate. Second, there is an equilibrium in which the liquidity provider establishes a spread wider than s^*_{LOB} (from which she earns positive profits), and if ever undercut then responds by immediately shifting to a spread of s^*_{LOB} thereafter. Other liquidity providers do not undercut, and indeed they would not obtain positive profits from doing so because they would own the best quotes for only an infinitesimal length of time. Both of these equilibria will fail to be plausible. Respectively, they will be ruled out by criteria that we refer to as "straightforward liquidity taking" and "competitive liquidity provision" below.

DEFINITION 1: *A WPBE is plausible if it satisfies* (*i*) *symmetry,* (*ii*) *stationarity,* (*iii*) *competitive liquidity provision,* (*iv*) *no unnecessary depth, and* (*v*) *straightforward liquidity taking (where these properties are defined below).*

We define *symmetry* to mean that the equilibrium is symmetric with respect to exchanges.

[25] Analogous statements could be made for the equilibrium selections corresponding to NDs and FBAs.

The model is symmetric in this way, and therefore it would seem that the focal equilibria should be symmetric as well.

Referring to the property that certain aspects of equilibrium behavior do not depend on calendar time, we formally define *stationarity* to mean the following two properties:

- Liquidity providers establish quotes at time zero and send no further orders unless another order has been processed by an exchange.
- The investor, if he is an information investor, conducts research with a fixed intensity r regardless of the time period in which he arrives.

We define *competitive liquidity provision* to mean that liquidity providers set quotes such that an ex-ante zero-profit condition holds order by order. Essentially the same assumption governs how quotes are priced in much of the rest of the rest of the literature (Glosten and Milgrom (1985), Kyle (1985), Glosten (1994)).

Referring to the property that liquidity providers quote only the minimal amount of depth required to satisfy liquidity investor demand, we formally define *no unnecessary depth* to mean the following two properties:

- Every initial quote established by a liquidity provider is accessed by a liquidity investor with positive probability on the equilibrium path.
- Liquidity providers, if at some point they believe that no future orders will originate from information investors, send cancellation orders for their remaining quotes.

Essentially, this criterion means that the presence of liquidity investors is necessary to induce liquidity providers to post quotes, which is consistent with the no-trade theorem logic of Milgrom and Stokey (1982).

Referring to the property that investors and snipers submit immediate-or-cancel orders only in accordance with their expected value for the security given their beliefs, we formally define *straightforward liquidity taking* to mean the following three properties:

- The investor, if he is a liquidity investor with a buying (selling) motive, submits to his home exchange an immediate-or-cancel order to buy (sell) a single share at a price β ($-\beta$), doing so in the period of arrival. Note that any other trading strategy would be weakly dominated for this trader type.
- The investor, if he is an information investor who learns that the value of the security is 1 (-1), submits immediate-or-cancel orders to buy (sell) at least one share at the price 1

(−1) to at least one exchange, doing so in the period of arrival. If he does not learn the value, then he does not send any orders.

- Each sniper, if and only if the expected value of the security given his beliefs is greater (less than) the ask (bid) at an exchange, submits immediate-or-cancel orders to buy (sell) at least one share to that exchange, where the price specified by the order is the expected value of the security given his beliefs.

PROPOSITION 1: *A WPBE is plausible if and only if it gives rise to a spread and research intensity as characterized by Proposition 1.*

Proof: Consider a WPBE and suppose that it is plausible. Let r^* denote the research intensity of information investors. (By stationarity, r^* is well defined.) Let T_1 denote the time period in which the first immediate-or-cancel order arrives at an exchange (if such an order arrives). Let s^* denote the initial spread set by liquidity providers. By stationarity, this spread is maintained until T_1, and by symmetry, it is the same for all exchanges. Most of the following arguments focus on the ask side of the LOB, but the bid side analysis is symmetric.

We first observe that all orders arriving at time T_1 must have been sent by the investor. To see this, it suffices to show that no sniper submits any order at any time before T_1. Applying Bayesian updating, the expected value of the security given HFT beliefs must be zero at all times before T_1. Moreover, by symmetry of bid and ask, the bid is nonpositive and the ask is nonnegative at all times before T_1. Combining these observations with straightforward liquidity taking for snipers, we obtain the desired conclusion.

Combining the above observation with the fact that a liquidity investor sends just a single order, we conclude that no liquidity investor order arrives after T_1. It follows from no unnecessary depth that liquidity providers send cancellations at time T_1 for any remaining quotes. In addition, by straightforward liquidity taking for liquidity investors, any liquidity investor orders that do arrive at T_1 are for exactly one share. It then also follows from no unnecessary depth that the initial depth quoted by liquidity providers must consist of one share on each side of the book for each exchange.

For the following arguments, we focus on the case in which $T_1 \in \{3\varepsilon, 4\varepsilon, 5\varepsilon, \ldots, 1\}$. Of course, it is also possible that $T_1 \in \{\varepsilon, 2\varepsilon\}$, but because this latter case arises with only infinitesimal probability, it can be ignored without loss. Fix any time $t \in \{3\varepsilon, 4\varepsilon, 5\varepsilon, \ldots, 1\}$. We consider various probabilities associated with orders arriving at a given exchange at such a time:

- We consider the probability that $T_1 = t$ and that a buy order sent by a liquidity investor arrives at a given exchange at that time. By straightforward liquidity taking for liquidity investors, the probability is $\frac{1}{2X(N+1)}(1-\lambda)$. In this event, the order is for exactly one share, and the expected value of the security is zero.
- We consider the probability that $T_1 = t$ and that a buy order sent by an information investor arrives at a given exchange at that time. By straightforward liquidity taking for information investors and symmetry, the probability is at least $\frac{1}{2X(N+1)}\lambda r^*$. In this event, the arriving order is for at least one share and the expected value of the security is one.

Suppose that at an immediate-or-cancel order to buy arrives at a given exchange at T_1. Applying Bayesian updating, the expected value of the security under HFT beliefs must be in the interval

$$\left[\frac{\lambda r^*}{1-\lambda+\lambda r^*}, 1\right].$$

Thus, by competitive liquidity provision, the initial quotes set by liquidity providers must be such that, at each exchange, exactly one share is quoted at the ask, where the ask is

$$\frac{s^*}{2} \in \left[\frac{\lambda r^*}{1-\lambda+\lambda r^*}, 1\right]. \tag{IA27}$$

Again, fix any time $t \in \{3\varepsilon, 4\varepsilon, 5\varepsilon, \ldots, 1\}$. Now consider instead various probabilities associated with orders arriving to the market:

- We consider the probability that $T_1 = t$, with a buy order sent by a liquidity investor arriving to the market at that time. By straightforward liquidity taking for liquidity investors, the probability is $\frac{1}{2(N+1)}(1-\lambda)$. In this event, the order is for exactly one share, exactly one such order arrives, and the expected value of the security is zero.
- We consider the probability that $T_1 = t$, with a buy order sent by an information investor arriving to the market at that time. By straightforward liquidity taking for information investors, the probability is at least $\frac{1}{2(N+1)}\lambda r^*$. In this event, each arriving order is for at least one share, multiple such orders may arrive simultaneously, and the expected value of the security is one.

Suppose that exactly one immediate-or-cancel order to buy arrives to the market at T_1. Applying Bayesian updating, the expected value of the security under HFT beliefs must be in the interval

$$\left[0, \frac{\lambda r^*}{1-\lambda+\lambda r^*}\right].$$

Comparing this to equation (IA27) and applying straightforward liquidity taking for snipers, we conclude that in this case, snipers do not send any orders at time T_1.

Suppose that two or more immediate-or-cancel orders to buy arrive to the market at T_1. Applying Bayesian updating, the expected value of the security under HFT beliefs must be one.[26] Comparing this to equation (IA27) and applying straightforward liquidity taking for snipers, we conclude that in this case, each sniper sends immediate-or-cancel orders at time T_1 to every exchange at which the initial quotes remain available.

Suppose the investor is an information investor who learns that the value of the security is $v = 1$. The investor earns a profit of $1 - s^*/2$ for every share that he buys, and so he optimally attempts to maximize his traded volume. Because one share is quoted at the ask of each exchange, the investor optimally selects a quantity of one for each of the orders that he sends. By symmetry, the only question is how many exchanges to target. Given the aforementioned behavior of liquidity providers and snipers, if he sends orders to y exchanges, then as in the proof of Proposition 1, he expects to receive the following number of fills:

$$F_{LOB}(y) = yp_I + y(1-p_I)^y + (1-p_H)(y-1)yp_I(1-p_I)^{y-1}.$$

By Lemma 1, this function is weakly increasing on the domain of positive integers, so it is indeed optimal to submit orders to target all exchanges. He then obtains $F(X) = X_I$ fills in expectation. And given the above, an information investor's expected profits conditional on a choice of research intensity r are

$$rX_I\left(1 - \frac{s^*}{2}\right) - c(r).$$

Thus, it is necessary for plausible equilibrium that

$$r^* \in \arg\max_{r\in[0,1]} \left\{ rX_I\left(1 - \frac{s^*}{2}\right) - c(r) \right\}. \tag{IA28}$$

Given the aforementioned behavior of the investor and snipers, then as in Section III.A of the main article, liquidity provider profits are zero only if

$$(1-\lambda)\frac{s^*}{2} = \lambda r^*(X_I + X_S)\left(1 - \frac{s^*}{2}\right).$$

[26] This does not formally follow from Bayesian updating if information investors always target only a single exchange. Nevertheless, if $X \geq 2$, then we can adapt some of the following arguments to rule out the possibility that information investors always target only a single exchange. Moreover, if $X = 1$, then HFT beliefs at such information sets are irrelevant.

