SIEPR Discussion Paper No. 15-017

High-Frequency Trading and Market Performance

By

Markus Baldauf and Joshua Mollner

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

High-Frequency Trading and Market Performance*

Markus Baldauf Joshua Mollner

May 13, 2015

Abstract

High-frequency trading has transformed financial markets in recent years. We study the consequences of this development using a model with multiple trading venues, costly information acquisition, and several types of traders. An increase in trading speed crowds out information acquisition by reducing the gains from trading against mispriced quotes. Thus, faster speeds have two effects on traditional measures of market performance. First, the bid-ask spread declines, since there are fewer informational asymmetries. Second, price efficiency deteriorates, since less information is available to be incorporated into prices. A general tradeoff exists between low spreads and price efficiency. We characterize the frontier of this tradeoff and evaluate several trading mechanisms within this framework. The prevalent limit order book mechanism generally does not induce outcomes on this frontier. We consider two alternatives: first, a small delay added to the processing of all orders except cancellations, and second, frequent batch auctions. Both induce equilibrium outcomes on this frontier.

*We are indebted to our advisors Timothy Bresnahan, Gabriel Carroll, Jonathan Levin, Monika Piazzesi, and Paul Milgrom. We would also like to thank Sandro Ambuehl, Eric Budish, Darrell Duffie, Liran Einav, Joseph Grundfest, Terrence Hendershott, Fuhito Kojima, Muriel Niederle, Alvin Roth, Ilya Segal, Andrzej Skrzypacz, and seminar participants at Stanford, as well as various industry experts 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. Mailing address: Stanford University Economics Department, 579 Serra Mall, Stanford, CA 94305. Email: baldauf@stanford.edu (Baldauf), jmollner@stanford.edu (Mollner).

1 Introduction

Financial markets have recently seen drastic improvements in speed by both traders and exchanges. For example, The New York Stock Exchange has slashed the amount of time it requires to process an order by two orders of magnitude, from one second in 2004 to five milliseconds in 2009 (NYSE, 2004, 2009). Furthermore, the minimum feasible round-trip travel time of communication between NASDAQ and the Chicago Mercantile Exchange, which is a measure of trader speed, has declined from over 14.5 milliseconds in 2010 (Adler, 2012) to under 8.1 milliseconds today (McKay Brothers, 2014). In addition, another important feature of modern trading is that it is dispersed across a large number of venues. Many large-cap stocks are now traded at forty or more venues, a number that is much larger than just a decade ago.

Given the multiplicity of trading venues, the recent improvements in speed have increased the effectiveness of certain strategies used by high-frequency traders, one of which is described next. For reasons including variable network traffic, the time required to send an order to an exchange is not perfectly predictable. Therefore, traders are unable to ensure simultaneous arrival of orders sent to several exchanges. If high-frequency traders are sufficiently fast, they may observe the trade generated by the first order to arrive and react on the other exchanges before the orders of the original trader arrive there. This practice is referred to as order anticipation, and it has significantly affected outcomes in these markets.[1] To illustrate, a typical modern trader who attempts to trade against posted quotes on six or more exchanges does so successfully only 25 percent of the time (Barclays, 2014).

Due to the sheer scale of these markets, order anticipation is responsible for large transfers within the financial system. Another concern deals with price efficiency. If the victims of order anticipation are traders who conduct research into fundamentals, then order anticipation may reduce the profitability of that research. Less research would then be conducted,

[1]This practice has also been referred to as "front-running" in the media. However, in the context of financial markets "front-running" is more appropriately applied to the illegal practice addressed in FINRA Rule 5270, which prohibits a broker-dealer from trading for its own account while taking advantage of knowledge of an imminent client block transaction. Order anticipation, on the other hand, is legal.

divorcing stock prices from the fundamental value of the underlying asset, which might generate further distortions in the wider economy.

In this paper, we present a theoretical model of order anticipation. We show that it may indeed harm price efficiency, but a positive effect is that it may reduce transaction costs, as measured by the bid-ask spread. We also use the model to evaluate the merits of the limit order book, the predominant mechanism used by exchanges today, relative to some alternative trading mechanisms. The model features an asset that is traded on multiple exchanges by three types of traders: an analyst, high-frequency traders, and investors. The analyst, through costly research, may become privately informed about the value of the asset. High-frequency traders may trade for profit by speculating or by facilitating transactions with other traders. Investors arrive at the market with exogenous liquidity motives to buy or sell one indivisible share, and they play the role of ordinary traders.

In equilibrium, the analyst submits orders to all exchanges upon attaining information about the value of the asset. However, random communication latency prevents these orders from arriving simultaneously. After the first order arrives at an exchange, high-frequency traders infer the analyst's information and respond by engaging in order anticipation at the remaining exchanges. In the equilibrium we identify, high-frequency traders sort into two roles. One, the liquidity provider, facilitates trade by posting quotes at all exchanges; she responds to the analyst's trade by attempting to cancel her remaining mispriced quotes. The others, stale-quote snipers, respond to the analyst's trade by free-riding on his information and attempting to trade in front of the analyst against the remaining mispriced quotes. This gives rise to winner-take-all races on the remaining exchanges, which may be won by either the analyst, the liquidity provider, or a stale-quote sniper.

As in Glosten and Milgrom (1985) and Budish, Cramton, and Shim (2013)—henceforth, BCS—a central feature of the model is that the liquidity provider faces adverse selection: the analyst and stale-quote snipers trade against her quotes only when those quotes are mispriced. To offset the losses from adverse selection, the liquidity provider must generate revenue from trades with investors, which she does by setting a bid-ask spread. Worse

adverse selection must be compensated by a larger spread.

Using this model, we evaluate the consequences of recent improvements in speed by exchanges and traders. Speed improvements enable high-frequency traders to be more successful at order anticipation, which reduces the amount of rent that the analyst can extract by trading on a piece of information. By reducing the incentives to conduct research, faster speeds lead to a lower equilibrium research intensity, which affects traditional measures of market performance in two ways. First, the bid-ask spread declines. Intuitively, since less research is being done, the liquidity provider is less exposed to adverse selection from the analyst, so she can afford to demand a smaller spread. This prediction is in line with a great deal of empirical evidence. Second, price efficiency diminishes, in the sense that prices become less correlated with the fundamental value of the underlying asset. Intuitively, since less research is being done, less information is available to be incorporated into prices. Notably, this second prediction highlights an omission by many empirical studies of the topic. Those studies have documented that information, *conditional on being incorporated into prices*, is incorporated more rapidly when exchanges and traders are faster. They have interpreted this as evidence that speed improves price efficiency. Those studies, however, do not control for the effect of speed on the amount of information that becomes incorporated into prices. On the other hand, our model does consider this channel and finds that it dominates the first, so that the net effect of speed improvements is to harm price efficiency.

To summarize, when trading is governed by the limit order book, speed improvements give rise to a tradeoff: price efficiency diminishes, but the bid-ask spread declines. We therefore proceed to study whether alternative trading mechanisms can be used to obtain improvements with respect this tradeoff between spreads and price efficiency. The analysis focuses on two specific proposals, one new to the literature and one familiar. First, we propose adding a small delay to the processing of all orders except cancellations. Second, we consider the performance of frequent batch auctions.

The first proposal is to add a small delay between arrival at an exchange and processing for all order types except cancellations. We refer to this as a *selective delay*. Intuitively, a

selective delay gives the liquidity provider the ability to cancel mispriced quotes before they can be exploited by snipers. It therefore prevents stale-quote snipers from free-riding on the information of the analyst, as they could with a limit order book. We also characterize the frontier of the tradeoff between spreads and price efficiency by formulating and solving a social planner's problem. We find that by eliminating this free-riding, a selective delay implements an equilibrium on the frontier of this tradeoff.

The second proposal is to replace continuous trading with frequent batch auctions, which are uniform-price sealed-bid double auctions conducted repeatedly at discrete time intervals. The batch auction proposal has received a great deal of recent attention, a notable example being BCS. We follow them by considering within our model the performance of frequent batch auctions in which batch intervals are "long" relative to communication latency and synchronized across exchanges. Like a selective delay, frequent batch auctions implement an equilibrium on the frontier of the tradeoff between spreads and price efficiency. However, the frequent batch auction equilibrium features a higher spread and higher price efficiency than the selective delay equilibrium. Intuitively, the batching not only prevents stale-quote snipers from free-riding on the information of the analyst, but also prevents the liquidity provider from canceling mispriced quotes before they can be exploited by the analyst. Research therefore becomes more valuable for the analyst, which induces a higher research intensity and higher price efficiency. On the other hand, the liquidity provider—by being unable to practice order anticipation—faces more adverse selection and therefore demands a larger spread.

The remainder of the paper is organized as follows. Section 2 discusses the related literature. Section 3 describes the model. Section 4 describes the equilibrium that prevails when trading is governed by the limit order book mechanism, and it also assesses theoretically the consequences of the recent increases in trading speed. Section 5 characterizes the outcomes prevailing under the two aforementioned alternative trading mechanisms and discusses how they compare to the limit order book. Section 6 characterizes the frontier of the tradeoff between spreads and price efficiency. Section 7 concludes. All proofs are

contained in Appendix A.

2 Related Literature

Our model fits into the branch of the literature that has focused on financial markets with asymmetric information. Some models in this class are Copeland and Galai (1983), Glosten and Milgrom (1985), Kyle (1985), Glosten (1994), and Back and Baruch (2004). More recently, BCS demonstrate how similar forces arise in a limit order book when multiple high-frequency traders react to the same piece of information.

Our model is also connected to the literature on information acquisition in financial markets. Central to that literature is Grossman and Stiglitz (1980), who study the incentives to engage in costly information acquisition, and the repercussions on price efficiency. In the equilibrium of their model, prices adjust to reflect the information of the informed, but only partially so.

The model of this paper lies at the confluence of these two literatures. It features an asset that is traded on multiple limit order books by several types of traders, one of which may pay a cost to acquire information. Using this model, we study how the incentives to acquire information are affected by features of the microstructure of the trading environment, including the speed of exchanges and traders, the number of exchanges in operation, and the mechanism that governs trading.

This paper is most closely related to Glosten and Milgrom (1985) and BCS, with two primary differences. First, we explicitly model a fragmented financial system in which several exchanges operate simultaneously. Many strategies used in practice by high-frequency traders (e.g. order anticipation) hinge crucially on the presence of multiple exchanges. This feature, therefore, allows these strategies to be explicitly modeled. Second, our model endogenizes the amount of information possessed by informed traders. This feature, therefore, allows price efficiency to depend upon the trading mechanism as well as parameters such as the speeds of traders and exchanges.

Whereas prior models have tended to focus on either spreads or price efficiency in isolation, our model endogenizes both quantities. We contribute to the literature by (*i*) demonstrating the existence of a general tradeoff between these two quantities, (*ii*) characterizing the frontier of this tradeoff, and (*iii*) evaluating several trading mechanisms within this framework.

Others have studied very different models of high-frequency trading. For example, Biais, Foucault, and Moinas (2013), Foucault, Hombert, and Roşu (2013), and Martinez and Roşu (2013) present models in which high-frequency traders possess an informational advantage over the liquidity provider, or the agent who makes the market. In our model, in contrast, the liquidity provider *is* a high-frequency trader.[2] Consequently, our model gives rise to very different predictions about the effects of high-frequency trade. In particular, their models predict that if high-frequency traders become faster or better informed, then adverse selection increases and the market becomes less liquid. However, in our model, since the liquidity provider is a high-frequency trader, the increase in speed helps her avoid adverse selection from the analyst. We therefore obtain the opposite prediction: spreads decline when high-frequency traders become faster.[3]

Additionally, several others have attempted to evaluate how the market would perform under alternative trading mechanisms. Frequent batch auctions have received the most attention, having been considered by BCS, as well as by others, including Madhavan (1992) and Wah and Wellman (2013). Our findings are most easily compared with those of BCS. In their model, batching reduces adverse selection from stale-quote snipers, which results in smaller spreads. However, our model also features a second source of adverse selection: an analyst who possesses private information. We show that batching actually increases this source of adverse selection; moreover, this effect dominates so that batching leads to larger spreads in our model. Furthermore, we also find that batching improves price efficiency.

[2]Our model is similar in this respect to BCS. Note also that this feature of the model is corroborated by empirical evidence, for example Menkveld (2013), who studies a large high-frequency trading firm and finds that 78% of its trades are passive (i.e. liquidity providing).

[3]This conclusion is also supported by empirical evidence. Two notable examples are Hasbrouck and Saar (2013) and Hendershott, Jones, and Menkveld (2011).

3 Model

The model features an asset that may be traded in multiple limit order books by three types of traders: an analyst, investors, and high-frequency traders. The details of the model build primarily upon the BCS framework, with two primary differences. First, we allow for multiple exchanges. Second, information arrives privately via costly research, as opposed to via exogenous public revelation.

3.1 Trading environment

Time. Time evolves over the interval $[0, T]$. We employ a continuous time construction, in which we allow for infinitesimal time intervals. Specifically, we index points in time by elements of the hyperreals, ${}^*\mathbb{R}$, which are an ordered field extension of the real numbers that contain nonzero infinitesimals.[4]

Certain aspects of the model, such as the processing time of exchanges and the communication latency of traders, are defined to occur on timescales measured in infinitesimals. This construction approximates the reality of incredibly fast speeds in modern markets. Moreover, it allows for a clean model by formalizing the following notion: traders and exchanges are unable to react instantaneously to the arrival of information, yet are able to react so quickly that additional information arrives before the reaction has completed with only a negligible probability.