Thus, it is also necessary for plausible equilibrium that

$$s^* = \frac{2\lambda r^*(X_I + X_S)}{1 - \lambda + \lambda r^*(X_I + X_S)}. \tag{IA29}$$

Together, conditions (IA28) and (IA29) imply the desired conclusion. □

Despite this result, it is worth clarifying that the equilibrium described in Section III.A of the main article is not the unique plausible equilibrium. For instance, another plausible equilibrium involves quotes on the various exchanges being maintained by different liquidity providers rather than by a single one.

We also conjecture that Proposition 1 could be strengthened. In particular, we conjecture that the proposition's conclusion would also apply to a weaker definition of plausible equilibrium that insists upon only competitive liquidity provision and straightforward liquidity taking.

IV. Strategic Exchanges

Earlier sections of this paper characterize equilibrium outcomes under several trading mechanisms. However, this raises the following question: if exchanges choose trading mechanisms for themselves, which of these mechanisms would they adopt? To address this question, we use the model to formulate a game among the exchanges in which they simultaneously choose trading mechanisms in a strategic fashion.[27] We use spread minimization as a proxy for profit maximization, and we assume that traders behave as described in the main text to determine the spreads that prevail under a profile of trading mechanisms. We find that it is a Nash equilibrium for all exchanges to use 1-ND, wherein all noncancellations are delayed by a random amount. However, the same is not true of the other mechanisms we consider.[28] In Internet Appendix IV.C, we attempt to reconcile this with the LOB's position as the prevailing industry standard.

A. *Exchange Game*

In this section, we define the "exchange game," in which exchanges choose trading mechanisms for themselves. To define this game, which we denote $\Gamma_{exchange}$, we specify players, strategy sets, and payoffs.

Players. The players of $\Gamma_{exchange}$ are the exchanges $\{1, \ldots, X\}$.

Strategy sets. A pure strategy in $\Gamma_{exchange}$ for an exchange is a trading mechanism that it selects for itself. For concreteness, we assume that the available mechanisms are limited to those already considered: LOB, qND for $q \in [0, 1]$, and FBAs, where the details of these mechanisms are as specified in the main text. Nevertheless, we conjecture that the results

[27] Of course, in practice exchanges must also make a number of other choices: the speed and trading infrastructure they provide, the fees they charge, etc. Addressing those issues, however, would require significant modifications to our existing model. For that reason, we focus on the choice of trading mechanism for the purpose of this extension.

[28] Budish, Lee, and Shim (2019) study a similar question, but both their findings and their approach differ from ours. In their model, the "exchange game" is a prisoner's dilemma: each exchange has a unilateral incentive to adopt a mechanism that eliminates sniping in response to public news (i.e., FBAs, yet equally NDs), but when all exchanges eliminate sniping they are no longer able to monetize co-location services, as they might if the LOB were maintained. Whereas we model the "exchange game" as a one-shot interaction, they model it as a repeated interaction and select the equilibrium in which the collusive outcome (i.e., LOBs) arises, maintained by Nash reversion. Thus, whereas our findings predict a trend toward NDs (and 1-ND in particular), they predict that—in lieu of any regulatory intervention—the LOB status quo will remain in place indefinitely.

stated here would continue to hold even if additional mechanisms were available. We use $m = (m_1, \ldots, m_X)$ to denote a profile of trading mechanisms.

Payoffs. Although our model does not provide a means to measure exchange profits, spread minimization seems to be a plausible proxy for profit maximization. According to an old industry saying, "liquidity begets liquidity." That is, a liquid exchange (in the model, an exchange with small spreads) attracts more investors and therefore becomes even more liquid.[29] Spread minimization (i.e., limiting adverse selection) can initiate this virtuous cycle and thereby attract investors, and it therefore seems likely to be a primary driver of an exchange's choice of mechanism.[30]

For any profile of trading mechanisms, we suppose that the investor and HFTs interact on the exchanges as in the baseline model. If exchanges all choose the same trading mechanism, then Propositions 1, 3, and 4 describe the resulting spreads.

If exchanges choose heterogeneous trading mechanisms, then additional work is needed to characterize the equilibrium of the continuation game so as to derive the resulting profile of spreads. As before, (*i*) snipers send immediate-or-cancel orders to each exchange after observing two trades, (*ii*) the liquidity provider sends cancellations to each exchange after observing one trade, (*iii*) the liquidity provider sets the spread so that she earns zero profits, (*iv*) each liquidity investor submits a single order to his home exchange in the period of his arrival, buying or selling in correspondence with the direction of his trading desire, and (*v*) each information investor chooses research intensity optimally and submits orders in the period of his arrival. However, there may be some subtleties in how information investors route orders. Before, when all exchanges used the same trading mechanism, an information investor could do no better than sending orders to all exchanges. Now, he may find it optimal to avoid an exchange if that exchange carries a relatively high risk of "tipping his hand" before he can obtain fills elsewhere.

For any profile of trading mechanisms $m = (m_1, \ldots, m_X)$, let $s(m) = (s_1(m), \ldots, s_X(m))$ denote the corresponding profile of spreads, which is pinned down by the play of the traders in the continuation game. We define the profile of payoffs in $\Gamma_{exchange}$ from m as $\pi(m) = -s(m)$.

[29] This force is not captured by the model of this paper because liquidity investors are quite inelastic: they divide equally among exchanges and, so long as the spread is below a certain level, trade only a fixed amount. For a model that does allow for some elasticity, see, for example Baldauf and Mollner (forthcoming).

[30] Another natural proxy for profit maximization is volume maximization. Different results would prevail if we were to define exchange payoffs using that proxy. In particular, the Pareto-dominant Nash equilibrium would be for all exchanges to use (synchronized) FBAs. However, since that result would rely quite heavily on the inelasticity of investor demand in the model, spread minimization is our preferred proxy.

B. *Results*

In this section, we analyze the exchange game $\Gamma_{exchange}$. We find that if exchanges are strategic in their choice of trading mechanism in the way described here, then the 1-ND equilibrium survives. Moreover, if certain conditions are satisfied, then the same is not true of the equilibria under the other mechanisms: LOB, qND for $q < 1$, and FBAs. Those equilibria fail to survive because exchanges, by deviating to 1-ND, can reduce their spread. One of these conditions is $c'(0) < 1$, which ensures that information investors always find it optimal to do some amount of research. Proposition 2 formalizes these statements.[31]

PROPOSITION 2: *The following statements are true.*

(i) It is a Nash equilibrium of $\Gamma_{exchange}$ *for all exchanges to use* 1*-ND.*

(ii) If $c'(0) < 1$ *and* $X \geq 2$*, then it is not a Nash equilibrium of* $\Gamma_{exchange}$ *for all exchanges to use qND for any* $q \in [0, 1)$*.*

(iii) If $c'(0) < 1$ *and* $X \geq 2$*, then it is not a Nash equilibrium of* $\Gamma_{exchange}$ *for all exchanges to use FBAs.*

(iv) If $c'(0) < 1$ *and* $X \geq 2$*, then is not a Nash equilibrium of* $\Gamma_{exchange}$ *for all exchanges to use the LOB.*

Proof of Proposition 2: *Part (i).* To prove that it is a Nash equilibrium for all exchanges to use 1-ND, it suffices to show that no single exchange has a profitable deviation. When all exchanges use 1-ND, an information investor is able to obtain a fill at only one exchange, no matter how he routes orders, snipers do not trade, and the equilibrium spread is s^*_{1ND}, as characterized by Proposition 3.

If one exchange deviates to any other trading mechanism, then the information investor would again be able to obtain a fill at only one exchange, no matter how he routes orders. Moreover, all exchanges would be alike in that none would feature adverse selection exerted by snipers.[32] Therefore, the equilibrium spread would again be s^*_{1ND}. Because this deviation would not affect the spreads, it is not profitable.

Parts (ii) to (iv). To establish these three claims, it suffices to show that 1-ND would be a profitable deviation in each scenario. Part of the argument is common to each of the three

[31] We conjecture that the following are also true. 1-ND is a dominant strategy in $\Gamma_{exchange}$ for each exchange. If $c'(0) < 1$ and $X \geq 2$, then 1-ND is the unique dominant strategy in $\Gamma_{exchange}$ for each exchange.

[32] Snipers submit orders only after two trades have taken place. Because the information investor can obtain only a single fill before the liquidity provider cancels all remaining quotes, snipers submit no orders on path. Thus, although the LOB permits snipers to exert adverse selection in general, this does not occur in the scenario in which one exchange deviates to the LOB while all other exchanges continue to use 1-ND.

claims, and we establish it here. We then establish the remaining parts of each argument separately below.

Suppose it is the case that (i) all exchanges use qND for $q < 1$, in which case the equilibrium spread is s^*_{qND}, as characterized by Proposition 3, (ii) all exchanges use FBAs, in which case the equilibrium spread is s^*_{FBA}, as characterized by Proposition 4, or (iii) all exchanges use the LOB, in which case the equilibrium spread is s^*_{LOB}, as characterized by Proposition 1. Similar to the proof of Corollary 2, the assumption that $c'(0) < 1$ implies that equilibrium research intensity is positive, which implies that the equilibrium spread is positive as well.

If one exchange deviates to 1-ND, then when an information investor routes orders, he must choose between (i) obtaining a fill only on the 1-ND exchange, and (ii) obtaining fills only on (some of) the other exchanges. There are then two cases. In the first case, the information investor never routes to the 1-ND exchange. In that case, the spread on that exchange is zero. Because the spread was positive before, the deviation is indeed profitable.

It therefore only remains to analyze the second case, in which the information investor sometimes routes to the 1-ND exchange. We let $\gamma^* \in (0, 1]$ denote the endogenously chosen probability that the information investor routes to the 1-ND exchange, so that with probability $1-\gamma^*$ he routes to all the other exchanges. We also let x denote the deviating exchange. Below, we characterize the resulting profile of spreads for each scenario, and we argue that in each, the deviation is indeed profitable for exchange x.

Part (ii). Define

$$X'_{qND} = q^{X-1} + \sum_{x=1}^{X-1} \binom{X-1}{x} (1-q)^x q^{X-1-x} \left(x p_H (1-p_I)^x + x \left[1 - p_H (1-p_I) \right] \right),$$

where X'_{qND} is the expected number of fills obtained by an information investor, conditional on learning the value of the security and routing orders to the $X-1$ exchanges still using qND. The profile of spreads in this scenario must satisfy

$$\begin{aligned}
s^*_x &= \frac{2\lambda\gamma^* r^* X}{1 - \lambda + \lambda\gamma^* r^* X} \\
s^*_{-x} &= \frac{2\lambda(1-\gamma^*) r^* \frac{X}{X-1} X'_{qND}}{1 - \lambda + \lambda(1-\gamma^*) r^* \frac{X}{X-1} X'_{qND}} \\
r^* &\in \arg\max_{r \in [0,1]} \left\{ r \left(1 - \frac{s^*_x}{2} \right) - c(r) \right\} \\
1 - \frac{s^*_x}{2} &\leq X'_{qND} \left(1 - \frac{s^*_{-x}}{2} \right).
\end{aligned}$$

From the fourth equation, which is necessary given that the information investor sometimes routes to exchange x, we obtain $\gamma^* \leq 1/X$.[33]

Next, define $s^*(\Omega)$ and $r^*(\Omega)$ as the solution to the system

$$s^* = \frac{2\lambda r^*[\Omega X_{qND} + (1-\Omega)]}{1-\lambda+\lambda r^*[\Omega X_{qND} + (1-\Omega)]} \tag{IA30}$$

$$r^* \in \arg\max_{r\in[0,1]} \left\{ r[\Omega X_{qND} + (1-\Omega)]\left(1 - \frac{s^*}{2}\right) - c(r) \right\}. \tag{IA31}$$

Since $c'(0) < 1$ rules out the possibility of a corner solution with no research, we must have $r^*(\Omega) > 0$. On that domain, s is, other things equal, strictly increasing in Ω (because $X_{qND} > 1$ when $q < 1$) and weakly increasing in r. Moreover, applying Topkis' Theorem to (IA31), we find that, other things equal, r is weakly increasing in Ω and weakly decreasing in s. By combining these observations, we conclude that $s^*(\Omega)$ is strictly increasing in Ω.

Notice that s^*_{qND} corresponds to $s^*(\Omega)$ evaluated at $\Omega = 1$. In addition, since $\gamma^* \leq 1/X$, we have that s^*_x is weakly less than $s^*(\Omega)$ evaluated at $\Omega = 0$. We therefore have $s^*_x < s^*_{qND}$, so the deviation is profitable in this case as well.

Part (iii). The argument is similar to that used in Part *(ii)* of the proof. The profile of spreads in this scenario must satisfy

$$\begin{aligned}
s^*_x &= \frac{2\lambda\gamma^* r^* X}{1-\lambda+\lambda\gamma^* r^* X} \\
s^*_{-x} &= \frac{2\lambda(1-\gamma^*) r^* X}{1-\lambda+\lambda(1-\gamma^*) r^* X} \\
r^* &\in \arg\max_{r\in[0,1]} \left\{ r\left(1 - \frac{s^*_x}{2}\right) - c(r) \right\} \\
1 - \frac{s^*_x}{2} &\leq (X-1)\left(1 - \frac{s^*_{-x}}{2}\right).
\end{aligned}$$

From the fourth equation, which is necessary given that the information investor sometimes routes to exchange x, we obtain $\gamma^* \leq 1/X$.[34]

[33] To see that $\gamma^* \leq 1/X$, observe that if $\gamma^* > 1/X$, then the left-hand side would be less than $\frac{1-\lambda}{1-\lambda+\lambda r}$, while the right-hand side would be greater than $\frac{(1-\lambda)X'_{qND}}{1-\lambda+\lambda r X'_{qND}}$, which is a contradiction because $X \geq 2$ implies that $X'_{qND} \geq 1$.

[34] To see that $\gamma^* \leq 1/X$, observe that if $\gamma^* > 1/X$, then the left-hand side would be less than $\frac{1-\lambda}{1-\lambda+\lambda r}$,

Next, define $s^*(\Omega)$ and $r^*(\Omega)$ as the solution to the system

$$s^* = \frac{2\lambda r^*[\Omega X + (1-\Omega)]}{1-\lambda+\lambda r^*[\Omega X + (1-\Omega)]} \tag{IA32}$$

$$r^* \in \arg\max_{r\in[0,1]} \left\{ r[\Omega X + (1-\Omega)]\left(1 - \frac{s^*}{2}\right) - c(r) \right\}. \tag{IA33}$$

Since $c'(0) < 1$ rules out the possibility of a corner solution with no research, we must have $r^*(\Omega) > 0$. On that domain, s is, other things equal, strictly increasing in Ω (because $X \geq 2$) and weakly increasing in r. Moreover, applying Topkis' Theorem to (IA33), we find that, other things equal, r is weakly increasing in Ω and weakly decreasing in s. By combining these observations, we conclude that $s^*(\Omega)$ is strictly increasing in Ω.

Notice that s^*_{FBA} corresponds to $s^*(\Omega)$ evaluated at $\Omega = 1$. In addition, since $\gamma^* \leq 1/X$, we have that s^*_x is weakly less than $s^*(\Omega)$ evaluated at $\Omega = 0$. We therefore have $s^*_x < s^*_{FBA}$, so the deviation is profitable in this case as well.

Part (iv). Define

$$X'_I = (X-1)\left(p_I + (1-p_I)^{X-1} + (1-p_H)(X-2)p_I(1-p_I)^{X-2}\right)$$
$$X'_S = (X-1)\left(1 - p_I - (1-p_I)^{X-1} - (1-p_H)(X-2)p_I(1-p_I)^{X-2} - p_H(X-2)p_I(1-p_I)^{X-2}\right),$$

where X'_I is the expected number of fills obtained by an information investor, conditional on learning the value of the security and routing orders to the $X-1$ exchanges still using the LOB. Likewise, X'_S is the expected number of fills obtained by snipers, conditional on an information investor arriving, learning the value of the security, and routing orders to the $X-1$ exchanges still using the LOB. The argument is similar to that used in Part *(ii)* of the

while the right-hand side would be greater than $\frac{(1-\lambda)(X-1)}{1-\lambda+\lambda r(X-1)}$, which is a contradiction because $X \geq 2$.

proof. The profile of spreads in this scenario must satisfy

$$
\begin{aligned}
s_x^* &= \frac{2\lambda\gamma^* r^* X}{1-\lambda+\lambda\gamma^* r^* X} \\
s_{-x}^* &= \frac{2\lambda(1-\gamma^*) r^* \frac{X}{X-1}(X_I'+X_S')}{1-\lambda+\lambda(1-\gamma^*) r^* \frac{X}{X-1}(X_I'+X_S')} \\
r^* &\in \arg\max_{r\in[0,1]} \left\{ r X_I' \left(1-\frac{s_{-x}^*}{2}\right) - c(r) \right\} \\
1-\frac{s_x^*}{2} &= X_I' \left(1-\frac{s_{-x}^*}{2}\right).
\end{aligned}
$$

From the fourth equation, which is necessary given that the information investor sometimes routes to exchange x, we obtain $s_x^* \leq s_{-x}^*$.

Next, define $s^*(\Omega)$ and $r^*(\Omega)$ as the solution to the system

$$
s^* = \frac{2\lambda r^* \left[\Omega(X_I+X_S) + (1-\Omega)\frac{X}{X-1}(X_I'+X_S')\right]}{1-\lambda+\lambda r^* X \left[\Omega(X_I+X_S) + (1-\Omega)\frac{X}{X-1}(X_I'+X_S')\right]} \tag{IA34}
$$

$$
r^* \in \arg\max_{r\in[0,1]} \left\{ r\left[\Omega X_I + (1-\Omega)X_I'\right]\left(1-\frac{s^*}{2}\right) - c(r) \right\}. \tag{IA35}
$$

Since $c'(0) < 1$ rules out the possibility of a corner solution with no research, we must have $r^*(\Omega) > 0$. On that domain, s is, other things equal, strictly increasing in Ω and weakly increasing in r.[35] Moreover, applying Topkis' Theorem to (IA35), we find that, other things equal, r is weakly increasing in Ω and weakly decreasing in s.[36] By combining these observations, we conclude that $s^*(\Omega)$ is strictly increasing in Ω.

Notice that s_{LOB}^* corresponds to $s^*(\Omega)$ evaluated at $\Omega = 1$. In addition, s_{-x}^* is weakly less than $s^*(\Omega)$ evaluated at $\Omega = 0$. We therefore have $s_{-x}^* < s_{LOB}^*$. In addition, as argued above, $s_x^* \leq s_{-x}^*$. We conclude that $s_x^* < s_{LOB}^*$, so the deviation is profitable in this case as well. □

These results are driven primarily by the observation that 1-ND is extremely effective at limiting adverse selection, for two reasons. First, it eliminates the adverse selection that comes from sniper orders. Second, it puts an information investor in a difficult situation when it comes to order routing. Because of the large variability in execution time under 1-

[35] To show that s is strictly increasing in Ω, it suffices to verify that $X_I + X_S > \frac{X}{X-1}(X_I'+X_S')$. After plugging in for these expressions, we see that it suffices to verify that $(X-1)(1-p_I)^{X-1}$ is strictly decreasing in X on the domain where $X \geq 2$ and $p_I \geq 0.5$, which is indeed the case.