Asset. There is a single asset whose fundamental value at time t is v_t. Trading begins at $t = 0$, at which point the fundamental value v_0 is public information. Trading ends at $t = T$.[5] During the interval $[0, T]$, v_t evolves as a compound Poisson jump process with

[4] An *infinitesimal* $\varepsilon \in {}^*\mathbb{R}$ is a number for which $|\varepsilon| < \frac{1}{n}\ \forall n \in \mathbb{N}$. The hyperreals are the objects used in a branch of mathematics known as nonstandard analysis (Robinson, 1966; Goldblatt, 1998). A key result of nonstandard analysis is the *transfer principle*, which states that a sentence is true over $\mathbb{R}$ if and only if a corresponding sentence is true over ${}^*\mathbb{R}$. This is useful for us because it allows us to define random variables and compute probabilities that involve the hyperreals in the natural way.

[5] For example, the asset may be a company, and the times $\{0, T\}$ may represent the dates of release of quarterly earnings reports. Changes in v_t may represent realizations of profits, which are not made public until after the release of the next quarterly earnings report.

arrival rate $\lambda_{jump} \in \mathbb{R}_+$.[6] When a jump arrives, v_t either increases or decreases by one, each with equal probability.

Exchanges. There are X exchanges, each of which allows shares of the asset to be traded throughout the interval $[0, T]$. Shares are indivisible. After trading has ended, v_T is made public, and all traders with a net position in the asset are compensated at v_T per share. In the baseline model, each exchange is organized as a limit order book, the structure of which is described below. We later consider alternative trading mechanisms. Order flow is non-anonymous and is publicly observed after an infinitesimal delay of length $\delta_E \in {}^*\mathbb{R}_+$.[7]

Limit order book. The benchmark trading environment in the model is a limit order book. A limit order book, at any point in time, is a collection of active limit orders. In what follows, we refer to four types of orders. A *limit order* consists of (*i*) the number of shares desired to transact, positive if the trader wishes to sell or negative if the trader wishes to buy, (*ii*) a price, and (*iii*) a time until when the order stays in force. Limit orders, unless otherwise specified, are assumed to be "good 'til cancelled." An *immediate or cancel (IOC) order* is a limit order with a time in force of zero. A *market order* may be thought of as an IOC order with a limit price of positive or negative infinity. A *cancellation order* instructs the exchange to remove an active order from the book.

Orders are processed sequentially, in the order they are received. Incoming limit orders are processed as follows. First, it is checked whether the incoming order specifies a price that allows trade with any orders residing in the book. If so, then the order leads to a trade at the price of the order in the book. If no match is found then the order is added to the book.

The *bid* is the highest price at which there exists an offer to buy. The *ask* is the lowest

[6] Throughout we use $\mathbb{R}_+$ to denote the set $\{x \in \mathbb{R} \mid x > 0\}$. Similarly, ${}^*\mathbb{R}_+ = \{x \in {}^*\mathbb{R} \mid x > 0\}$.

[7] Depending upon the rules of the particular exchange, anonymous trading may or may not be allowed in practice. At exchanges where the identities of traders are not immediately observable, traders use sophisticated statistical methods to attempt to infer the true identities of anonymous traders. Therefore, while the model does not directly apply to exchanges that allow anonymous trading, we believe our results to be indicative of what would transpire in such an environment. Moreover, the assumption of non-anonymous trading is not uncommon in the literature, for example Sannikov and Skrzypacz (2014).

price at which there exists an offer to sell. The *mid-price* is the average of the bid and ask. The *spread* is the difference between the bid and ask. The spread is a measure of transaction costs, and in this model determines the welfare of ordinary traders.

3.2 Traders

There are three types of traders: an analyst, investors, and high-frequency traders.[8] The analyst obtains private information about v_t through costly effort. Investors wish to buy or sell for exogenous reasons that may include hedging, saving, borrowing, or liquidity motives. High-frequency traders facilitate trade with investors and the analyst. All traders are risk-neutral, do not discount the future, and act to maximize the standard part of their expected utility.[9]

Analyst. There is a single analyst. At each point in time t, he chooses a research intensity $r_t \in [0,1]$ at the flow cost $c(r_t)$. Conditional on a jump of v_t occurring at time t, the analyst observes the jump with probability r_t. We assume $c(r)$ is continuous. The analyst's objective is to maximize profits net of research costs.

At any time t, the action space of the analyst is (i) a choice of research intensity, and (ii) whether to submit any orders. We place two restrictions on orders that the analyst may send. First, he is restricted to using IOC orders.[10] Second, the analyst is restricted to sending orders to buy (sell) only at times when there was an upward (downward) jump in the value of the security, which prevents him from engaging in trade-based market manipulation, a practice that is prohibited in most countries.[11]

[8] High-frequency traders are modeled as being similar to the market makers of BCS, and investors are also similar to their counterparts in that paper. However, the analyst, through whom the amount of information is endogenized, has no counterpart in BCS.

[9] In nonstandard analysis, the *standard part* of a number $x \in {}^*\mathbb{R}$ is the unique real number whose difference from x is an infinitesimal. In effect, we assume that agents treat events with infinitesimal probabilities as though they have probability zero.

[10] This is a technical restriction, which ensures that the analyst does not provide liquidity, and it is standard in the literature, for example Glosten and Milgrom (1985). Later, we impose the same technical restriction on investors, which is also standard, for example Glosten and Milgrom (1985) and BCS.

[11] In particular, trade-based market manipulation is illegal in the United States under Sections 10(b) and 9(a)(2) of the Securities Exchange Act of 1934, as well as SEC Rule 10b-5 (Nelemans, 2008). For example, Section 10(b) states, "It is unlawful ... [t]o use or employ, in connection with the purchase or sale of any

Investors. Investors arrive at Poisson rate $\lambda_{invest} \in \mathbb{R}_+$, at which point they randomly select one of the exchanges.[12] With equal probability, an investor is either a "buyer," who wishes to buy one share, or a "seller," who wishes to sell one share. From acquiring a portfolio consisting of x shares and y dollars between the time of arrival and time T, a buyer receives utility $u_B(x, y) = y + xv_T + \theta \mathbb{1}\{x = 1\}$, and a seller receives utility $u_S(x, y) = y + xv_T + \theta \mathbb{1}\{x = -1\}$. That is, an investor's utility is determined by his trading profits, in addition to a utility bonus of θ if his trading need is satisfied.

The action space of an investor who arrives at time t, at any time $t' \geq t$, is whether to send any orders. We place two restrictions on orders that investors may send. First, they are restricted to using IOC orders. Second, an investor is restricted to sending orders to his selected exchange.

High-frequency traders. There is an infinite number of high-frequency traders.[13] Their objectives are to maximize profits. At any time t, the action space of a high-frequency trader is whether to submit any orders. High-frequency traders may use any type of limit order, as well as cancellations.

Communication latency. Communication latency, which is the amount of time needed for a trader to send an order to an exchange, is a random variable that is measured on infinitesimal time scales.[14] Formally, the amount of time required for a trader to send an order to an exchange is drawn from a shifted exponential distribution. For the analyst, this distribution has the minimum δ_A and mean $\delta_A + \mu_A$. For high-frequency traders, these values

security . . . any manipulative or deceptive device or contrivance" (United States Code, 1934). Moreover, such violations are often detected and punished. For example, Aggarwal and Wu (2006) identify 142 instances of SEC litigation concerning trade-based market manipulation occurring between 1990 and 2001. Therefore, this restriction may be thought of as coming from optimal behavior on the part of the analyst if at some point in the future he would be audited and punished for manipulation.

[12] See Baldauf and Mollner (2015) for a similar model in which investors, rather than choosing exchanges randomly, choose according to trading conditions at the various exchanges.

[13] In practice the number of high-frequency traders is quite large. For example, Baron, Brogaard, and Kirilenko (2012) identify 65 separate high-frequency trading firms that actively trade the E-mini S&P contract in August 2010. Furthermore, since each firm may employ several different high-frequency trading algorithms, the effective number of competitors may be even higher.

[14] In practice, communication latency may not be perfectly predictable for several reasons, including the amount of traffic in the network, equipment glitches, and static.

are δ_H and $\delta_H + \mu_H$. For investors, these values are δ_I and $\delta_I + \mu_I$. These six parameters are assumed to be infinitesimals that are "on the order of" some fixed positive infinitesimal $\varepsilon \in {}^*\mathbb{R}_+$ in the sense that they are neither infinitely larger nor infinitely smaller than ε.[15]

Let $L_{i,x,t}$ denote the amount of latency for an order submitted by trader i to exchange x at time t. We assume the following correlation structure: $L_{i,x,t} = L_{i',x',t'}$ if $i = i'$, $x = x'$, and $|t - t'|$ is an infinitesimal; they are otherwise independent.[16] That is, communication latencies are independent except for messages sent by the same trader to the same exchange at "almost" the same time.

3.3 Assumptions

Most results that follow rely on Assumptions 1 and 2, which are stated below. These assumptions place restrictions on the parameter space, which are sufficient to guarantee the existence of equilibria in which (*i*) investors trade, and (*ii*) the analyst trades each time he observes a jump.

Assumption 1 (investor participation)**.** $\dfrac{2\lambda_{jump}X}{\lambda_{invest} + \lambda_{jump}X} \leq 2\theta.$

Assumption 2 (analyst participation)**.** $\dfrac{2\lambda_{jump}X}{\lambda_{invest} + \lambda_{jump}X} \leq 1.$

Assumption 1 is sufficient to ensure that the equilibrium spread is not so large that it exceeds 2θ and therefore crowds out all trades by investors. If the spread did exceed 2θ, then the market would shut down due to adverse selection, as only informed trading would occur.

Assumption 2 guarantees that the equilibrium spread is not larger than the size of a jump, which is sufficient to ensure that the analyst finds it optimal to trade each time he observes a jump. If the spread did exceed the size of a jump, then the analyst might prefer

[15] An element $a \in {}^*\mathbb{R}$ is said to be *infinitely larger* than an element $b \in {}^*\mathbb{R}$ iff a is nonzero and $\frac{b}{a}$ is an infinitesimal. Similarly, a is said to be *infinitely smaller* than b iff b is nonzero and $\frac{a}{b}$ is an infinitesimal.

[16] In the language of non-standard analysis, when $|t - t'|$ is an infinitesimal, t and t' are said to be *infinitely close*. The role of this assumption is to rule out the possibility that an order sent at time t could arrive after an order sent by the same trader to the same exchange at some time $t' > t$.

to wait for several jumps to accumulate before trading, which would present technical issues by breaking the stationary nature of the equilibrium.

4 Limit Order Book Equilibrium

In this section, we study the baseline model in which each exchange is organized as a limit order book. We describe equilibrium trading behavior in this environment, and we discuss the comparative statics of this equilibrium.

4.1 Equilibrium Description

This section considers equilibrium behavior of the analyst, investors, and high-frequency traders within the limit order book environment. Theorem 1 characterizes the outcomes that arise in equilibrium, focusing on two outcome variables: (*i*) the bid-ask spread, s^*_{LOB}, and (*ii*) the research intensity, r^*_{LOB}. The equilibrium we identify is stationary, in which these outcome variables are constant throughout $[0, T]$.

Theorem 1. *Under Assumptions 1 and 2, there exists a Nash equilibrium of the limit order book mechanism in which the spread, s^*_{LOB}, and research intensity, r^*_{LOB}, are constants given by any solution to*

$$s^*_{LOB} = \frac{2r^*_{LOB}\lambda_{jump}X}{\lambda_{invest} + r^*_{LOB}\lambda_{jump}X} \tag{1}$$

$$r^*_{LOB} \in \arg\max_{r\in[0,1]} \left\{ r\lambda_{jump}\left[X - (X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]\left(1 - \frac{s^*_{LOB}}{2}\right) - c(r)\right\} \tag{2}$$

All proofs are deferred to Appendix A. While a complete description of the strategies that support these outcomes in Nash equilibrium is given in the proof of this result, we sketch these strategies here.[17]

[17] It can be shown that the Nash equilibrium we identify also survives a continuous time version of the perfect Bayesian equilibrium refinement, in which the other traders form beliefs about the analyst's information by observing his trades. When they observe an episode in which the analyst buys, then they infer that he has observed an upward jump. Similarly, when they observe an episode in which the analyst sells, then they infer that he has observed a downward jump. These beliefs are indeed consistent with the analyst's strategy, and all strategies are optimal given these beliefs.

Investors submit orders to buy or sell according to their private transaction motives. They submit these orders immediately upon arrival and to their selected exchanges. The analyst submits orders to buy or sell according to the directions of the jumps he observes. He submits these orders immediately upon observing a jump and to all exchanges. While the analyst's orders are sent simultaneously to all exchanges, the randomness of communication latency prevents these orders from arriving simultaneously, and high-frequency traders react to the trade triggered by the first of a series of his orders. The way in which high-frequency traders react resembles the way in which market makers react to public information in BCS. The nature of their reaction is determined by which of two roles they play. One is the "liquidity provider." The rest are "stale-quote snipers." The liquidity provider maintains quotes of one unit at both the bid and the ask at all exchanges. She reacts to the first in a series of orders from the analyst by attempting to cancel her remaining quotes, which she now knows to be mispriced, and she also submits updated quotes. The stale-quote snipers remain inactive until the first order in a series, at which point they free-ride on the information of the analyst by attempting to trade in front of him against the remaining mispriced quotes.