[36] To show that r is weakly increasing in Ω, it suffices to verify that $X_I \geq X_I'$. This is a consequence of Lemma 1.

ND, it is only with infinitesimal probability that he can achieve an execution at both a 1-ND exchange and any other exchange. This reduces the number of fills that he can obtain on 1-ND exchanges and, in addition, makes those exchanges relatively less attractive to him. In consequence, 1-ND also tends to reduces the adverse selection that comes from information investor orders.

The first part of the proposition is proven by showing that if all exchanges are using 1-ND, then an exchange cannot further reduce its spread by deviating to another mechanism. The other parts are proven by showing that if all exchanges are using another mechanism, then an exchange can reduce its spread by deviating to 1-ND.

For the negative results, we must assume there are $X \geq 2$ exchanges, for otherwise all mechanisms would give rise to the same equilibrium spread. Similarly, the role of assuming $c'(0) < 1$ is to eliminate the possibility of zero research intensity in equilibrium, in which case there would be no spread and therefore no room for a profitable deviation.

C. Discussion

Consistent with the result of this section, there does seem to be some momentum building for mechanisms resembling 1-ND. Indeed, several industry participants have already proposed modifications to their order matching rules that would incorporate random delays of non-cancellations or similar categories of orders. (See Internet Appendix VI.A for details.)

Nevertheless, the LOB remains the standard industry practice. While this would seem to be in conflict with our result, the discrepancy might be explained by the fact that the industry has been recently transformed by the rise of HFT and a drastic increase in fragmentation. Indeed, because all mechanisms lead to the same outcome in the absence of fragmentation (i.e., if $X = 1$), one might conclude that the LOB became a disequilibrium choice only recently, and exchanges may not yet have had sufficient time to adjust to the new trading conditions. Moreover, a number of forces—regulatory barriers, switching costs, resistance from clientele—may be hindering adjustments in the trading mechanism. As a result, we cautiously interpret Proposition 2 as a statement about what to expect in the long run.

This result also has policy implications. If the financial industry will ultimately settle at an equilibrium in which all exchanges adopt 1-ND, then a regulator might wish to speed the transition to that equilibrium by endorsing that mechanism. This would have the benefit of reducing the amount of time spent at the inefficient LOB equilibrium, which is off the frontier of the tradeoff between liquidity and information production.

V. Benefits of Informative Prices

This appendix discusses rationales for the social value of informative prices—and in turn fundamental research. In the model, traders are indifferent to the timing of resolution of uncertainty. However, to the extent that in practice traders prefer earlier resolution of uncertainty, they would benefit from more informative prices.

Moreover, informative prices may also be a positive externality for economic agents not trading the security. By conveying information to real-world decision-makers, more informative prices can improve the efficiency of resource allocation in the wider economy. The literature has identified several channels through which this may occur.

The idea that informative prices are vital to the efficient distribution of resources dates back to at least Hayek (1945). This is especially true in the case of "equity-dependent" firms (i.e., firms able to raise funds only through equity issuance). Such a firm may be discouraged from undertaking an efficient investment in the event that its stock price—and thus its cost of capital—falls far below its fundamental value, which is less likely if prices are more informative, for two reasons. First, with less information, prices are less closely tied to the fundamental value, which raises the probability of a large difference between the two. Second, with less information, prices may be depressed overall, as investors demand a higher return to hold the security, as in Easley and O'Hara (2004). Myers and Majluf (1984) provide a formal model in which imperfect pricing can lead a firm to bypass an efficient project. Moreover, Baker, Stein, and Wurgler (2003) and Chen, Goldstein, and Jiang (2007) find empirical evidence for such effects. In addition, because lenders use stock prices in setting the terms at which they will lend to a firm (e.g., Merton (1974)), this force might not be limited to equity-dependent firms, but rather might apply more generally.

Next, the *incentive channel* refers to the idea that more informative prices assist a board of directors in gauging a manager's performance, thereby enabling them to provide better incentives for the manager and in turn raising the manager's effort. This idea was first proposed by Baumol (1965), and it was later formalized in models by Diamond and Verrecchia (1982), Fishman and Hagerty (1989); and Holmström and Tirole (1993). Furthermore, Kang and Liu (2008) and Ferreira, Ferreira, and Raposo (2011) find empirical evidence consistent with the predictions of these models.

Baumol (1965) also proposed the *learning channel*, whereby more informative prices provide better feedback to firm managers, thus enabling them to make better decisions. Dow and Gorton (1997), Subrahmanyam and Titman (1999) and Lin, Liu, and Sun (2019)

provide theoretical models of this channel. Luo (2005), Chen, Goldstein, and Jiang (2007), Kau, Linck, and Rubin (2008), Bakke and Whited (2010), Foucault and Fresard (2014), and Lin, Liu, and Sun (2019) find empirical evidence consistent with the operation of this channel.

There may also exist other channels through which informative prices increase economic efficiency beyond those specifically discussed above. In particular, the information contained in prices may be used by other real-world decision-makers, including employees, customers, suppliers, regulators, and blockholders, all of whom take actions that may influence the efficiency of resource allocation.[37] See Bond, Edmans, and Goldstein (2012) for a thorough review of the literature on the effects of financial markets on the real economy.

[37]For example, Subrahmanyam and Titman (2001) develop a model of some of these dependencies, and Faure-Grimaud and Gromb (2004) argue that more informative prices increase the incentives of blockholders to take actions that increase the value of the company. Additionally, several papers document a relationship between informative prices and economic efficiency without identifying a particular channel. Examples include Wurgler (2000) and Durnev, Morck, and Yeung (2004).

VI. Alternative Mechanisms in Practice

This appendix reports on real-world examples of trading mechanisms that resemble either NDs or FBAs. None of the mechanisms used in practice match exactly the theoretical proposals that we analyze in this paper, but not all those differences are of economic import. Generally speaking, departures from our theoretical analysis tend to be cosmetic in nature in the case of NDs but economically meaningful in the case of auctions.

A. Noncancellation Delay Mechanisms in Practice

Mechanisms related to ND have been recently implemented in practice by two Canadian exchanges. (In Table IA.III we summarize these and other ND implementations and proposals.) Aequitas NEO Exchange, a new Canadian equities exchange that opened in March 2015, applies a random delay of 3 to 9 milliseconds to immediate-or-cancel orders from traders whom they have classified as "latency sensitive traders" (Aequitas NEO Exchange (2016)). Six months later, the incumbent, TMX Group, followed suit on one of its platforms, the TSX Alpha Exchange. That exchange applies a random delay of 1 to 3 milliseconds to all orders except post-only orders and cancellations thereof (Alpha Trading Systems Limited Partnership (2016)).[38] In Europe, the Eurex exchange has recently announced that they will begin a pilot study in which a deterministic delay of 1 millisecond (for German equity options) or 3 milliseconds (for French equity options) will be applied to all liquidity-removing orders (Deutsche Börse Group (2019)). Similarly, the London Metal Exchange has announced plans to implement a deterministic delay of 8 milliseconds for all orders except for cancellations on their LMEprecious platform (London Metal Exchange (2019)). The Moscow Exchange launched an experimental order book for the USD/RUB currency pair, which features a randomized delay of between 2 and 5 milliseconds for non-cancellation orders (Moscow Exchange (2019)).

In the U.S., Intercontinental Exchange recently submitted a proposal to implement a delay of 3 milliseconds to liquidity-taking orders for gold and silver daily futures (ICE Futures U.S. (2019)). Before its acquisition by NYSE, the Chicago Stock Exchange submitted a proposal to implement a delay of 350 microseconds to market orders, marketable limit orders, and certain related cancel messages (Chicago Stock Exchange (2016, 2017)). More recently, EDGA has proposed a delay of four milliseconds for all liquidity-removing orders (Cboe

[38] See Chen et al. (2017) and Anderson et al. (2018) for empirical analyses of the TSX Alpha delay.

EDGA Exchange (2019)). In addition, NASDAQ PHLX once proposed to implement a delay of 5 milliseconds to marketable orders (NASDAQ OMX PHLX (2012)), although they subsequently withdrew that proposal. Additionally, Interactive Brokers advocated in an open letter that the SEC mandate a random delay of 10 to 200 milliseconds for orders that would remove liquidity (Peterffy (2014)).

These implementations and proposals differ both from each other and from our proposal in terms of the exact orders that are delayed. In particular, while we propose to delay all noncancellations, most of these proposals exempt nonmarketable limit orders from the delay. While this has the advantage of subjecting fewer orders to delay, it also complicates the mechanism, because the decision of whether to delay may hinge on the current state of the LOB, rather than simply on the order type. Nevertheless, exempting non-marketable limit orders from the delay would not alter equilibrium outcomes within the model.

Also similar in spirit to NDs is the 350-microsecond delay that IEX applies to all incoming orders (IEX Group (2018a)). A popular class of order types provided by IEX are peg orders, which rest in the IEX LOB at prices defined in reference to the national best bid and offer (NBBO). Because the NBBO is not subject to the delay but instead updated in real time, those orders are protected from snipers. A difference is that IEX's design only protects peg orders in this way, whereas NDs protect standard limit orders as well. IEX's design therefore has drawbacks: (*i*) fewer orders are protected from aggressive-side order anticipation, and (*ii*) an incentive is created for traders to switch from standard limit orders to peg orders, which could disrupt the price discovery process. IEX initially operated as an alternative trading system, but in 2016 it became an exchange, the primary effect of which was that its quotes then received trade-through protection.[39] Similarly, the NYSE American exchange subsequently adopted virtually the same mechanism (NYSE (2018)).

Also similar to ND is the "ideal latency floor" mechanism (Melton (2015)), which was adopted in 2016 by Thomson Reuters Matching, a foreign exchange venue. The mechanism operates by batching noncancellation orders, and then releasing them to the LOB in a randomized sequence.

NDs are also similar in spirit to "last look," a common practice in several other foreign exchange markets, whereby dealer platforms have the ability to retroactively cancel trades within a short window of time. Last look has recently been subject to criticism. One argument against the practice is that it enables information leakage: a market maker can observe its clients' intentions, acquiring any information therein, even without filling their

[39] See Hu (2019) for an empirical analysis of when IEX became an exchange.

orders (Foreign Exchange Professionals Association (2015)). NDs, in contrast, are immune to this criticism. With NDs, a trader observes the orders of its counterparties only if a trade occurs.

B. Auction-Based Trading Mechanisms in Practice

Mechanisms related to the FBA design of Section V.B of the main article are currently in use by a small number of venues. In what follows we describe such implementations and contrast them with the proposal of Budish, Cramton, and Shim (2015) ("the proposal"), which also guides our analysis in the text. To our knowledge, the implementations closest to the proposal are:

(*i*) the Cboe Europe *Equities Periodic Auctions Book* (Cboe Global Markets (2015)),
(*ii*) the London Stock Exchange *Turquoise Plato Lit Auctions* (London Stock Exchange (2019)),
(*iii*) the Goldman Sachs *SIGMA X Auction Book* (Goldman Sachs (2018)),
(*iv*) the ITG *POSIT Auction* (Investment Technology Group (2017)), and
(*v*) the Frankfurt Stock Exchange *Continuous Auction with Specialist* (Deutsche Börse Group (2017)).

These implementations differ from the proposal in certain specifics. These auctions do not occur at fixed points in time but rather, variously, at randomized intervals or endogenously-chosen times. And in many cases, the auctions are not sealed-bid in that they make available indicative price and quantity information prior to clearing. Nevertheless, these implementations resemble the proposal in that they replace LOBs and that auctions can happen repeatedly over an extended window of time. See Besson, Lasnier, and Falck (2019) for an empirical analysis of the effects of some of these auction mechanisms.

In addition, some venues have replaced continuous trading with auctions for illiquid securities, including Euronext (2018), SETSqx at the London Stock Exchange (2015), and the Tel Aviv Stock Exchange (2018). Several other exchanges have implemented auction mechanisms to complement continuous trading, for example, to determine prices at the open (in the morning or after a trading halt) and at the close. However, these implementations depart considerably from the proposal: (*i*) they are not frequent, crossing only a handful of times per day, and (*ii*) they are not sealed-bid, publishing indicative information about price or volumes before clearing.

Table IA.III
Mechanisms Related to ND Proposed or Implemented in Practice

Venue	Length of Delay	Targets of Delay	Effective Date
Implementations			
Aequitas NEO Exchange	3–9 milliseconds	Immediate-or-cancel orders from IDs classified as "latency sensitive traders"	March 27, 2015 – present
TSX Alpha Exchange	1–3 milliseconds	All but post-only orders and cancellations thereof	September 21, 2015 – present
Eurex Exchange	1 or 3 milliseconds	Liquidity-removing orders for German and French equity options	June 3, 2019 – present
IEX	350 microseconds	All orders (but not the NBBO)	October 25, 2013 – present
Moscow Exchange	2–5 milliseconds	All but cancellation orders for USD/RUB currency pair	April 22, 2019 – present
NYSE American	350 microseconds	All orders (but not the NBBO)	July 24, 2017 – present
Thomson Reuters	0–3 milliseconds	All but cancellation orders	June 5, 2016 – present
Proposals			
Cboe EDGA Exchange	4 milliseconds	Liquidity-removing orders	TBD
Chicago Stock Exchange	350 microseconds	Market orders, marketable limit orders, and certain related cancel messages	N/A
ICE Futures U.S.	3 milliseconds	Liquidity-removing orders	TBD
London Metal Exchange	8 milliseconds	All but cancellation orders for LMEprecious	TBD
NASDAQ OMX PHLX	5 milliseconds	Marketable orders	N/A
Interactive Brokers	10–200 milliseconds	Liquidity-removing orders	N/A

VII. Alternative Welfare Criteria

In Section IV of the main article, we characterize the set of feasible outcomes with respect to two criteria: (*i*) research intensity, which might be thought of as a sufficient statistic for the positive externalities of financial markets and therefore the welfare of unmodeled agents, and (*ii*) liquidity investor welfare. Although we intend liquidity investor welfare to be interpreted as a measure for liquidity—indeed, for mechanisms with a spread, it is related to the spread through $w = \beta - s/2$—it obviously can also be interpreted as the welfare of a single type of modeled agent: liquidity investors. In this appendix, we demonstrate that our main findings carry over even if we modify this second criterion so as to account for the welfare of other types of modeled agents as well.

In Internet Appendix VII.A we depart from the previous analysis by taking total investor welfare (instead of liquidity investor welfare) for the second criterion. Thus, this analysis also values rents earned from information acquisition. Nevertheless, we obtain analogues of all key results. Likewise, in Internet Appendix VII.B, we take total trader welfare for the second criterion. Thus, this analysis also values not only rents earned from information acquisition but also any profits that HFTs may accrue. Provided that a certain subtlety is addressed in an appropriate way, we again obtain analogues of all key results.

A. Total Investor Welfare

In this appendix, we conduct analysis similar to that of Section IV of the main article, characterizing the set of feasible outcomes, but with respect to a different set of criteria, research intensity and total investor welfare.

As before, the social planner recommends a research intensity to the investor and allocates resources (dollars and shares of the security) among the traders, subject to the same set of constraints. The *feasible set*, which we denote $\mathcal{F}'$, consists of the set of research intensities, r, and investor welfares, w, that can be implemented in this way. Using the same notation as in Section IV of the main article, the feasible set is defined as follows. The key difference is in the first constraint, where the difference is due to using w to denote expected investor welfare instead of liquidity investor welfare:

$$\mathcal{F}' = \left\{ (r, w) \;\middle|\; \begin{array}{l} \exists y(\theta), \exists z(\theta), \exists \{y_h(\theta)\}_{h \in \mathcal{H}}, \exists \{z_h(\theta)\}_{h \in \mathcal{H}} \text{ such that} \\ \text{(W}'\text{), (BB–1), (BB–2), (IR–I), (IR–H), (O)} \end{array} \right\},$$

where

$$
\begin{aligned}
&\text{(W}'\text{)} \quad w = \mathbb{E}_r[u(y(\theta), z(\theta)|\theta)] - \lambda c(r) \\
&\text{(BB–1)} \quad (\forall \theta \in \Theta) : y(\theta) + \textstyle\sum_{h\in\mathcal{H}} y_h(\theta) = 0 \\
&\text{(BB–2)} \quad (\forall \theta \in \Theta) : z(\theta) + \textstyle\sum_{h\in\mathcal{H}} z_h(\theta) = 0 \\
&\text{(IR–I)} \quad (\forall \theta \in \Theta) : u(y(\theta), z(\theta)|\theta) \geq 0 \\
&\text{(IR–H)} \quad (\forall h \in \mathcal{H}) : \mathbb{E}_r[u_h(y_h(\theta), z_h(\theta)|\theta)] \geq 0 \\
&\text{(O)} \quad r \in \arg\max_{\hat{r}\in[0,1]} \left[\tfrac{\hat{r}}{2} u(y(1), z(1)|1) + \tfrac{\hat{r}}{2} u(y(-1), z(-1)|-1) + (1-\hat{r}) u(y(0), z(0)|0) - c(\hat{r})\right]
\end{aligned}
$$

PROPOSITION 3: *The feasible set is* $\mathcal{F}' = \{(r, w) \mid r \in [0,1], w \in [\lambda r c'(r) - \lambda c(r), (1-\lambda)\beta - \lambda c(r)]\}$.

Proof of Proposition 3: As in the proof of Proposition 2, an initial observation is that the combination of (BB–1), (BB–2), and (IR–H) is equivalent to the following single constraint:

$$
\frac{1-\lambda}{2} y(B) + \frac{1-\lambda}{2} y(S) + \frac{\lambda r}{2} [y(1) + z(1)] + \frac{\lambda r}{2} [y(-1) - z(-1)] + \lambda(1-r) y(0) \leq 0. \tag{BB}
$$

The remainder of the proof consists of two parts. First, we show that the set defined in the proposition constitutes an outer bound for $\mathcal{F}'$. Second, we show that it constitutes an inner bound for $\mathcal{F}'$.