High-frequency traders react as soon as the exchange processes the first in a series of orders sent by the analyst. The processing time of the exchange is δ_E. Furthermore, an infinite number of high-frequency traders react, so one is sure to achieve the minimum communication latency of δ_H at each exchange. Thus, the analyst receives fills for all orders except those that arrive after the first order by $\delta_H + \delta_E$ or more. When the analyst sends orders to all X exchanges, he therefore expects to receive $X - (X-1)e^{-(\delta_H+\delta_E)/\mu_A}$ fills. Furthermore, the analyst's expected profit per fill is $1 - s^*_{LOB}/2$; that is, the size of the jump minus the half spread, which he must pay to the liquidity provider. Equation (2) in the theorem therefore ensures that the analyst chooses research intensity optimally.

As in BCS, free entry into high-frequency trading leads us to focus on equilibria in which the liquidity provider earns zero profits in expectation.[18] Equation (1) in the theorem follows

[18]GETCO (KCG since its merger with Knight Capital Group in 2013) is a representative, significant

from this zero-profit condition. At any instant, one of two things may set off a chain of events that affect her profits: an investor may arrive or the analyst may observe a jump. The arrival rate of investors is λ_{invest}, and conditional on the arrival of an investor, the liquidity provider earns the half-spread, $s^*_{LOB}/2$. On the other hand, the arrival rate of information to the analyst is $r^*_{LOB}\lambda_{jump}$. Because she races against an infinite number of stale-quote snipers, the liquidity provider is never able to cancel her mispriced quotes before they are picked off. Therefore, conditional on information arriving, the liquidity provider loses $1 - s^*_{LOB}/2$ on each exchange. The zero-profit condition of the liquidity provider is therefore

$$\lambda_{invest}\frac{s^*_{LOB}}{2} - r^*_{LOB}\lambda_{jump}X\left(1 - \frac{s^*_{LOB}}{2}\right) = 0. \tag{3}$$

Notice that the equilibrium spread must be such that it balances the revenue from trading with investors against the costs of adverse selection (i.e. trading losses to the analyst and stale-quote snipers). Solving the zero-profit condition (3) for s^*_{LOB} yields equation (1).

4.2 Comparative Statics

This section uses the characterization of equilibrium outcomes given in Theorem 1 to study how these outcomes vary with the parameters of the model. This exercise provides answers to policy-relevant questions such as "What happens when exchanges become faster?", "What happens when traders become faster?", and "What happens when trade becomes fragmented across more exchanges?"

Formally, let S^*_{LOB} and R^*_{LOB} denote the set of equilibrium spreads and research intensities that occur in equilibria of the form described in Theorem 1. The comparative statics of

global player in high-frequency trading and in market making of equities. Moreover, until recently it was the only such firm to be publicly traded and therefore the only such firm for which annual SEC filings are available. Its 2013 Form S-4 filing with the SEC reveals that its net income decreased by 41.9 percent from $232.0 million in 2007 to $167.2 million in 2011 (KCG, 2013, p. 31). For Q2 2013, its market making division even posted a loss of $1.9 million compared to a profit of 9.3 million in the previous year (KCG, 2013, Exhibit 99.2, p. 8). To the extent that excessive profits accrued to high-frequency traders during the previous decade, they were short-lived.

these sets with respect to the parameters of the model are given by the following theorem, and for additional convenience are also summarized in Table 1.

Theorem 2. *Within the set of parameters that satisfy Assumptions 1 and 2, the limit order book equilibrium sets of bid-ask spreads, S^*_{LOB}, and research intensities, R^*_{LOB}, have the following comparative statics (in the strong set order):*

(i) *S^*_{LOB} is nondecreasing in δ_E, nondecreasing in δ_H, nonincreasing in μ_A, nonincreasing in λ_{invest}, nondecreasing in λ_{jump}, and nondecreasing in X.*

(ii) *R^*_{LOB} is nondecreasing in δ_E, nondecreasing in δ_H, nonincreasing in μ_A, nondecreasing in λ_{invest}, and nondecreasing in λ_{jump}.*

Table 1: Summary of predictions of Theorem 2

	δ_E	δ_H	μ_A	λ_{invest}	λ_{jump}	X
S^*_{LOB}	+	+	−	−	+	+
R^*_{LOB}	+	+	−	+	+	

Two particularly interesting sets of comparative statics are those with respect to the latency of exchanges, δ_E, and the minimum latency of high-frequency traders, δ_H. According to the theorem, a decrease in either parameter (i.e. an increase in speed) reduces equilibrium research intensity. Intuitively, a decrease in either parameter leads to more order anticipation, reducing the number of fills that the analyst receives. This reduces the incentive to conduct research, leading to a lower equilibrium research intensity. Additionally, the lower research intensity decreases adverse selection, which allows the liquidity provider to quote a smaller spread. This conclusion is in line with the bulk of the empirical evidence on the relationship between the spread and trading speeds.[19]

[19]Evidence in support of the prediction that improvements in high-frequency trading reduce the spread (or more generally, improve liquidity) is found by Boehmer, Fong, and Wu (2014), Brogaard (2010), Brogaard, Hagströmer, Nordén, and Riordan (2013), Hasbrouck and Saar (2013), Hendershott, Jones, and Menkveld (2011), Malinova, Park, and Riordan (2013), and Menkveld (2013). Evidence in support of the prediction that improvements in exchange speed reduce the spread is found by Easley, Hendershott, and Ramadorai (2014) and Riordan and Storkenmaier (2012).

When the analyst's communication latency becomes more dispersed—that is, when μ_A increases—it becomes harder for the analyst to coordinate the arrivals of his orders. Fewer of the analyst's orders are therefore converted into fills, which disincentivizes research. A lower research intensity decreases the adverse selection faced by the liquidity provider, who quotes a smaller spread in response.

When the arrival rate of investors, λ_{invest}, increases, adverse selection becomes relatively less important, which allows the liquidity provider to quote a smaller spread. The smaller spread also increases the profitability of each of the analyst's trades, which incentivizes higher research intensity.

When the arrival rate of jumps, λ_{jump}, increases, the benefits of research increase: for a fixed level of research, the analyst observes more jumps. This incentivizes a higher research intensity. There is then an increase in the rate of observed jumps, which raises the adverse selection faced by the liquidity provider, who then quotes a larger spread in response.

Finally, another highly relevant set of comparative statics are those with respect to the number of exchanges, X. According to the theorem, an increase in X increases the equilibrium spread but has an ambiguous effect on equilibrium research intensity. Intuitively, the addition of another exchange increases the depth of the aggregate book, since the liquidity provider must offer one share at both the bid and the ask at each exchange in order to serve investors.[20] This therefore increases the number of venues at which an informed trader (either a directly informed analyst or an indirectly informed stale-quote sniper) may trade after observing a jump. The liquidity provider therefore faces more costs from adverse selection, and she must charge a larger spread to compensate. On the other hand, the response of research intensity to an increase in X is theoretically ambiguous, which is a result of two competing effects. The direct effect of an increase in X is to create more opportunities for the analyst to trade on any piece of information, which tends to increase research intensity. However, as previously argued, an increase in X also increases the spread. Larger spreads

[20] That the addition of a trading venue increases the depth of the aggregate book is in line with empirical evidence. For example, see Boehmer and Boehmer (2003), Fink, Fink, and Weston (2006), and Foucault and Menkveld (2008).

make each trade less profitable for the analyst, so the indirect effect of an increase in X tends to reduce research intensity.[21]

4.3 Price Efficiency

Another outcome of interest is price efficiency, or the extent to which prices reflect the value of the underlying asset. Clearly, research intensity is an important determinant of price efficiency, since a jump in the value of the asset can be incorporated into prices only if it is observed. In fact, in the equilibria of the limit order book that are identified in Theorem 1, a jump in the value of the asset is incorporated into prices after a non-infinitesimal amount of time if and only if the jump was observed by the analyst.

Of the reasons the literature has advanced for the social value of price efficiency (*cf.* Appendix C), none are affected in any significant way by price changes at the incredibly small timescales on which communication latency is measured. An apt measure of the aspects of price efficiency that are socially valuable is therefore the probability that a jump is incorporated into prices after a non-infinitesimal amount of time, which is the research intensity. Hence, it immediately follows from Theorem 2 that the socially valuable aspects of price efficiency are negatively affected by improvements in the speed of exchanges and high-frequency traders (i.e. decreases in δ_E and δ_H).

As argued, the socially valuable aspects of price efficiency depend upon price changes on longer timescales, on which price efficiency can be summarized by research intensity. However, when price changes on small timescales are considered, then price efficiency may also depend directly upon other features of the trading environment, such as the speed of exchanges and high-frequency traders. In particular, on very small timescales, improvements in the speed of exchanges or high-frequency traders (i.e. decreases in δ_E or δ_H) have two effects. First is the direct effect. Improvements in speed increase price efficiency, in the sense that jumps, conditional on being observed, are incorporated into prices sooner, which improves price efficiency. Second is the indirect effect, through research intensity. By

[21] Several factors may influence which of the two effects dominates. For example, when δ_H is smaller, the direct effect is smaller and the indirect effect is more likely to dominate.

Theorem 2, improvements in speed reduce the equilibrium research intensity, which harms price efficiency since fewer jumps are observed. The net effect of an improvement of speed on price efficiency is controlled by the relative magnitudes of these two effects. On longer timescales the direct effect is zero, and the indirect effect dominates so that price efficiency is summarized by research intensity. However, it is possible that the direct effect may dominate on very small timescales.

This conclusion—that improvements in speed are harmful to the aspects of price efficiency that are of social value—highlights an omission by many empirical studies on the subject, which have reached the opposite conclusion. For example, Hendershott, Jones, and Menkveld (2011) and Riordan and Storkenmaier (2012) study episodes in which there were improvements in, respectively, the speed of high-frequency traders and the speed of exchanges. They find that price changes become less correlated with trades after the upgrade, which is evidence that liquidity providers become better at adjusting their mispriced quotes before others can trade against them, and therefore that available information is incorporated into prices faster. They then interpret this evidence as indicating that prices are more efficient under the faster speeds that prevail after the upgrade.[22] This evidence is exactly the direct effect predicted by our model. However, it does not account for the indirect channel, which is how speeds affect the amount of information that ultimately becomes available. Since our analysis suggests that the indirect channel dominates—and moreover goes in the opposite direction—it may not be correct to interpret this evidence in the way that they do.

5 Alternative Trading Mechanisms

Dissatisfaction with current outcomes has ignited a wide-ranging policy debate involving industry experts, regulators, and academics. Those involved in this debate have proposed or

[22] In addition, other empirical papers to conclude that faster traders or faster exchanges are beneficial to price efficiency include Carrion (2013), Chaboud, Chiquoine, Hjalmarsson, and Vega (2013), Boehmer, Fong, and Wu (2014), Brogaard (2010), Brogaard, Hendershott, and Riordan (2014), and Hendershott and Moulton (2011).

considered a number of trading mechanisms as alternatives to the limit order book. In this section we evaluate the performance of two alternative trading mechanisms—one new to the literature and one familiar—within the context of our model. First, we propose adding a small delay to the processing of all orders except cancellations. Second, we consider the performance of frequent batch auctions.

5.1 Selective Delay

One alternative trading mechanism is to implement what we refer to as a *selective delay*. Under a selective delay, exchanges process cancellations upon arrival, but process all other order types only after a small delay. This is in contrast to a limit order book, in which orders are processed in the order received. Similar proposals have also been advanced by industry participants.[23] Yet to our knowledge, we are the first to study this mechanism in the literature.

The specific proposal that we consider in this section is the following. All exchanges process cancellations immediately. However, all other order types are processed only after a delay. To have the desired effect, this delay should be small, yet should exceed the maximum difference in reaction time that may occur between two high-frequency traders responding to the same event.[24] In the language of this paper, this corresponds to a delay whose length is an infinitesimal that is infinitely larger than ε. That is, the length of the delay should be "one order of magnitude larger" than communication latency.

The limit order book allows stale-quote snipers to engage in order anticipation, in which they free-ride on the information of the analyst. As shown formally in Corollary 7, this free-riding is a wedge, which prevents the limit order book from implementing an outcome on

[23] Several industry participants have advocated for similar types of delays. For example, Aequitas Innovations, which is planning to enter as a stock exchange serving the Canadian market, is considering a delay of randomized duration of between 3 and 9 milliseconds (Aequitas, 2013). In addition, the incumbent, TMX Group, has recently announced similar plans for one of their platforms, the Alpha Exchange. They are considering a delay of randomized duration of between 5 and 25 milliseconds (Alpha Exchange, 2014). Finally, in an open letter to the SEC, Peterffy (2014) advocates for a delay of randomized duration of between 10 and 200 milliseconds. While all these proposals advocate for randomization in the delay as an additional means of blunting the advantages of speed, randomization does not lead to additional benefits in our model, and a deterministic duration suffices.

[24] Budish, Cramton, and Shim (2014) indicate that this may be about 100 microseconds in practice.

the frontier of the tradeoff between spreads and price efficiency. However, a selective delay eliminates this free-riding by allowing the liquidity provider's cancellations to be processed before any stale-quote snipers can successfully trade against a mispriced quote. In doing so, the selective delay mechanism implements an outcome on the frontier of this tradeoff.

Theorem 3 characterizes the outcomes that arise in equilibrium under a selective delay whose length is an infinitesimal infinitely larger than ε. As before, the theorem focuses on two outcome variables: (i) the bid-ask spread, s^*_{SD}, and (ii) the research intensity, r^*_{SD}.

Theorem 3. *Under Assumptions 1 and 2, there exists a Nash equilibrium of the selective delay mechanism in which the spread, s^*_{SD}, and research intensity, r^*_{SD}, are constants given by any solution to*

$$s^*_{SD} = \frac{2r^*_{SD}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + r^*_{SD}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]} \tag{4}$$

$$r^*_{SD} \in \arg\max_{r\in[0,1]} \left\{ r\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]\left(1 - \frac{s^*_{SD}}{2}\right) - c(r)\right\} \tag{5}$$

The strategies that support these outcomes in Nash equilibrium are roughly as follows. As in the limit order book, investors submit orders to buy or sell according to their trading desires, and the analyst submits orders to buy or sell according to the directions of the jumps he observes. Also as before, one high-frequency trader plays the role of liquidity provider, and reacts to the first in a series of orders from the analyst by attempting to cancel her remaining mispriced quotes. In contrast to the limit order book, there are no stale-quote snipers. This is because the selective delay eliminates the possibility that a high-frequency trader could snipe a mispriced quote before it is cancelled by the liquidity provider.

The liquidity provider reacts as soon as the exchange processes the first in a series of orders sent by the analyst. The processing time of the exchange is δ_E, and the minimum communication latency of the liquidity provider is δ_H. Thus, the analyst receives fills for all orders that arrive within $\delta_H + \delta_E$ of the first order. For orders that arrive after that, their probability of being filled is determined by the dispersion of the analyst's communication

latency relative to that of the liquidity provider. When the analyst sends orders to all X exchanges, he therefore expects to receive $X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}$ fills. Furthermore, the analyst's expected profit per fill is $1 - s^*_{SD}/2$. Equation (2) in the theorem therefore ensures that the analyst chooses research intensity optimally.

As in the limit order book, the equilibrium is one in which the liquidity provider earns zero profits. The liquidity provider's revenue from investors is $\lambda_{invest} s^*_{SD}/2$. These must be balanced by the costs of adverse selection. Since there are no stale-quote snipers in this equilibrium, the liquidity provider faces adverse selection only from the analyst. Her zero-profit condition is therefore

$$\lambda_{invest}\frac{s^*_{SD}}{2} - r^*_{SD}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]\left(1 - \frac{s^*_{SD}}{2}\right) = 0, \tag{6}$$

which yields equation (4) in the theorem.

As shown in Section 6, a selective delay implements a point on the frontier of the tradeoff between the spread and price efficiency. As an aside, variants of the selective delay mechanism also perform well in the BCS model. In that model, if the delay is positive but less than a certain threshold, then a selective delay implements the equilibrium spread that they achieve with "short" batch intervals.[25] Furthermore, if the delay exceeds that threshold, then a selective delay implements the equilibrium spread that they achieve with "long" batch intervals.

[25] In the language of their model, the threshold is $\delta_{slow} - \delta_{fast}$. A subtlety is that for the purposes of that model, the delay should be applied only to orders that would trigger an immediate trade (rather than for all non-cancellation orders). This is to eliminate the possibility of an investor arriving before the liquidity provider can post new quotes. Note that this possibility is not a concern in this paper because the delay is only for an infinitesimal amount of time. Furthermore, because it would in practice be simpler to condition the delay on the order type rather than on both the order type and the state of the book in conjunction, we would in practice advocate for the proposal described in the main text: delaying all non-cancellation orders.

5.2 Frequent Batch Auctions

Frequent batch auctions are a policy intervention that has recently received a great deal of attention, notably from BCS, and also from several others.[26] Frequent batch auctions are uniform-price sealed-bid double auctions conducted repeatedly at discrete time intervals.[27] Batch auctions differ from the limit order book along several dimensions, most notably in that they break the continuous nature of trading.

In this section, we consider the batch auction design most closely in line with the BCS proposal. They advocate a batch length that is "long" relative to communication latency. The natural analogue of this in the language of this paper is a batch length, which is an infinitesimal that is infinitely larger than ε. Additionally, we focus on batch auctions that are synchronized across exchanges, which they also identify as an attractive property.[28]

Like a selective delay, frequent batch auctions also eliminate the informational free-riding that stale-quote snipers do when they engage in order anticipation. They thereby also implement an outcome on the frontier of the tradeoff between spreads and price efficiency. However, batch auctions of this nature also prevent the liquidity provider from engaging in order anticipation; that is, they preventing her from canceling mispriced quotes. The reason is that with batch auctions, the first of the analyst's orders to arrive generates a trade not immediately, but only at the end of the current batch interval. With batch lengths that are infinitely larger than ε, this allows the analyst's orders to arrive at all exchanges before any trade occurs. There is therefore no scope for any high-frequency traders to react, which allows the analyst to convert all his orders into fills. As shown formally in Theorem 5, this incentivizes a higher research intensity than the other trading mechanisms. Furthermore, the higher research intensity increases the adverse selection faced by the liquidity provider,

[26] Other papers that promote frequent batch auctions include Madhavan (1992) and Wah and Wellman (2013). Additionally, batch auctions have received mention from policymakers in, for example, SEC (2010), Foresight (2012), Schneiderman (2014), and White (2014).

[27] For a more detailed exposition of the batch auction design, see Section 7.1 of BCS. The same authors discuss implementation details in Budish, Cramton, and Shim (2014).

[28] "[T]he following [is an] attractive property: traders [...] have information about the time $t-1$ auction outcomes from *all* locations (e.g. Chicago, New York, London, Tokyo), and have information about the time t auction outcomes from *no* locations" (Budish, Cramton, and Shim, 2014).

who responds by setting a larger spread than under the other mechanisms.

Theorem 4 characterizes the outcomes that arise in equilibrium under frequent batch auctions that are synchronized across exchanges and have a batch length that is an infinitesimal infinitely larger than ε. As before, the theorem focuses on two outcome variables: (i) the bid-ask spread, s^*_{FBA}, and (ii) the research intensity, r^*_{FBA}.

Theorem 4. *Under Assumptions 1 and 2, there exists a Nash equilibrium of the frequent batch auction mechanism in which the spread, s^*_{FBA}, and research intensity, r^*_{FBA}, are constants given by any solution to*

$$s^*_{FBA} = \frac{2r^*_{FBA}\lambda_{jump}X}{\lambda_{invest} + r^*_{FBA}\lambda_{jump}X} \tag{7}$$

$$r^*_{FBA} \in \arg\max_{r\in[0,1]} \left\{ r\lambda_{jump}X\left(1 - \frac{s^*_{FBA}}{2}\right) - c(r) \right\} \tag{8}$$

The strategies that support these outcomes in Nash equilibrium are roughly as follows. As with the limit order book, investors submit orders to buy or sell according to their trading desires, and the analyst submits orders to buy or sell according to the directions of the jumps he observes.

As with the limit order book, equation (8) in the theorem ensures that the analyst chooses research intensity optimally. Also as before, the equilibrium is one in which the liquidity provider earns zero profits. As with the limit order book, the liquidity provider is never able to cancel her mispriced quotes before they are picked off. The zero-profit condition of the liquidity provider is therefore again

$$\lambda_{invest}\frac{s^*_{FBA}}{2} - r^*_{FBA}\lambda_{jump}X\left(1 - \frac{s^*_{FBA}}{2}\right) = 0,$$

which yields equation (7) in the theorem.

It should be noted that in this model, the outcomes prevailing with frequent batch auctions could also be implemented in a less intrusive way with a "universal delay." That is, by requiring exchanges to wait for an interval before announcing their trades but otherwise

maintaining the limit order book mechanism. Formally, if δ_E were an infinitesimal that is infinitely larger than ε, then inspection of Theorem 1 reveals that the limit order book would deliver the batch auction equilibrium outcome identified in Theorem 4.

Moreover, there are several reasons to think that this alternative would be preferable to frequent batch auctions. First, by virtue of being so near the status quo, it would be easier to implement and therefore also less likely to suffer from glitches, loopholes, or other complications.[29] Second, there are some legal questions pertaining to whether it is possible for frequent batch auctions to operate simultaneously on multiple exchanges in a way that satisfies laws as they are currently written, particularly Regulation NMS in the United States. Third, frequent batch auctions require synchronization across exchanges, which might be difficult to implement in practice since exchanges are competitors. On the other hand, a delay in the processing of orders could be implemented in a decentralized way.

5.3 Comparison of Equilibria

We have so far characterized the outcomes that prevail in equilibrium under three trading mechanisms: (*i*) the limit order book, (*ii*) a selective delay, and (*iii*) frequent batch auctions. This section compares these outcomes. We find that a selective delay results in a higher research intensity than the limit order book. Additionally, frequent batch auctions result in the largest spread and the highest research intensity of the three mechanisms.

Formally, let S^*_{SD} and R^*_{SD} denote the set of equilibrium spreads and research intensities that occur in a equilibria of the form described in Theorem 3. Similarly, let S^*_{FBA} and R^*_{FBA} denote the set of equilibrium spreads and research intensities that occur in equilibria of the form described in Theorem 4. The way in which these sets compare to each other and to the corresponding sets for the limit order book are as described in the following result.

Theorem 5. *Under Assumptions 1 and 2, the equilibrium spread and research intensity*

[29] Delaying the announcement of trades would be quite easy to implement by, for example, forcing announcements to travel through additional lengths of fiber-optic cable. A similar scheme is already used in practice by the alternative trading system IEX, which implements a 350-microsecond delay by simply forcing all incoming orders to travel through 38 miles of coiled cable before proceeding to their matching engine (IEX Group, 2014).

prevailing under the limit order book, selective delay, and frequent batch auction mechanisms can be ranked in the following way (in the strong set order):

$$S^*_{FBA} \geq S^*_{SD} \cup S^*_{LOB}$$

$$R^*_{FBA} \geq R^*_{SD} \geq R^*_{LOB}$$

The intuition for the equilibrium research intensity being higher under a selective delay than under the limit order book is as follows. In the limit order book, stale-quote snipers free-ride on the information of the analyst. However, a selective delay eliminates this free-riding and allows the analyst to capture more of the gains from information, which incentivizes a higher research intensity. However, the ranking of equilibrium spreads under these two trading mechanisms is ambiguous. This is due to two competing channels. A selective delay may reduce adverse selection by eliminating stale-quote sniping, but on the other hand it may increase adverse selection by raising the intensity of the analyst's research.

Intuition for the ordering among the equilibrium outcomes of frequent batch auctions and the other mechanisms is as follows. Frequent batch auctions enable the analyst to convert all his orders into fills. They therefore maximize the gains from research and induce the highest level of research intensity. However, frequent batch auctions also maximize the adverse selection faced by the liquidity provider: research intensity—therefore also the arrival rate of informed orders—is at its highest, and in addition she is never able to cancel a mispriced quote. The liquidity provider therefore responds by quoting the largest spread.

An interesting contrast between our findings and those of BCS is the following. In their model, frequent batch auctions reduce the spread, and in fact batch auctions with "long" batch intervals implement a spread of zero. However, in our model frequent batch auctions implement the *largest* spread. The crucial difference between the two models, which generates this divergence, is the source of information. In BCS, information is exogenous and publicly revealed to all traders. Batching then helps a liquidity provider adjust her mispriced quotes before they are picked off. On the contrary, in the model of this paper,

information is endogenous and privately revealed to the analyst. Batching then has the opposite effect: it prevents a liquidity provider from inferring the the analyst's information and adjusting her mispriced quotes before they are picked off.

6 Optimal Outcomes

In this section, we characterize a "possibilities frontier" of feasible outcomes. In effect, outcomes are evaluated according to two criteria: the bid-ask spread and price efficiency. There exists a tradeoff between these two criteria. We characterize the frontier of this tradeoff by formulating and solving a social planner's problem. The selective delay and frequent batch auction outcomes lie on this frontier. The limit order book outcome in general does not.

6.1 Performance Criteria

We use two criteria to evaluate the performance of a trading mechanism: the utility of investors and price efficiency. There are, of course, other criteria that could be used to evaluate performance, for example those that also involve the utility of high-frequency traders and the analyst. However, we choose to focus on investor utility and price efficiency because they are the aspects of our model that tie most closely to the stated objectives of most regulatory bodies.[30]

Investor utility. Investor protection is the primary stated goal of the main regulatory bodies of these markets. For example, the SEC states that its mission is "to protect investors, maintain fair, orderly, and efficient markets, and facilitate capital formation," and stresses that "as more and more first-time investors turn to the markets to help secure their futures, pay for homes, and send children to college, our investor protection mission is more

[30]For example, the SEC has affirmed, "Where the interests of long-term investors and short-term professional traders diverge, the Commission repeatedly has emphasized that its duty is to uphold the interests of long-term investors" (SEC, 2010). Nevertheless, many of our results about investor utility also pertain to total utility.

compelling than ever" (SEC, 2013). For posted-price mechanisms, investor utility and the bid-ask spread are equivalent criteria in the model, since they are related through $U = \theta - \frac{s}{2}$.