Part One (Outer Bound). Rewriting (W′), we obtain

$$
\begin{aligned}
w &= \frac{1-\lambda}{2} [y(B) + \beta \mathbb{1}\{z(B) = 1\}] + \frac{1-\lambda}{2} [y(S) + \beta \mathbb{1}\{z(S) = -1\}] \\
&\quad + \frac{\lambda r}{2} [y(1) + z(1)] + \frac{\lambda r}{2} [y(-1) - z(-1)] + \lambda(1-r) y(0) - \lambda c(r) \\
&\leq (1-\lambda)\beta + \frac{1-\lambda}{2} y(B) + \frac{1-\lambda}{2} y(S) \\
&\quad + \frac{\lambda r}{2} [y(1) + z(1)] + \frac{\lambda r}{2} [y(-1) - z(-1)] + \lambda(1-r) y(0) - \lambda c(r).
\end{aligned}
$$

Applying (BB), we obtain

$$
w \leq (1-\lambda)\beta - \lambda c(r). \tag{IA36}
$$

This establishes the desired upper bound on w. To establish the corresponding lower bound, we begin by rewriting (O) as

$$
r \in \arg\max_{\hat{r}\in[0,1]} \left\{ \frac{\hat{r}}{2} [y(1) + z(1)] + \frac{\hat{r}}{2} [y(-1) - z(-1)] + (1-\hat{r}) y(0) - c(\hat{r}) \right\},
$$

which implies[40]

$$r\left(\frac{y(1)+z(1)}{2}+\frac{y(-1)-z(-1)}{2}-y(0)\right)\geq rc'(r). \tag{IA37}$$

Next, we again rewrite (W′)

$$\begin{aligned}w &= \frac{1-\lambda}{2}\left[y(B)+\beta\mathbb{1}\{z(B)=1\}\right]+\frac{1-\lambda}{2}\left[y(S)+\beta\mathbb{1}\{z(S)=-1\}\right]\\ &\quad+\frac{\lambda r}{2}\left[y(1)+z(1)\right]+\frac{\lambda r}{2}\left[y(-1)-z(-1)\right]+\lambda(1-r)y(0)-\lambda c(r)\\ &= \frac{1-\lambda}{2}\left[y(B)+\beta\mathbb{1}\{z(B)=1\}\right]+\frac{1-\lambda}{2}\left[y(S)+\beta\mathbb{1}\{z(S)=-1\}\right]\\ &\quad+\lambda y(0)+\lambda r\left(\frac{y(1)+z(1)}{2}+\frac{y(-1)-z(-1)}{2}-y(0)\right)-\lambda c(r).\end{aligned}$$

Applying (IR–I) for $\theta\in\{B,S,0\}$, we obtain

$$w\geq\lambda r\left(\frac{y(1)+z(1)}{2}+\frac{y(-1)-z(-1)}{2}-y(0)\right)-\lambda c(r).$$

Using (IA37), this becomes

$$w\geq\lambda rc'(r)-\lambda c(r). \tag{IA38}$$

Thus, (IA36) and (IA38) imply $\mathcal{F}'\subset\{(r,w)\mid w\in[\lambda rc'(r)-\lambda c(r),(1-\lambda)\beta-\lambda c(r)]\}$, as desired.

Part Two (Inner Bound). For this part of the proof, we argue that any element of the set defined in the proposition can be implemented by a contract satisfying all of the constraints of $\mathcal{F}'$. In what follows, we use $r_{\max}$ to denote the largest r such that there exists a w for which $(r,w)\in\mathcal{F}'$. It is defined implicitly by

$$(1-\lambda)\beta=\lambda r_{\max}c'(r_{\max}).$$

[40] Define $\Delta=\left[\frac{y(1)+z(1)}{2}+\frac{y(-1)-z(-1)}{2}-y(0)\right]$. By assumption, $c(\cdot)$ is C^1. Any solution to the maximization problem in (O) must therefore satisfy one of the three conditions (*i*) $r=0$ and $c'(0)\geq\Delta$, (*ii*) $c'(r)=\Delta$, or (*iii*) $r=1$ and $c'(1)\leq\Delta$. In any of these three cases, the claimed inequality holds.

Suppose $r \in [0, r_{\max}]$ and suppose $w \in [\lambda r c'(r) - \lambda c(r), (1-\lambda)\beta - \lambda c(r)]$. Let

$$\begin{aligned} y(B) = y(S) &= \frac{w + \lambda c(r) - r c'(r)}{1-\lambda} - \beta \\ z(B) = -z(S) &= 1 \\ y(1) = -y(-1) &= 0 \\ z(1) = -z(-1) &= c'(r) \\ y(0) &= 0 \\ z(0) &= 0. \end{aligned}$$

We now argue that these contracts satisfy the constraints (W′), (BB), (IR–I), and (O):

(W′) Plugging in, $\mathbb{E}_r[u(y(\theta), z(\theta)|\theta)] - c(r) = (1-\lambda)\left(\frac{w + \lambda c(r) - \lambda r c'(r)}{1-\lambda} - \beta + \beta\right) + \lambda r c'(r) - \lambda c(r) = w$.

(BB) Plugging in, $\frac{1-\lambda}{2} y(B) + \frac{1-\lambda}{2} y(S) + \frac{\lambda r}{2}[y(1) + z(1)] + \frac{\lambda r}{2}[y(-1) - z(-1)] + \lambda(1 - r) y(0) = w + \lambda c(r) - (1-\lambda)\beta$, which is nonpositive by assumption.

(IR–I) First, $u(y(B), z(B)|B) = u(y(S), z(S)|S) = \frac{w + \lambda c(r) - \lambda r c'(r)}{1-\lambda}$, which is nonnegative by assumption. Second, $u(y(1), z(1)|1) = u(y(-1), z(-1)|-1) = c'(r) \geq 0$. Third, $u(y(0), z(0)|0) = 0$.

(O) Plugging in, (O) becomes $r \in \arg\max_{\hat{r} \in [0,1]} \{\hat{r} c'(r) - c(\hat{r})\}$. Then by convexity of $c(\cdot)$, the optimality of conducting research with intensity r follows from checking the first-order condition. □

We use this characterization of the feasible set to determine whether the various trading mechanisms implement points on or off the frontier of the set. Results along these lines are analogous to those stated in Sections IV.C and V.C of the main article.

First, Corollary 4 is an analogue of Corollary 2. It states that the LOB generally does not implement an outcome on the frontier of this alternative feasible set $\mathcal{F}'$. Second, Corollary 5 is an analogue of Corollary 5. It states that NDs and FBAs both do implement outcomes on the frontier of $\mathcal{F}'$.

COROLLARY 4: *If $X > 2$, $p_I < 1$, and $c'(0) < X_I$, then the LOB outcome is not on the frontier of $\mathcal{F}'$.*

Proof of Corollary 4: In the LOB equilibrium, investor welfare is related to the spread and research intensity through $w^*_{LOB} = (1-\lambda)(\beta - s^*_{LOB}/2) + \lambda r^*_{LOB} X_I (1 - s^*_{LOB}/2) - \lambda c(r^*_{LOB})$.

By Proposition 3, an equilibrium outcome lies on the frontier of $\mathcal{F}'$ if and only if $w = (1-\lambda)\beta - \lambda c(r)$. Therefore, the LOB outcome lies on the frontier of $\mathcal{F}'$ only if the following relationship holds:

$$s^*_{LOB} = \frac{2\lambda r^*_{LOB} X_I}{1-\lambda+\lambda r^*_{LOB} X_I}. \tag{IA39}$$

As before, r^*_{LOB} is characterized by the fixed point of the correspondence

$$R_{LOB}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{(1-\lambda) r X_I}{1-\lambda+\lambda\hat{r}(X_I+X_S)} - c(r) \right\},$$

where X_I and X_S are as defined in the statement of Proposition 1. The assumption that $c'(0) < X_I$ ensures that at $\hat{r} = 0$, the maximization problem on the right-hand side of the expression above for $R_{LOB}(\hat{r})$ does not have a solution at zero. Consequently, zero is not a fixed point of that correspondence, and so $r^*_{LOB} > 0$. Restating (1), we also have

$$s^*_{LOB} = \frac{2\lambda r^*_{LOB}(X_I+X_S)}{1-\lambda+\lambda r^*_{LOB}(X_I+X_S)}.$$

Given that $\lambda > 0$ and $r^*_{LOB} > 0$, comparing (IA39) to (1), we conclude that the LOB equilibrium outcome is on the frontier only if $X_S = 0$. Restating the expression for X_S, we obtain

$$\begin{aligned} X_S &= X(1-p_I) - X(1-p_I)^X - (X-1)Xp_I(1-p_I)^{X-1} \\ &= X(1-p_I)\left[\sum_{x=2}^{X-1}\binom{X-1}{x} p_I^x (1-p_I)^{X-1-x}\right]. \end{aligned}$$

Given that $p_I \geq 0.5$, $X_S = 0$ only if either $p_I = 1$ or $X \leq 2$ (or both). □

COROLLARY 5: *The following are true:*

(i) for all $q \in [0,1]$, the qND outcome is on the frontier of $\mathcal{F}'$;
(ii) the FBA outcome is on the frontier of $\mathcal{F}'$; and
(iii) if in addition either $X \leq 2$ or $p_I = 1$, then the LOB outcome is on the frontier of $\mathcal{F}'$.

Proof of Corollary 5: In the equilibria of the LOB, of NDs, and of FBAs, the equilibrium spread is related to investor welfare and research intensity through $w^* = (1-\lambda)(\beta - s^*/2) + \lambda r^* X^* (1 - s^*/2) - \lambda c(r)$, where X^* denotes the expected number of trades made by an

information investor conditional on arriving and learning the value of the security. By Proposition 3, an equilibrium outcome lies on the frontier of $\mathcal{F}'$ if and only if $w = (1-\lambda)\beta - \lambda c(r)$. Therefore, an equilibrium outcome of one of these trading mechanisms lies on the frontier of $\mathcal{F}'$ if the following relationship holds:

$$s^* = \frac{2\lambda r^* X^*}{1-\lambda+\lambda r^* X^*}. \tag{IA40}$$

Restating equations (1), (4), and (6), we also have

$$\begin{aligned} s^*_{LOB} &= \frac{2\lambda r^*_{LOB}(X_I + X_S)}{1-\lambda+\lambda r^*_{LOB}(X_I + X_S)} \\ s^*_{qND} &= \frac{2\lambda r^*_{qND} X_{qND}}{1-\lambda+\lambda r^*_{qND} X_{qND}} \\ s^*_{FBA} &= \frac{2\lambda r^*_{FBA} X}{1-\lambda+\lambda r^*_{FBA} X}. \end{aligned}$$

Because X_{qND} is the expected number of trades made by an information investor conditional on arriving and learning the value of the security under qND, equation (IA40) implies that the outcome of qND, for all $q \in [0,1]$, is on the frontier. Similarly, the outcome of FBAs is on the frontier. Finally, because X_I is the expected number of trades made by an information investor conditional on arriving and learning the value of the security under the LOB, the LOB outcome is on the frontier if $X_S = 0$, which is the case if either $p_I = 1$ or $X \leq 2$ (or both). □

B. Total Trader Welfare

In this appendix, we again characterize the set of feasible outcomes, but with respect to yet a different set of criteria, research intensity and total trader welfare.

There is, however, one subtlety that must be addressed in order to conduct this analysis, namely how to compute the criterion of total trader welfare. In the LOB equilibrium, positive trading profits are earned by HFTs, but they are divided among the infinite number of snipers that are active in equilibrium. It is therefore not obvious how to compute the aggregate profits of HFTs. Mathematically, this question is about how to take the limit that was alluded to in footnote 9 of the main article. As discussed in that footnote, one should think of our model as the limit of a sequence of models with finite numbers of HFTs. But what is the precise method for taking that limit?

Potential answers to this question include two "reasonable extremes." One possibility is to consider a sequence of models with a finite number of snipers $N_{HFT} \in \mathbb{N}$, each of whom can participate in the market without needing to pay an entry cost, and then consider the limit as $N_{HFT} \to \infty$. Along this sequence, and thus in the limit, all trades are transfers. So as long as each liquidity investor trades, total trader welfare equals $(1-\lambda)\beta - \lambda c(r)$. Thus, the feasible set collapses to a one-dimensional space, and each equilibrium of the model is trivially on the frontier.

We prefer a different possibility, however. In particular, we consider a sequence of models with a positive entry cost $c \in \mathbb{R}$ and where the number of active HFTs is determined by a free-entry condition. We then consider the limit as $c \to 0$. Along this sequence, and thus in the limit, aggregate HFT profits are zero.[41] It follows that the characterization of the feasible set boils down to that of Internet Appendix VII.A, that is, we continue to obtain analogues of the key results stated in Sections IV.C and V.C of the main article.

[41] This approach is consistent with an (unmodeled) HFT arms race in which all potential rents are dissipated by expenditures on speed technology (as in Budish, Cramton, and Shim (2015)).

VIII. Mathematical Notation

Name	Description
Model Parameters	
X	Number of exchanges*
λ	Probability of information investor
β	Liquidity investor private transaction motive
p_H	Probability an HFT obtains the lower latency
p_I	Probability the investor obtains the lower latency
Noncancellation Delay Parameters	
δ_{ND}	Constant component of delay
F_{ND}	Distribution of random component of delay
q	Probability of delay having a random component
Other Notation	
ε	Length of a time period†
$\mathcal{T}$	Set of time periods $\{0, \varepsilon, 2\varepsilon, \ldots, 1\}$
v	Fundamental value of security
s	Bid-ask spread‡
r	Research intensity‡
w	Liquidity investor welfare‡
$\mathcal{F}$	The set of feasible ordered pairs (r, w)
y, z	Number of dollars and number of shares in a portfolio
θ	Investor type
$u(y, z\|\theta)$	Investor utility (gross of research costs)
$c(r)$	Cost of research

* We also use X_I and X_S to denote the expected number of trades made by an information investor and snipers, respectively, under the LOB conditional on an information investor arriving and learning the value of the security. We further use X_{qND} to denote the expected number of trades made by an information investor under qND conditional on arriving and learning the value of the security.

† We also use N to denote the reciprocal of ε.

‡ We use stars to denote equilibrium values.

REFERENCES

Aequitas NEO Exchange, 2016, Trading functionality guide, https://www.aequitasneo.com/documents/en/trading-data/neo-exchange-trading-functionality-guide.pdf.

Aisen, Daniel, Bradley Katsuyama, Robert Park, John Schwall, Richard Steiner, Allen Zhang, and Thomas L. Popejoy, 2015, Synchronized processing of data by networked computing resources, US Patent 8,984,137.

Alpha Trading Systems Limited Partnership, 2016, Trading policy manual, http://www.tsx.com/resource/en/1069.

Anderson, Lisa, Emad Andrews, Baiju Devani, Michael Mueller, and Adrian Walton, 2018, Speed segmentation on exchanges: Competition for slow flow, Working paper, Bank of Canada.

Back, Kerry, and Shmuel Baruch, 2004, Information in securities markets: Kyle meets Glosten and Milgrom, *Econometrica* 72, 433–465.

Baker, Malcolm, Jeremy C. Stein, and Jeffrey Wurgler, 2003, When does the market matter? Stock prices and the investment of equity-dependent firms, *Quarterly Journal of Economics* 118, 969–1005.

Bakke, Tor-Erik, and Toni M. Whited, 2010, Which firms follow the market? An analysis of corporate investment decisions, *Review of Financial Studies* 23, 1941–1980.

Baldauf, Markus, and Joshua Mollner, forthcoming, Trading in fragmented markets, *Journal of Financial and Quantitative Analysis*.

Barclays Capital Inc. (Barclays), 2014, Why the smartness of order routing matters in options trading: How order routing performance impacts the bottom line, https://www.scribd.com/doc/204351010.

Bartlett, Robert P., and Justin McCrary, 2017, How rigged are stock markets? Evidence from microsecond timestamps, Working paper, University of California, Berkeley.

BATS Global Markets, Inc. (BATS), 2016, Bats system performance, http://cdn.batstrading.com/resources/features/bats_exchange_Latency.pdf Accessed: September 1, 2016.

Baumol, William J., 1965, *The Stock Market and Economic Efficiency* (Fordham University Press).

Berge, Claude, 1963, *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity* (Courier Dover Publications).

Bertsekas, Dimitri P., and Robert G. Gallager, 1992, *Data Networks (2nd Edition)* (Prentice Hall).

Besson, Paul, Matthieu Lasnier, and Antoine Falck, 2019, The benefits of European periodic auctions beyond MiFID dark trading caps, *The Journal of Investing*.

Bond, Philip, Alex Edmans, and Itay Goldstein, 2012, The real effects of financial markets, *Annual Review of Financial Economics* 4, 339–360.

Budish, Eric, Peter Cramton, and John Shim, 2014, Implementation details for frequent batch auctions: Slowing down markets to the blink of an eye, *American Economic Review: Papers & Proceedings* 104, 418–424.

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.

Budish, Eric, Robin Lee, and John Shim, 2019, Will the market fix the market? A theory of stock exchange competition and innovation, Working paper, University of Chicago.

Cboe EDGA Exchange, 2019, Notice of filing of a proposed rule change to introduce a liquidity provider protection on EDGA, *SEC Release No 34-86168* https://www.sec.gov/rules/sro/cboeedga/2019/34-86168.pdf.

Cboe Global Markets, 2015, Cboe Europe equities guidance note, periodic auctions book, http://cdn.batstrading.com/resources/participant_resources/BCE-GuidanceNote_Periodic_Auctions_Final.pdf Accessed: 17 June 2018.

Chen, Haoming, Sean Foley, Michael A. Goldstein, and Thomas Ruf, 2017, The value of a millisecond: Harnessing information in fast, fragmented markets, Working paper, University of New South Wales.

Chen, Qi, Itay Goldstein, and Wei Jiang, 2007, Price informativeness and investment sensitivity to stock price, *Review of Financial Studies* 20, 619–650.

Chicago Stock Exchange, 2016, Notice of filing of proposed rule change to adopt the CHX liquidity taking access delay, *SEC Release No. 34-78860* https://www.sec.gov/rules/sro/chx/2016/34-78860.pdf.

Chicago Stock Exchange, 2017, Notice of filing of proposed rule change to adopt the CHX liquidity enhancing access delay, *SEC Release No. 34-80041* https://www.sec.gov/rules/sro/chx/2017/34-80041.pdf.

Corvil, 2014, Electronic trading system performance, http://corvil.com/content/resources/04-white-papers/03-electronic-trading-system-performance/wp-electronic-trading-system-performance.pdf.

Deutsche Börse Group, 2017, Xetra release 17.0, market model continuous auction, http://www.xetra.com/blob/3160700/34bdd76de0cf73ca0e9121fca2edb0d5/data/Release-17-MM-Continuous-auction_e.pdf Accessed: 31 October 2018.

Deutsche Börse Group, 2018, Insights into trading system dynamics, http://www.eurexchange.com/blob/238346/3d4aa6471c4a964aacc091ced89ffc61/data/presentation_insights-into-trading-system-dynamics_en.pdf.

Deutsche Börse Group, 2019, Passive liquidity protection: Pilot phase and setup information, https://www.eurexchange.com/resource/blob/1506528/0374d7b91e93e8cab7d230247f68ef15/data/er19027e.pdf.

Diamond, Douglas W., and Robert E. Verrecchia, 1982, Optimal managerial contracts and equilibrium security prices, *Journal of Finance* 37, 275–287.

Dow, James, and Gary Gorton, 1997, Stock market efficiency and economic efficiency: Is there a connection?, *Journal of Finance* 52, 1087–1129.

Durnev, Art, Randall Morck, and Bernard Yeung, 2004, Value-enhancing capital budgeting and firm-specific stock return variation, *Journal of Finance* 59, 65–105.

Easley, David, and Maureen O'Hara, 2004, Information and the cost of capital, *Journal of Finance* 59, 1553–1583.

Egginton, Jared F., Bonnie F. Van Ness, and Robert A. Van Ness, 2016, Quote stuffing, *Financial Management* 45, 583–608.

Euronext, 2018, Euronext rule book book I: Harmonised rules, https://www.euronext.com/sites/www.euronext.com/files/harmonised_rulebook_en_2018_07_31_euronext_dublin.pdf Accessed: 31 October 2018.