Price efficiency. Another criterion that is often used to gauge the performance of a market is price efficiency, or the extent to which market prices reflect the fundamental value of the underlying assets. For example, Fama (1970) states, "the ideal is a market in which prices ... at any time 'fully reflect' all available information." Although price efficiency is not innately valuable to the traders in our model, the literature has identified a number of channels through which price efficiency is a positive externality of financial exchanges, valuable to non-traders. For example, higher price efficiency may enable a firm's board of directors to provide managers with better incentives and thereby ameliorate an agency problem. Additionally, higher price efficiency may provide better feedback to firm managers, enabling them to make better decisions. A more detailed discussion of the benefits of price efficiency is given in Appendix C.

6.2 Social Planner Problem

We now outline the social planner's problem, which is stated formally in the next section. Roughly speaking, the social planner evaluates outcomes according to the criteria discussed in Section 6.1: investor utility and price efficiency. While there are no prices in this abstract formulation, price efficiency is measured by the analyst's observation of jumps.

The social planner is allowed to allocate resources (shares of the asset and dollars) among the agents, and is also allowed to recommend a research intensity to the analyst. However, there are informational constraints, since the planner cannot observe the analyst's research intensity. Thus, to incentivize the analyst, the planner must make the allocation a function of the analyst's reports.

This framework allows the social planner to optimize over a wide range of trading mechanisms, including a limit order book, a selective delay, and frequent batch auctions. It also allows for other possibilities, such as "infrequent batch auctions," in which trade is

allowed to occur only a small number of times.

To be more precise, the social planner's objective includes two criteria: (*i*) the expected value of aggregate investor utility, and (*ii*) the expected number of jumps that are unobserved by the analyst. The social planner maximizes a welfare function that puts weight $\alpha \in [0,1]$ on the first criterion (which enters positively) and weight $1-\alpha$ on the second criterion (which enters negatively).

In addition, the social planner faces several constraints: (*i*) budget balance; (*ii*) ex-ante individual rationality for investors; (*iii*) ex-ante individual rationality for high-frequency traders; (*iv*) time T individual rationality for the analyst: conditional on any history, the expected value of the analyst's portfolio at time T, just before v_T is announced, must be nonnegative;[31] (*v*) effort choice incentive compatibility for the analyst: the analyst must always find it optimal to follow the instructions of the social planner with respect to research intensity; and (*vi*) truthful reporting incentive compatibility for the analyst: the analyst must always find it optimal to report the jumps he observes.

6.3 Formalities

For readers who are interested in the mathematical details, a formal statement of the social planner's problem is contained in this section. The next section characterizes properties of solutions to the problem.

We first establish some notation, which is needed to formalize the social planner's problem. Let J_t^+ and J_t^- denote, respectively, the set of times in the interval $[0,t]$ at which the analyst observes upward and downward jumps. Let $J_t = (J_t^+, J_t^-)$ and $\tilde{v}_t = v_0 + |J_t^+| - |J_t^-|$. Let $\hat{J}_T^+$ and $\hat{J}_T^-$ denote, respectively, the set of times in the interval $[0,T]$ at which the analyst reports observing upward and downward jumps. By the "no market manipulation"

[31] We impose time T individual rationality for the analyst, as opposed to ex-ante individual rationality, to prevent the social planner from levying a lump sum tax on the analyst equal to his expected profits. We do so because only realized capital gains are taxed in practice. Note that the analyst's portfolio may contain "lottery tickets" that pay off after v_T is announced, and we do not require the value of the analyst's portfolio to be nonnegative after resolution of the lottery. Therefore, this constraint should not be interpreted as an inability of the analyst to take on debt (i.e. limited liability), but instead as an inability of the social planner to tax the analyst on gains that he may not earn.

assumption, the analyst is restricted to reports of the form $\hat{J}_T^+ \subseteq J_T^+$ and $\hat{J}_T^- \subseteq J_T^-$.[32] Let $\hat{J}_T = (\hat{J}_T^+, \hat{J}_T^-)$ and $\hat{v}_T = v_0 + |\hat{J}_T^+| - |\hat{J}_T^-|$. In addition, let I_t^B and I_t^S denote, respectively, the set of times in the interval $[0, t]$ at which an investor arrives with a desire to buy or to sell. Let $I_t = (I_t^B, I_t^S)$.

We consider contracts between the social planner, the analyst, investors, and high-frequency traders. These contracts specify a research intensity for the analyst as a function of time. In addition, they may specify payments, as functions of $\hat{J}_T$ and I_T to the analyst, the investors who arrive, and high-frequency traders. These payments may be in the form of either shares of the security, dollars, or both. Because agents do not discount, it is without loss of generality to think of these payments as occurring at time T (i.e. just before v_T is announced). At this point in time, the expected value of a share of the security, conditional on all available information, is $\tilde{v}_T$.

Let $x_A(\hat{J}_T, I_T)$ and $y_A(\hat{J}_T, I_T)$ denote, respectively, the number of shares and the number of dollars paid to the analyst if $\hat{J}_T$ is reported and investors arrive according to I_T. Let $x_{B,t}(\hat{J}_T, I_T)$ and $y_{B,t}(\hat{J}_T, I_T)$ denote the corresponding quantities for an investor who arrives at time t with a desire to buy. Similarly, let $x_{S,t}(\hat{J}_T, I_T)$ and $y_{S,t}(\hat{J}_T, I_T)$ denote the corresponding quantities for an investor who arrives at time t with a desire to sell. Finally, let $x_{H,h}(\hat{J}_T, I_T)$ and $y_{H,h}(\hat{J}_T, I_T)$ denote the corresponding quantities for the hth high-frequency trader.

A formal statement of the social planner's problem is then

$$\max_{\substack{r, x_A, y_A, x_{H,h}, y_{H,h} \\ x_{B,t}, y_{B,t}, x_{S,t}, y_{S,t}}} \Bigg\{ \alpha \Bigg(\frac{\lambda_{invest}}{2} \int_0^T \mathbb{E}_r \left[x_{B,t}(J_T, I_T)\tilde{v}_T + y_{B,t}(J_T, I_T) + \theta \mathbb{1}\{x_{B,t}(J_T, I_T) = 1\} \right] dt + \frac{\lambda_{invest}}{2} \int_0^T \mathbb{E}_r \left[x_{S,t}(J_T, I_T)\tilde{v}_T + y_{S,t}(J_T, I_T) + \theta \mathbb{1}\{x_{S,t}(J_T, I_T) = -1\} \right] dt \Bigg) - (1 - \alpha) \int_0^T \lambda_{jump}[1 - r(t)]\, dt \Bigg\}$$

subject to the constraints

[32] As before, this restriction may be thought of as coming from optimal behavior on the part of the analyst if at some point in the future he would be audited and punished for making false reports (see Footnote 11).

(BB–1) $\forall J_T\ \forall I_T,\ x_A(J_T, I_T) + \sum_{t \in I_T^B} x_{B,t}(J_T, I_T) + \sum_{t \in I_T^S} x_{S,t}(J_T, I_T) + \sum_{h=1}^{\infty} x_{H,h}(J_T, I_T) = 0$

(BB–2) $\forall J_T\ \forall I_T,\ y_A(J_T, I_T) + \sum_{t \in I_T^B} y_{B,t}(J_T, I_T) + \sum_{t \in I_T^S} y_{S,t}(J_T, I_T) + \sum_{h=1}^{\infty} y_{H,h}(J_T, I_T) = 0$

(IR–B) $\forall t,\ \mathbb{E}_r\left[x_{B,t}(J_T, I_T)\tilde{v}_T + y_{B,t}(J_T, I_T) + \theta \mathbb{1}\{x_{B,t}(J_T, I_T) = 1\}\right] \geq 0,$

(IR–S) $\forall t,\ \mathbb{E}_r\left[x_{S,t}(J_T, I_T)\tilde{v}_T + y_{S,t}(J_T, I_T) + \theta \mathbb{1}\{x_{S,t}(J_T, I_T) = -1\}\right] \geq 0,$

(IR–H) $\forall h,\ \mathbb{E}_r\left[x_{H,h}(J_T, I_T)\tilde{v}_T + y_{H,h}(J_T, I_T)\right] \geq 0,$

(IR–A) $\forall J_T\ \forall I_T,\ x_A(J_T, I_T)\tilde{v}_T + y_A(J_T, I_T) \geq 0,$

(IC–1) at all $t \in [0, T]$ and conditional on any J_t and any I_t, the analyst finds it optimal to conduct research with intensity $r(t)$,

(IC–2) $\forall J_T\ \forall I_T$, the analyst finds it optimal to report $\hat{J}_T = J_T$.

6.4 Outcome Frontier

We now characterize properties of solutions to the social planner's problem. This characterization is useful because it enables an evaluation of how various trading mechanisms perform within the tradeoff between investor utility and price efficiency. We find that the selective delay and frequent batch auction equilibria are on the frontier of this tradeoff. However, the limit order book outcome in general is not.

The *outcome* of a contract consists of two quantities, which are rescalings of the two quantities that enter the social planner's objective: (*i*) average expected utility of investors, denoted by U, and (*ii*) average research intensity, denoted by r. The *frontier* is the union of outcomes of contracts that solve the social planner's problem for some $\alpha \in [0, 1]$. This section characterizes the frontier.

To obtain a clean characterization of the frontier, we require some additional regularity conditions on the cost of research function. We assume continuous differentiability, strict

convexity, and convexity of $rc'(r)$. The latter is used to ensure that the social planner prefers to implement a research intensity that is a constant function of time.[33]

Assumption 3. $c(r)$ is continuously differentiable, strictly convex, and $rc'(r)$ is convex.

We now state the main result of this section, Theorem 6, which provides an elegant characterization of the frontier. Given the complexity of the social planner's problem, the simplicity of this characterization is somewhat striking. Intuitively, the constraints of the problem can be combined in such a way that they distill to a single tradeoff: the analyst can be incentivized to conduct research, but only if he is paid with resources taken from investors. It is this tradeoff that determines the frontier.

Theorem 6. *Let* $\bar{r} = \max\{r \in [0,1] : rc'(r) \leq \theta\lambda_{invest}\}$. *Under Assumption 3, an outcome* (r, U) *is on the frontier if and only if research intensity is a constant* $r \in [0, \bar{r}]$ *and the average expected utility of investors is* $U = \theta - \frac{rc'(r)}{\lambda_{invest}}$.

In words, an outcome is on the frontier if research intensity is not too high, and if investor utility is related to research intensity in a particular way. This relationship is downward sloping, so that there exists a tradeoff between these two quantities. Intuitively, the incentives required to implement a high research intensity are costly and must be funded through a "tax" on investors. Furthermore, there is a limit to the amount that the social planner can tax investors, which implies an upper bound on the research intensity that can be incentivized.

While a complete proof is in Appendix A, an outline of the argument is as follows. We begin by relaxing several constraints. We then demonstrate that the frontier of the relaxed problem is as described in the statement of the theorem. A key step in this process uses a result from Holmström and Milgrom (1987) to characterize the (IC–1) constraint in such a way that the dynamic problem collapses to a static one. Finally, we verify that this is the frontier of the original problem by identifying, for any point described in the statement of

[33] This condition is relatively mild and would be satisfied by most natural parametrizations of the cost function. Also note that the condition $c'''(r) \geq 0$ is sufficient to ensure that it holds.

the theorem, a contract that both implements that point and satisfies all the constraints of the original problem.

6.5 Trading Mechanisms and the Frontier

In this section, we study whether the previously considered trading mechanisms implement outcomes on the frontier. We find that the selective delay and frequent batch auction outcomes lie on the frontier. However, the limit order book outcome in general does not.

Corollary 7 formalizes these statements. Parts (*i*) and (*ii*) of the corollary state that both a selective delay and frequent batch auctions implement outcomes on the frontier. Part (*iii*) states that the limit order book does not, unless there is only a single exchange or exchanges are very slow relative to communication latency. These results depend upon the following assumption, which roughly states that it would be infinitely costly for the analyst to ensure observation of all jumps.

Assumption 4. $\lim_{r \to 1} c(r) = \infty$.

Corollary 7. *Under Assumptions 1, 2, 3, and 4,*

(i) the selective delay mechanism implements an outcome on the frontier,

(ii) the frequent batch auction mechanism implements an outcome on the frontier, and

(iii) the limit order book mechanism implements an outcome on the frontier if and only if either $X = 1$ *or* δ_E *is infinitely larger than* ε.

We prove these results by comparing the equilibria characterized by Theorems 1, 3 and 4 against the frontier characterized by Theorem 6.

When there are multiple exchanges and δ_E is not infinitely larger than ε, the limit order book equilibrium involves stale-quote snipers free-riding on the information of the analyst. This free-riding contributes to adverse selection, which raises the spread, without providing incentives to conduct research. This wedge therefore prevents the limit order book equilibrium from being on the frontier. However, when there is a single exchange or when

δ_E is infinitely larger than ε, there is no free-riding (in fact there is no order anticipation at all) and the limit order book equilibrium does lie on the frontier. Moreover, there is never free-riding with either a selective delay or with frequent batch auctions, and the equilibria of both mechanisms lie on the frontier.

The role of Assumption 4 is to eliminate the possibility that the analyst is at a corner solution, choosing to conduct research with the maximal intensity. If the analyst were at such a corner solution, selective delay and frequent batch auctions might implement outcomes off the frontier because the analyst's rents could be reduced without reducing his choice of research intensity. With the assumption, the analyst cannot be at such a corner solution, which ensures that selective delay and frequent batch auctions implement points on the frontier.

Furthermore, if a policymaker is free to adjust the number of exchanges in operation, then both a selective delay and frequent batch auctions can be used to implement much of the frontier. Corollary 8 demonstrates that, by adjusting the number of exchanges, either a selective delay or frequent batch auctions can be used to implement all points on the frontier except, perhaps, for a region involving very high research intensities. For the purposes of this result, we do not restrict the number of exchanges to the integers. If, however, we did impose this restriction, then these trading mechanisms could still be used to implement a rich set of points on the frontier, especially if the social planner were also free to adjust other parameters of the model, such as the speed of exchanges, δ_E. This result can be interpreted as a partial converse to the first two parts of Corollary 7.

Corollary 8. *Let* $\bar{r}' = \max\left\{r \in [0,1] : rc'(r) \leq \lambda_{invest} \min\left\{\frac{1}{2}, \theta\right\}\right\}$. *If* (r,U) *is an outcome on the frontier in which* $r \in [0,\bar{r}']$, *then under Assumptions 3 and 4,*

(i) there exists $X' \geq 0$ *such that* (r,U) *is the equilibrium outcome of the selective delay mechanism with* X' *exchanges, and*

(ii) there exists $X'' \geq 0$ *such that* (r,U) *is the equilibrium outcome of the frequent batch auction mechanism with* X'' *exchanges.*

Furthermore, it can be shown that the r' defined in Corollary 8 is at least r^*_{LOB}. As a consequence, for either selective delay or frequent batch auctions, there exists some number of exchanges under which the mechanism would implement an outcome on the frontier dominating the limit order book outcome along both dimensions (i.e. featuring both a higher research intensity and higher investor utility).

It should be noted that in our formulation investors possess no demand for immediacy. That is, they receive a utility bonus of θ if their desire to transact is satisfied at any point before time T, regardless of the exact time at which that occurs. In practice, investors are likely to prefer transacting sooner. However, a delay cost would not substantially detract from the performance of either a selective delay or frequent batch auctions because they both allow investors to transact with only infinitesimal delays. On the other hand, a delay cost could substantially detract from the performance of other types of trading mechanisms, such as infrequent batch auctions. Thus, incorporating delay costs into the model would only strengthen the result about the desirability of the selective delay and frequent batch auction mechanisms.

6.6 Comparative Statics

The frontier consists of the feasible outcomes that are optimal for some weighting of the objectives. This section describes how the optimal point on the frontier varies with the parameters of the model, as well as with the social planner's weights.

Formally, let U^*_{OPT} and R^*_{OPT} denote the set of average expected utilities of investors and the set of average research intensities that prevail under solutions to the social planner's problem. The comparative statics of these sets are given by the following result, and for additional convenience are also summarized in Table 2.

Proposition 9. *Under Assumption 3, the sets of optimal average investor utility, U^*_{OPT}, and average research intensity, R^*_{OPT} have the following comparative statics (in the strong set order):*

(i) U^*_{OPT} *is nondecreasing in* α*, nondecreasing in* λ_{invest}*, nonincreasing in* λ_{jump}*, and nondecreasing in* θ*; and*

(ii) R^*_{OPT} *is nonincreasing in* α*, nondecreasing in* λ_{invest}*, nondecreasing in* λ_{jump}*, and nondecreasing in* θ*.*

Table 2: Summary of predictions of Proposition 9

	α	λ_{invest}	λ_{jump}	θ
U^*_{OPT}	+	+	−	+
R^*_{OPT}	−	+	+	+

As α increases, the social planner places more weight on investor utility and less weight on price efficiency (or equivalently, on research). The optimal point on the frontier therefore shifts accordingly. An increase in λ_{jump} raises the marginal return of research, since there are more jumps to be observed. The social planner responds by increasing research intensity. This must also reduce investor utility, since funding of the additional research requires investors to be "taxed" at a higher level. Finally, an increase in either λ_{invest} or θ raises the total gains from trade available in the economy, which allows both investor utility and research intensity to rise.

Proposition 9 indicates that one trading mechanism may not be appropriate for all securities. For example, it may be desirable to tailor trading mechanisms differently for securities for which the benefits of price efficiency (*cf.* Appendix C) are small or large (i.e. high or low α). Similarly, it may be desirable to tailor trading mechanisms differently for thickly and thinly traded securities (i.e. high and low λ_{invest}) or for volatile and nonvolatile securities (i.e. high and low λ_{jump}).

7 Conclusion

Each month, financial markets facilitate trillions of dollars of transactions. Even small changes in technology or the trading mechanism may therefore have considerable implica-

tions, and as such deserve careful analysis. The model presented in this paper provides a framework that enables an evaluation of both the consequences of recent changes—such as increase in speed by both exchanges and traders—and the effects of several regulatory shifts currently being debated.

The model predicts two main consequences of an increase in speed by either traders or exchanges: lower spreads and lower price efficiency. The analysis of alternative trading mechanisms focuses on two specific proposals: the addition of a selective delay, applied to all orders except cancellations, and the implementation of frequent batch auctions. We find that both alternatives implement equilibria that lie on the frontier of the tradeoff between spreads and price efficiency. In that sense, they represent an improvement over the limit order book mechanism, which, despite being the current industry practice, does not generally implement an equilibrium on this frontier.

A Proofs

Proof of Theorem 1. The proof proceeds in three parts. First, we argue that there exists at least one solution to (1) and (2). Second, we describe the equilibrium strategies of the traders. Third, we verify that no trader has a profitable deviation.

Part One (Existence): R^*_{LOB}, the set of research intensities that appear in any solution to (1) and (2), is the intersection of $R_{LOB}(\hat{r})$, given below, with the forty-five degree line:

$$R_{LOB}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X-(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest}+\hat{r}\lambda_{jump}X} - c(r) \right\}$$

By definition, $\min R_{LOB}(0) \geq 0$ and $\max R_{LOB}(1) \leq 1$. By the maximum theorem (Berge, 1963), $R_{LOB}(\hat{r})$ is upper hemicontinuous. Thus, the intermediate value theorem implies that it intersects the forty-five degree line at some point, and therefore R^*_{LOB} is nonempty. Finally, S^*_{LOB}, the set of spreads that appear in any solution to (1) and (2), can be easily derived from R^*_{LOB} and (1).

Part Two (Description): One high-frequency trader plays the role of a "liquidity provider." The remaining high-frequency traders play the role of a "stale-quote sniper."

The strategy of the liquidity provider is as follows. At time zero, she submits to each exchange a limit order to buy one share at the bid $b_0 = v_0 - \frac{s^*_{LOB}}{2}$ and a limit order to sell one share at the ask $a_0 = v_0 + \frac{s^*_{LOB}}{2}$. If at any time t one of her standing limit orders is filled by an investor, then she immediately submits an identical order to replace it. If at any time t one of her standing buy (sell) orders is filled by the analyst, and if the analyst's last trade was an non-infinitesimal length of time in the past, then she immediately submits to each exchange (*i*) cancellations for her remaining limit orders, (*ii*) a limit order to buy one share at $b_{t^+} = b_{t^-} - 1$ ($b_{t^+} = b_{t^-} + 1$), and (*iii*) a limit order to sell one share at $a_{t^+} = a_{t^-} - 1$ ($a_{t^+} = a_{t^-} + 1$).[34]

The strategy of a stale-quote sniper is as follows. If at any time t, a standing buy (sell) order is filled by the analyst at a particular exchange, and if the analyst's last trade was an non-infinitesimal length of time in the past, then she immediately submits to the other exchanges an IOC order to buy (sell) at the price $m_{t^-} + \frac{s^*_{LOB}}{2}$ ($m_{t^-} - \frac{s^*_{LOB}}{2}$), where $m_t = \frac{b_t+a_t}{2}$.

The strategy of an investor who arrives at time t with a private motive to buy (sell) is to submit immediately an IOC order to buy (sell) one share at the price $m_t + \frac{s^*_{LOB}}{2}$ ($m_t - \frac{s^*_{LOB}}{2}$).

The strategy of the analyst is as follows. He conducts research with intensity r^*_{LOB} at all times. If at any time t he observes an upward (downward) jump in the value of the asset, then he immediately submits to each exchange IOC orders to buy (sell) one share at the price $m_{t^-} + \frac{s^*_{LOB}}{2}$ ($m_{t^-} - \frac{s^*_{LOB}}{2}$). He submits no orders otherwise.

Part Three (Verification): We now argue that the investors do not have profitable deviations. Given the behavior of the other traders, the expected difference, conditional on any

[34] For any continuous time variable X_t, we use the shorthand X_{t^+} to denote $\lim_{s\to t^+} X_s$ and X_{t^-} to denote $\lim_{s\to t^-} X_s$.

investor's information, between the liquidation value of a share, v_T, and the midprice at any exchange is zero at all times. Moreover, from Assumption 1, the half-spread $\frac{s^*_{LOB}}{2}$ does not exceed θ. Since investors receive a utility benefit of θ from trading, it is optimal for them to trade. Additionally, since the half-spread is stationary and investors are risk-neutral, they also do not have an incentive to delay trading.

We now argue that the liquidity provider does not have a profitable deviation. The arguments are similar to those in Budish, Cramton, and Shim (2013, Proof of Proposition 1). As argued in Section 4.1, she earns zero profits in the equilibrium. It therefore remains to show that she does not possess a deviation that would yield positive profits. It is not profitable to deviate by quoting a larger spread, since, because of the limit prices specified by the other traders, she would then never participate in any trades. It is also not profitable to deviate by quoting a smaller spread, since that would result in negative expected profits. Finally, it is also not profitable to deviate by quoting more than a single unit at either the bid or the ask, since her benefits would be the same (only one unit at each is needed to satisfy investor demand) but her adverse selection costs would increase (since more units are exposed to adverse selection from the analyst and stale-quote snipers).

We now argue that the stale-quote snipers do not have profitable deviations. The arguments are also similar to those in Budish, Cramton, and Shim (2013, Proof of Proposition 1). An infinite number of snipers compete with each other for the opportunity to trade against the same mispriced quotes. Each individual sniper therefore earns zero profits. It therefore remains to show that none of them possesses a deviation that would yield positive profits. It is not profitable to attempt to provide liquidity at a larger spread than the liquidity provider, since these orders would never be filled. It is also not profitable to attempt to provide liquidity at a smaller spread than the liquidity provider, since that would result in negative expected profits. It is also not profitable to attempt to provide liquidity at the same spread as the liquidity provider, since these quotes have the same adverse selection costs (from analyst and stale-quote sniper orders) that the liquidity provider faces in equilibrium but only half the benefits (from investor orders), and therefore would result in negative expected profits.

We now argue that the analyst does not have a profitable deviation. By assumption, the analyst does not send orders to buy (sell) without observing an upward (downward) jump in the value of the asset. It therefore only remains to check that it is optimal for the analyst to send orders to each exchange each time the asset jumps.

To see that this is the case, let $V(j,k)$ denote the analyst's expected continuation profits if he were to know that he would observe j more jumps and if current midprices differ from his expectation of the value of the asset by k. For the purposes of this argument, we let $\hat{X} = X - (X-1)e^{-(\delta_H+\delta_E)/\mu_A}$ denote the number of fills that the analyst expects to receive when he attempts to trade on all X exchanges. If the analyst sends orders to each exchange each time the asset jumps, then his expected profit from observing j jumps are $\hat{X}j\left(1-\frac{s^*_{LOB}}{2}\right)$. It therefore suffices to show that $V(j,0) \leq \hat{X}j\left(1-\frac{s^*_{LOB}}{2}\right)$. We in fact prove the stronger statement: $V(j,k) \leq \hat{X}\left[j\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{k}{2}\right]$ for all $j \in \mathbb{N}$, $k \in \mathbb{N}$. The proof is by induction on j. The result trivially holds for $j = 0$ and all $k \in \mathbb{N}$, since the analyst never gets another opportunity to trade (recall that by assumption the analyst can trade only at times when a jump occurred, and only in the direction of the jump even then). Now, assuming

that the inequality holds for $j-1$ and all $k \in \mathbb{N}$, we prove it for j and all $k \in \mathbb{N}$. If $k = 0$,

$$\begin{aligned}
V(j,0) &\leq \max\left\{V(j-1,1), \max_{X' \in [1,\hat{X}]}\left[X'\left(1-\frac{s^*_{LOB}}{2}\right)\right] + V(j-1,0)\right\} \\
&\leq \max\left\{\hat{X}\left[(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{1}{2}\right], \hat{X}\left[\left(1-\frac{s^*_{LOB}}{2}\right)+(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)\right]\right\} \\
&= \hat{X} j\left(1-\frac{s^*_{LOB}}{2}\right),
\end{aligned}$$

as desired. The first step uses the fact that after observing the first jump, the analyst may either abstain from trading or may take an action that produces $X' \in [1, \hat{X}]$ fills in expectation. The second step uses the induction hypothesis. The third step uses Assumption 2, which ensures that $s^*_{LOB} \leq 1$. Similarly, if $k \geq 1$,

$$\begin{aligned}
V(j,k) &\leq \frac{1}{2}\max\left\{V(j-1,k-1), \max_{X' \in [1,\hat{X}]}\left[X'\left(1-k-\frac{s^*_{LOB}}{2}\right)\right] + V(j-1,k)\right\} \\
&\quad + \frac{1}{2}\max\left\{V(j-1,k+1), \max_{X' \in [1,\hat{X}]}\left[X'\left(k+1-\frac{s^*_{LOB}}{2}\right)\right] + V(j-1,k)\right\} \\
&\leq \frac{1}{2}\max\left\{\hat{X}\left[(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{k-1}{2}\right], \left(1-k-\frac{s^*_{LOB}}{2}\right)+\hat{X}\left[(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{k}{2}\right]\right\} \\
&\quad + \frac{1}{2}\max\left\{\hat{X}\left[(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{k+1}{2}\right], \hat{X}\left[\left(1-\frac{s^*_{LOB}}{2}\right)+(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{k}{2}\right]\right\} \\
&= \frac{\hat{X}}{2}\left[(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{k}{2}\right] + \frac{\hat{X}}{2}\left[\left(1-\frac{s^*_{LOB}}{2}\right)+(j-1)\left(1-\frac{s^*_{LOB}}{2}\right)\right] \\
&\leq \hat{X}\left[j\left(1-\frac{s^*_{LOB}}{2}\right)+\frac{k}{2}\right],
\end{aligned}$$

as desired. The logic of the steps is almost as in the case of $k = 0$. The main difference is that we now consider two cases, since the jump may either increase or decrease the distance between current midprices and the analyst's expectation of the value of the asset.