Faure-Grimaud, Antoine, and Denis Gromb, 2004, Public trading and private incentives, *Review of Financial Studies* 17, 985–1014.

Ferreira, Daniel, Miguel A. Ferreira, and Clara C. Raposo, 2011, Board structure and price informativeness, *Journal of Financial Economics* 99, 523–545.

Fishman, Michael J., and Kathleen M. Hagerty, 1989, Disclosure decisions by firms and the competition for price efficiency, *Journal of Finance* 44, 633–646.

Foreign Exchange Professionals Association, 2015, Focus on last look, https://fxpa.org/wp-content/uploads/2016/01/FXPA-lastlook-final.pdf.

Foucault, Thierry, and Laurent Fresard, 2014, Learning from peers' stock prices and corporate investment, *Journal of Financial Economics* 111, 554–577.

Gai, Jiading, Chen Yao, and Mao Ye, 2013, The externalities of high frequency trading, Working paper, University of Illinois at Urbana-Champaign.

Gider, Jasmin, Simon N. M. Schmickler, and Christian Westheide, 2019, High-frequency trading and price informativeness, Working paper, Tilburg University.

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable?, *Journal of Finance* 49, 1127–1161.

Glosten, Lawrence R., 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.

Goldman Sachs, 2018, Goldman Sachs electronic trading: SIGMA X MTF participant manual, https://gset.gs.com/Sigmaxmtf/Public/GetDocument/12d3f6c6-df1c-48fb-a141-0285af3f0054?compName=ParticipantDocs Accessed: 22 July 2019.

Harris, Larry, 2013, What to do about high-frequency trading, *Financial Analysts Journal* 69, 6–9.

Hayek, Friedrich A., 1945, The use of knowledge in society, *American Economic Review* 35, 519–530.

Hirschey, Nicholas H., 2019, Do high-frequency traders anticipate buying and selling pressure?, Working paper, London Business School.

Holmström, Bengt, and Jean Tirole, 1993, Market liquidity and performance monitoring, *Journal of Political Economy* 101, 678–709.

Hu, Edwin, 2019, Intentional access delays, market quality, and price discovery: Evidence from IEX becoming an exchange, DERA Working Paper, U.S. Securities and Exchange Commission.

ICE Futures U.S., 2019, Amendments to Rule 4.26 order execution (new passive order protection functionality) submission pursuant to Section 5c(c)(1) of the act and Regulation 40.6(a), Open Letter to CFTC, https://www.cftc.gov/sites/default/files/2019-02/ICEFuturessPassiveOrder020119.pdf.

IEX Group, 2018a, Investors Exchange rule book, https://iextrading.com/docs/Investors%20Exchange%20Rule%20Book.pdf.

IEX Group, 2018b, Router stats, https://www.iextrading.com/stats/#router-stats.

Investment Technology Group, 2017, POSIT MTF: FIX and connectivity guidance, https://www.virtu.com/uploads/documents/POSIT-MTF-FIX-and-Connectivity-guide_Updated-06.25.2018.pdf Accessed: 22 July 2019.

Ixia, 2012, Measuring latency in equity transactions, https://intl.ixiacom.com/sites/default/files/resources/whitepaper/lowlatencywhitepaperbooklet.pdf.

Kakutani, Shizuo, 1941, A generalization of Brouwer's fixed point theorem, *Duke Mathematical Journal* 8, 457–459.

Kang, Qiang, and Qiao Liu, 2008, Stock trading, information production, and executive incentives, *Journal of Corporate Finance* 14, 484–498.

Kau, James B., James S. Linck, and Paul H. Rubin, 2008, Do managers listen to the market?, *Journal of Corporate Finance* 14, 347–362.

Kay, Rony, 2009, Pragmatic network latency engineering fundamental facts and analysis, https://pdfs.semanticscholar.org/a19d/53a533fe78d01f0423f628e4d1688be23e6d.pdf.

KCG Holdings, Inc. (KCG), 2014, The need for speed II.

Kirilenko, Andrei A., and Gui Lamacie, 2015, Latency and asset prices, Working paper, University of Cambridge.

Korajczyk, Robert A., and Dermot Murphy, 2019, High frequency market making to large institutional trades, *Review of Financial Studies* 32, 1034–1067.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Lehr, Michael, 2016, The latency difference between depth of book and BBO feeds, http://www.maystreet.com/api/files/mst_drive/public/TheLatencyDifferenceBetweenDepthAndBBO-MayStreet.pdf.

Lewis, Michael, 2014, *Flash Boys: A Wall Street Revolt* (W. W. Norton & Company).

Lin, Tse-Chun, Qi Liu, and Bo Sun, 2019, Contractual managerial incentives with stock price feedback, *American Economic Review* 109, 2446–68.

London Metal Exchange, 2019, Technical change to LMEselect FIX message processing for the LMEprecious market to introduce a fixed minimum delay https://www.lme.com/-/media/Files/News/Notices/2019/05/19-165-Technical-change-to-LMEselect-FIX-message-processing-for-the-LMEprecious-marke pdf.

London Stock Exchange, 2015, Factsheet, further improvements to auctions on the SETSqx trading service, https://www.lseg.com/sites/default/files/content/documents/Further%20improvements%20to%20auctions%20on%20the%20SETSqx%20Trading%20Service.pdf Accessed: 31 October 2018.

London Stock Exchange, 2019, Turquoise trading service description, https://www.lseg.com/sites/default/files/content/documents/Turquoise%20Trading%20Services%20Description%203.34.9i.PDF Accessed: 6 February 2019.

Luo, Yuanzhi, 2005, Do insiders learn from outsiders? Evidence from mergers and acquisitions, *Journal of Finance* 60, 1951–1982.

Malinova, Katya, and Andreas Park, 2017, Does high frequency trading add noise to prices?, Working paper, McMaster University.

Melton, Hayden Paul, 2015, Ideal latency floor, US Patent App. 14/533,543.

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

Mercer, R.L., and P.F. Brown, 2016, System and method for executing synchronized trades in multiple exchanges, US Patent App. 14/451,356.

Merton, Robert C., 1974, On the pricing of corporate debt: The risk structure of interest rates, *Journal of Finance* 29, 449–470.

MIAX Miami International Securities Exchange, 2018, MIAX infrastructure and latest performance numbers, https://www.miaxoptions.com/node/89.

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

Moscow Exchange, 2019, Moscow Exchange to expand FX offering https://www.moex.com/n23386.

Myers, Stewart C., and Nicholas S. Majluf, 1984, Corporate financing and investment decisions when firms have information that investors do not have, *Journal of Financial Economics* 13, 187–221.

Nanex Research (Nanex), 2014, Perfect pilfering, http://www.nanex.net/aqck2/4661.html.

NASDAQ OMX PHLX, 2012, Notice of filing of proposed rule change to modify Exchange Rule 3307 to institute a five millisecond delay in the execution time of marketable orders on NASDAQ OMX PSX, *SEC Release No. 34-67680* https://www.sec.gov/rules/sro/phlx/2012/34-67680.pdf.

NASDAQ OMX Group (NASDAQ OMX), 2012, INET technology and the Nasdaq stock market, http://www.nasdaqtrader.com/Trader.aspx?id=Latencystats Accessed: July 5, 2017.

NYSE, 2018, NYSE Pillar, https://www.nyse.com/pillar.

NYSE Euronext (NYSE), 2018, Equities rules, http://wallstreet.cch.com/AmericanTools/PlatformViewer.asp?SelectedNode=chp_1_5&manual=/american/rules/american-rules/.

Peterffy, Thomas, 2014, Interactive Brokers Group proposal to address high frequency trading, Open Letter to SEC, https://www.interactivebrokers.ca/download/SEC_proposal_high_frequency_trading.pdf.

Shkilko, Andriy, and Konstantin Sokolov, 2019, Every cloud has a silver lining: Fast trading, microwave connectivity and trading costs, Working paper, Wilfrid Laurier University.

Simon, Leo K., and Maxwell B. Stinchcombe, 1989, Extensive form games in continuous time: Pure strategies, *Econometrica* 57, 1171–1214.

SIX Swiss Exchange, 2016, Excellent latency and capacity, http://www.six-swiss-exchange.com/download/participants/trading/x-stream_inet_performance_measurement_details.pdf.

Subrahmanyam, Avanidhar, and Sheridan Titman, 1999, The going-public decision and the development of financial markets, *Journal of Finance* 54, 1045–1082.

Subrahmanyam, Avanidhar, and Sheridan Titman, 2001, Feedback from stock prices to cash flows, *Journal of Finance* 56, 2389–2413.

Tel Aviv Stock Exchange, 2018, TASE trading schedule, illiquid securities and maintenance lists, https://info.tase.co.il/Eng/trading/trading_schedule/Pages/trading_schedule.aspx Accessed: 31 October 2018.

TMX Group, 2014, Tmx group readies launch of tmx quantum xa on toronto stock exchange, press release, https://www.tmx.com/newsroom/press-releases?id=59&year=2014.

Topkis, Donald M., 1978, Minimizing a submodular function on a lattice, *Operations Research* 26, 305–321.

van Kervel, Vincent, 2015, Competition for order flow with fast and slow traders, *Review of Financial Studies* 28, 2094–2127.

van Kervel, Vincent, and Albert J. Menkveld, 2019, High-frequency trading around large institutional orders, *Journal of Finance* 74, 1091–1137.

Weller, Brian M., 2018, Does algorithmic trading reduce information acquisition?, *Review of Financial Studies* 31, 2184–2226.

Wurgler, Jeffrey, 2000, Financial markets and the allocation of capital, *Journal of Financial Economics* 58, 187–214.