Finally, given that the analyst acts on information by trading after every observed jump, his expected flow profits from a choice of r are

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

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

Proof of Theorem 2. As is evident from equation (1), the bid-ask spread is *ceteris paribus*, (*i*) nondecreasing in research intensity, (*ii*) constant in δ_E, (*iii*) constant in δ_H, (*iv*) constant in μ_A, (*v*) nonincreasing in λ_{invest}, (*vi*) nondecreasing in λ_{jump}, and (*vii*) nondecreasing in X. Applying Topkis' Theorem (Topkis, 1978) to equation (2), we see that research intensity is *ceteris paribus*, (*i*) nonincreasing in the spread, (*ii*) nondecreasing in δ_E, (*iii*) nondecreasing in δ_H, (*iv*) nonincreasing in μ_A, (*v*) constant in λ_{invest}, (*vi*) nondecreasing in λ_{jump},

and (*vii*) nondecreasing in X. By combining these observations, we establish all claimed comparative statics for S^*_{LOB} and R^*_{LOB} except that of R^*_{LOB} with respect to λ_{jump}.

Recall that the set of equilibrium research intensities, R^*_{LOB}, is the intersection of $R_{LOB}(\hat{r})$, given below, with the forty-five degree line:

$$R_{LOB}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X - (X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + \hat{r}\lambda_{jump}X} - c(r) \right\}$$

By Topkis' theorem, $R_{LOB}(\hat{r})$ is nonincreasing in $\hat{r}$ and nondecreasing in λ_{jump} (in the strong set order). We conclude that R^*_{LOB} has the claimed comparative static with respect to λ_{jump}. □

Proof of Theorem 3. The proof proceeds in three parts. First, we argue that there exists at least one solution to (4) and (5). Second, we describe the equilibrium strategies of the traders. Third, we verify that no trader has a profitable deviation.

Part One (Existence): R^*_{SD}, the set of research intensities that appear in any solution to (4) and (5), is the intersection of $R_{SD}(\hat{r})$, given below, with the forty-five degree line:

$$R_{SD}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + \hat{r}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]} - c(r) \right\}$$

As in the proof of Theorem 1, R^*_{SD} is nonempty, and S^*_{SD} can be derived from R^*_{SD} and (4).

Part Two (Description): One high-frequency trader plays the role of a "liquidity provider." The remaining high-frequency traders never submit any orders. The strategies of the liquidity provider, investors, and the analyst are analogous to those described in the proof of Theorem 1. The only differences are that s^*_{SD} and r^*_{SD} assume the roles played by s^*_{LOB} and r^*_{LOB}.

Part Three (Verification): That the investors do not have profitable deviations is as in the proof of Theorem 1.

We now argue that the liquidity provider does not have a profitable deviation. As argued in Section 5.1, the liquidity provider earns zero profits in the equilibrium. It therefore remains to show that she does not possess a deviation that would yield positive profits. As in the proof of Theorem 1, it is not profitable to deviate by quoting a larger spread, by quoting a smaller spread, or by quoting more than a single unit at either the bid or the ask.

All other high-frequency traders also earn zero profits in equilibrium. They also have no deviations that would yield positive profits. As in the proof of Theorem 1, it is not profitable to deviate by attempting to provide liquidity at a larger spread, a smaller spread, or the same spread as the liquidity provider. Finally, it is not profitable to become a stale-quote sniper, since the selective delay eliminates the possibility of earning trading profits in this way: the liquidity provider always cancels her mispriced quotes before a stale-quote sniper could pick them off.

Lastly, the argument for why the analyst does not have a profitable deviation follows exactly as in the proof of Theorem 1. □

Proof of Theorem 4. The proof proceeds in three parts. First, we argue that there exists at least one solution to (7) and (8). Second, we describe the equilibrium strategies of the traders. Third, we verify that no trader has a profitable deviation.

Part One (Existence): R^*_{FBA}, the set of research intensities that appear in any solution to (7) and (8), is the intersection of $R_{FBA}(\hat{r})$, given below, with the forty-five degree line:

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

As in the proof of Theorem 1, R^*_{FBA} is nonempty, and S^*_{FBA} can be derived from R^*_{FBA} and (7).

Part Two (Description): One high-frequency traders plays the role of a "liquidity provider." The remaining high-frequency traders never submit any orders. The strategies of the liquidity provider, investors, and the analyst are analogous to those described in the proof of Theorem 1. The only differences are that s^*_{FBA} and r^*_{FBA} assume the roles played by s^*_{LOB} and r^*_{LOB}.

Part Three (Verification): That the investors do not have profitable deviations is as in the proof of Theorem 1.

We now argue that the liquidity provider does not have a profitable deviation. As argued in Section 5.2, she earns zero profits in the equilibrium. It therefore remains to show that she does not possess a deviation that would yield positive profits. As in the proof of Theorem 1, it is not profitable to deviate by quoting a larger spread, by quoting a smaller spread, or by quoting more than a single unit at either the bid or the ask.

All other high-frequency traders also earn zero profits in equilibrium. They also have no deviations that would yield positive profits. As in the proof of Theorem 1, it is not profitable to deviate by attempting to provide liquidity at a larger spread, a smaller spread, or the same spread as the liquidity provider. Finally, it is not profitable to become a stale-quote sniper, since batching allows the analyst to trade against all mispriced quotes, leaving none for stale-quote snipers.

Lastly, the argument for why the analyst does not have a profitable deviation follows exactly as in the proof of Theorem 1. □

Proof of Theorem 5. For convenience, we restate here that the sets of equilibrium research intensities R^*_{LOB}, R^*_{SD}, and R^*_{FBA} are defined, respectively, by the intersections of the following correspondences with the forty-five degree line:

$$R_{LOB}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X - (X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + \hat{r}\lambda_{jump}X} - c(r) \right\}$$

$$R_{SD}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + \hat{r}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]} - c(r) \right\}$$

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

The proof proceeds in three parts. First, we show $S^*_{FBA} \geq S^*_{SD}$ and $R^*_{FBA} \geq R^*_{SD}$. Second, we show $S^*_{FBA} \geq S^*_{LOB}$. Third, we show $R^*_{SD} \geq R^*_{LOB}$.

Part One ($S^*_{FBA} \geq S^*_{SD}$ and $R^*_{FBA} \geq R^*_{SD}$). Define $S^*(\Omega)$ and $R^*(\Omega)$ as the set of solutions to the system

$$s^* = \frac{2r^*\lambda_{jump}\Omega}{\lambda_{invest} + r^*\lambda_{jump}\Omega} \tag{9}$$

$$r^* \in \arg\max_{r\in[0,1]} \left\{ r\lambda_{jump}\Omega\left(1 - \frac{s^*}{2}\right)\right\} \tag{10}$$

Notice that S^*_{FBA} and R^*_{FBA} correspond to $S^*(\Omega)$ and $R^*(\Omega)$ evaluated at $\Omega = X$. Similarly, S^*_{SD} and R^*_{SD} correspond to $S^*(\Omega)$ and $R^*(\Omega)$ evaluated at $\Omega = X - \frac{\mu_A}{\mu_A+\mu_H}(X - 1)e^{-(\delta_H+\delta_E)/\mu_A}$. It therefore suffices to show that $S^*(\Omega)$ and $R^*(\Omega)$ are both nondecreasing in Ω.

Notice from equation (9) that s^* is *ceteris paribus* (*i*) nondecreasing in r^* and (*ii*) nondecreasing in Ω. Furthermore, applying Topkis' Theorem to equation (10), r^* is *ceteris paribus* (*i*) nonincreasing in s^* and (*ii*) nondecreasing in Ω. By combining these observations, we establish that $S^*(\Omega)$ is nondecreasing. Next, notice that $R^*(\Omega)$ is defined by the intersection of the following correspondence with the forty-five degree line $r = \hat{r}$:

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

By Topkis' Theorem, $R(\hat{r}, \Omega)$ is nondecreasing in Ω. Since the correspondence is also nonincreasing in $\hat{r}$, this implies that $R^*(\Omega)$ is nondecreasing.

Part Two ($S^*_{FBA} \geq S^*_{LOB}$). By another application of Topkis' Theorem, $R_{FBA}(\hat{r}) \geq R_{LOB}(\hat{r})$. Since both correspondences are nonincreasing in $\hat{r}$, we conclude that $R^*_{FBA} \geq R^*_{LOB}$. Next, letting $s(r) = \frac{2r\lambda_{jump}X}{\lambda_{invest}+r\lambda_{jump}X}$, we know that $S^*_{LOB} = s(R^*_{LOB})$ and $S^*_{FBA} = s(R^*_{FBA})$. Because $s(r)$ is nondecreasing, we conclude that $S^*_{FBA} \geq S^*_{LOB}$.

Part Three ($R^*_{SD} \geq R^*_{LOB}$). By another application of Topkis' Theorem, $R_{SD}(\hat{r}) \geq R_{LOB}(\hat{r})$. Since both correspondences are nonincreasing in $\hat{r}$, we conclude that $R^*_{SD} \geq R^*_{LOB}$. □

Proof of Theorem 6. Our approach is to demonstrate that the outcomes described in the statement of the theorem constitute the frontier of a relaxed problem. We then show that outcomes described in the statement of the theorem can be implemented with contracts that satisfy all original constraints, and therefore they actually constitute the frontier of the original problem.

Part One (Relaxed Problem). We relax the social planner's problem in three ways. First, we eliminate (IC–2). Second, we replace (IR–B) and (IR–S) with the less demanding condition $U \geq 0$, which we denote (IR), where U is defined as the average expected utility

of investors:

$$U = \frac{1}{2T}\int_0^T \mathbb{E}_r\left[x_{B,t}(J_T, I_T)\tilde{v}_T + y_{B,t}(J_T, I_T) + \theta \mathbb{1}\{x_{B,t}(J_T, I_T) = 1\}\right] dt$$
$$+ \frac{1}{2T}\int_0^T \mathbb{E}_r\left[x_{S,t}(J_T, I_T)\tilde{v}_T + y_{S,t}(J_T, I_T) + \theta \mathbb{1}\{x_{S,t}(J_T, I_T) = -1\}\right] dt$$

Third, we replace (BB–1), (BB–2), and (IR–H) with the less demanding condition stated below, which we denote (BB):

$$\mathbb{E}_r\left[x_A(J_T, I_T)\tilde{v}_T + y_A(J_T, I_T)\right] + \frac{\lambda_{invest}}{2}\int_0^T \mathbb{E}_r\left[x_{B,t}(J_T, I_T)\tilde{v}_T + y_{B,t}(J_T, I_T)\right] dt$$
$$+ \frac{\lambda_{invest}}{2}\int_0^T \mathbb{E}_r\left[x_{S,t}(J_T, I_T)\tilde{v}_T + y_{S,t}(J_T, I_T)\right] dt \leq 0$$

With (IC–2) relaxed, shares of the security are interchangeable with dollars at the rate $\tilde{v}_T$ for all parties, with the exception of investors, who derive a utility "bonus" of θ from $x_{B,t}(J_T, I_T) = 1$ and $x_{S,t}(J_T, I_T) = -1$. It is therefore optimal to set $x_{B,t}(J_T, I_T) = 1$ and $x_{S,t}(J_T, I_T) = -1$. Furthermore, a sufficient statistic for the "sharing rule" from the social planner to the analyst is $s(J_T, I_T) = x_A(J_T, I_T)\tilde{v}_T + y_A(J_T, I_T)$. The relaxed problem can then be written

$$\max_{r(t), s(J_T, I_T), U} \left\{\alpha T \lambda_{invest} U - (1-\alpha)\int_0^T \lambda_{jump}[1 - r(t)]\, dt\right\}$$

subject to the constraints

(BB) $\mathbb{E}_r\left[s(J_T, I_T)\right] + T\lambda_{invest}(U - \theta) \leq 0$,

(IR) $U \geq 0$,

(IR–A) $\forall J_T \ \forall I_T,\ s(J_T, I_T) \geq 0$,

(IC–1) at all $t \in [0, T]$ and conditional on any J_t, the analyst finds it optimal to conduct research with intensity $r(t)$.

Note that (BB) must hold with equality at the optimum; otherwise it would be possible to increase U and improve the social planner's objective. Using that equality, we can rewrite (IR) as $\mathbb{E}_r[s(J_T, I_T)] \leq T\lambda_{invest}\theta$. We can also write the planner's objective as

$$\alpha T \lambda_{invest}\theta - \alpha \mathbb{E}_r[s(J_T, I_T)] - (1-\alpha)T\lambda_{jump} + (1-\alpha)\lambda_{jump}\int_0^T r(t)\, dt,$$

which, by dropping constant terms, is equivalent to

$$(1-\alpha)\lambda_{jump}\int_0^T r(t)\, dt - \alpha \mathbb{E}_r\left[s(J_T, I_T)\right].$$

Following arguments made in Holmström and Milgrom (1987) and Breuer (1995), (IC–1) is equivalent to the following condition: $s(J_T, I_T)$ can be written in the form $s(J_T, I_T) = s_0 +$

$\sum_{t\in J_T^+\cup J_T^-} s(t)$, and for all $t \in [0,T]$, $r(t) \in \arg\max_{\hat{r}\in[0,1]}\{s(t)\lambda_{jump}\hat{r} - c(\hat{r})\}$. (See Appendix D for details.) Taking all this into account, the social planner's problem becomes

$$\max_{r(t),s_0,s(t)} \left\{(1-\alpha)\lambda_{jump}\int_0^T r(t)\,dt - \alpha\mathbb{E}_r\left[s_0 + \sum_{t\in J_T^+\cup J_T^-} s(t)\right]\right\}$$

subject to the constraints

$$\text{(IR)}\quad \mathbb{E}_r\left[s_0 + \textstyle\sum_{t\in J_T^+\cup J_T^-} s(t)\right] \le T\lambda_{invest}\theta,$$

$$\text{(IR–A)}\quad \forall J_T\ \forall I_T,\ s_0 + \textstyle\sum_{t\in J_T^+\cup J_T^-} s(t) \ge 0,$$

$$\text{(IC–1)}\quad \forall t \in [0,T],\ r(t) \in \arg\max_{\hat{r}\in[0,1]}\{s(t)\lambda_{jump}\hat{r} - c(\hat{r})\}.$$

It follows from (IC–1) that for all t, $r(t)c'(r(t)) = r(t)\lambda_{jump}s(t)$.[35] Additionally, (IR–A) implies $s_0 \ge 0$. Moreover this must hold with equality; otherwise the planner could induce the same research intensity with a less expensive scheme. Evaluating the expectation, (IR) is equivalent to $s_0 + \int_0^T r(t)\lambda_{jump}s(t)\,dt \le T\lambda_{invest}\theta$, which, combining the previous observations, can be rewritten as $\int_0^T r(t)c'(r(t))\,dt \le T\lambda_{invest}\theta$. The social planner's objective can be rewritten in a similar way, so that the problem becomes

$$\max_{r(t)} \int_0^T \left[(1-\alpha)\lambda_{jump}r(t) - \alpha r(t)c'(r(t))\right]dt$$
$$\text{subject to}\quad \int_0^T r(t)c'(r(t))\,dt \le T\lambda_{invest}\theta$$

By Assumption 4, $rc'(r)$ is a convex function. Jensen's inequality (Jensen, 1906) therefore implies that the optimum can be achieved with a constant function. The social planner's problem therefore reduces to the following static optimization:

$$\max_{r\in[0,1]} \left\{(1-\alpha)\lambda_{jump}r - \alpha rc'(r)\right\}$$
$$\text{subject to}\quad rc'(r) \le \lambda_{invest}\theta$$

Let $\bar{r} = \max\{r \in [0,1] : rc'(r) \le \lambda_{invest}\theta\}$. Notice that $r = 0$ is an optimum for $\alpha = 1$ and $r = \bar{r}$ is an optimum for $\alpha = 0$. Then by the maximum theorem, any $r \in [0,\bar{r}]$ is optimal for some $\alpha \in [0,1]$. Furthermore, given any $r \in [0,\bar{r}]$, the flow rate of aggregate investor expected utility can be recovered as

$$U = \theta - \frac{rc'(r)}{\lambda_{invest}},$$

[35] By Assumption 4, $c(\cdot)$ is strictly convex and C^1. Therefore this maximization problem has a unique solution, which satisfies one of the three conditions (*i*) $r(t) = 0$ and $c'(0) \ge \lambda_{jump}s(t)$, (*ii*) $c'(r(t)) = \lambda_{jump}s(t)$, or (*iii*) $r(t) = 1$ and $c'(1) \le \lambda_{jump}s(t)$. It is never optimal for the social planner to set $s(t) \ge \frac{c'(1)}{\lambda_{jump}}$ because she could induce the same research intensity with a less expensive scheme. This leaves cases (*i*) and (*ii*), for either of which the claimed equality holds.

which establishes that the outcomes described in the statement of the theorem constitute the frontier of the relaxed problem.

Part Two (Implementation). In this step we complete the proof by arguing that any outcome described in the statement of the theorem can be implemented by a contract satisfying all constraints.

Let $r \in [0, \bar{r}]$, and let $U = \theta - \frac{rc'(r)}{\lambda_{invest}}$. Then consider the contract described below. At every time t, the analyst engages in research with intensity r. Furthermore, the analyst reports $\hat{J}_T^+ = J_T^+$ and $\hat{J}_T^- = J_T^-$. One high-frequency trader is identified as the budget breaker; the other high-frequency traders never receive any payments. The shares and dollars transferred to the analyst, investors, and the budget breaker are computed as follows. If the analyst reports a jump at time t, then $\frac{c'(r)}{\lambda_{jump}}$ dollars are transferred from the budget breaker's account to the analyst's account. If an investor arrives at time t with a desire to buy, one share is transferred from the budget breaker's account to his account, and $\hat{v}_t + \frac{rc'(r)}{\lambda_{invest}}$ dollars are transferred in the opposite direction. If an investor arrives at time t with a desire to sell, one share is transferred from his account to the budget breaker's account, and $\hat{v}_t - \frac{rc'(r)}{\lambda_{invest}}$ dollars are transferred in the opposite direction.

It is straightforward to verify that this contract implements constant research intensity r and average expected utility of investors U. Moreover, it can be verified that this contract satisfies all original constraints. Verification of (BB–1), (BB–2), (IR–B), (IR–S), (IR–A), and (IC–2) is immediate. (IR–H) holds since the flow profits of the budget breaker are zero:

$$\lambda_{invest} \frac{rc'(r)}{\lambda_{invest}} - r\lambda_{jump} \frac{c'(r)}{\lambda_{jump}} = 0.$$

It is also easy to see that (IC–1) holds, since $r = \arg\max_{\hat{r}} \left\{ \hat{r}\lambda_{jump} \frac{c'(r)}{\lambda_{jump}} - c(\hat{r}) \right\}$. □

Proof of Proposition 9. As demonstrated in the proof of Theorem 6, the research intensity in an optimal point on the frontier satisfies

$$r^* \in \arg\max_{r \in [0,1]} \left\{ (1-\alpha)\lambda_{jump} r - \alpha r c'(r) \right\}$$
$$\text{subject to} \quad rc'(r) \leq \lambda_{invest}\theta$$

For the comparative statics with respect to α and λ_{jump}, the claims about research intensity follow immediately from Topkis' Theorem. Furthermore, the claims about investor utility follow from the relationship $U^* = \theta - \frac{r^* c'(r^*)}{\lambda_{invest}}$.

For the comparative statics with respect to θ and λ_{invest}, there are two cases. First, the constraint might be binding. In this case, a small increase in either θ or λ_{invest} adds slack to the constraint, which allows for a small increase in r^*, while U^* remains constant at zero. Second, the constraint might not be binding. In this case, a small increase in either θ or λ_{invest} does not affect r^*, but raises U^*. □

Proof of Corollary 7. In the limit order book, selective delay, and frequent batch auction equilibria, the average expected utility of investors is determined by the spread through $U = \theta - \frac{s}{2}$. Therefore, by Theorem 6, an equilibrium outcome of one of these trading

mechanisms lies on the frontier if and only if the equilibrium spread, s^*, is related to the equilibrium research intensity, r^*, through the equation

$$r^* c'(r^*) = \lambda_{invest} \frac{s^*}{2}. \tag{11}$$

In fact, the equilibrium research intensities r^*_{LOB}, r^*_{SD}, and r^*_{FBA} are characterized, respectively, by the intersections of the following functions with the forty-five degree line (note that with the addition of Assumption 3, the problem is concave, and so the argmax is now unique):

$$r_{LOB}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X - (X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + \hat{r}\lambda_{jump}X} - c(r) \right\}$$

$$r_{SD}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + \hat{r}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]} - c(r) \right\}$$

$$r_{FBA}(\hat{r}) = \arg\max_{r\in[0,1]} \left\{ \frac{r\lambda_{jump}\lambda_{invest}X}{\lambda_{invest} + \hat{r}\lambda_{jump}X} - c(r) \right\}$$

And with Assumption 4, we then obtain the following equations:[36]

$$r^*_{LOB}c'(r^*_{LOB}) = \frac{r^*_{LOB}\lambda_{jump}\lambda_{invest}\left[X - (X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + r^*_{LOB}\lambda_{jump}X}$$

$$r^*_{SD}c'(r^*_{SD}) = \frac{r^*_{SD}\lambda_{jump}\lambda_{invest}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + r^*_{SD}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}$$

$$r^*_{FBA}c'(r^*_{FBA}) = \frac{r^*_{FBA}\lambda_{jump}\lambda_{invest}X}{\lambda_{invest} + r^*_{FBA}\lambda_{jump}X}$$

In addition, from equations (1), (4), and (7), we also have

$$s^*_{LOB} = \frac{2r^*_{LOB}\lambda_{jump}X}{\lambda_{invest} + r^*_{LOB}\lambda_{jump}X}$$

$$s^*_{SD} = \frac{2r^*_{SD}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + r^*_{SD}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}$$

$$s^*_{FBA} = \frac{2r^*_{FBA}\lambda_{jump}X}{\lambda_{invest} + r^*_{FBA}\lambda_{jump}X}$$

Comparing these expressions to the condition for being on the frontier, equation (11), we find that the selective delay equilibrium and the frequent batch auction equilibrium lie on

[36] By Assumptions 3 and 4, $c(\cdot)$ is strictly convex, C^1, and $\lim_{r\to 1} c(r) = \infty$. Therefore this maximization problem has a unique solution, which is either a corner solution at zero or an interior solution. These equalities hold in either case.

the frontier. Furthermore, the limit order book equilibrium lies on the frontier if and only if $(X-1)e^{-(\delta_H+\delta_E)/\mu_A} = 0$. Since δ_H and μ_A are both assumed to be "on the order of ε," this is the case if and only if either $X = 1$ or δ_E is infinitely larger than ε. □

Proof of Corollary 8. Suppose that (r, U) is an outcome on the frontier in which $r \in [0, \bar{r}']$. By Theorem 6, $U = \theta - \frac{rc'(r)}{\lambda_{invest}}$.

Then consider frequent batch auctions with $X'' = \frac{\lambda_{invest}c'(r)}{\lambda_{jump}[\lambda_{invest} - rc'(r)]}$ exchanges. While Assumption 1 and 2 may not hold when the number of exchanges is X'', their role was only to ensure that the spread did not exceed $\min\{1, 2\theta\}$. We therefore suppose for the moment that this is the case, and we verify it later. Given Assumptions 3 and 4, we can then follow arguments from the proof of Corollary 7 to see that frequent batch auctions with X'' exchanges will implement a research intensity r^* that satisfies

$$r^*c'(r^*) = \frac{r^*\lambda_{jump}\lambda_{invest}X''}{\lambda_{invest} + r^*\lambda_{jump}X''} = rc'(r)$$

By strict convexity of $c(\cdot)$, this uniquely pins down $r^* = r$. Furthermore, the spread will be

$$s^* = \frac{2r\lambda_{jump}X''}{\lambda_{invest} + r\lambda_{jump}X''} = \frac{2rc'(r)}{\lambda_{invest}}, \tag{12}$$

which implies that the average expected utility of investors in this equilibrium is U. It only remains to verify that $s^* \leq \min\{1, 2\theta\}$. However, this follows immediately from equation (12) and the fact that $r \in [0, \bar{r}']$. We conclude that the outcome (r, U) can be implemented by frequent batch auctions with X'' exchanges.

Finally, it can be shown through similar methods that the outcome (r, U) can be implemented by a selective delay with X' exchanges, where

$$X' = \frac{X'' - \frac{\mu_A}{\mu_A+\mu_H}e^{-(\delta_H+\delta_E)/\mu_A}}{1 - \frac{\mu_A}{\mu_A+\mu_H}e^{-(\delta_H+\delta_E)/\mu_A}}.$$ □

B Additional Results

In this appendix, we augment the results discussed in the main text with some additional results that may be of interest. Appendices B.1 and B.2 discuss how the equilibrium outcomes prevailing under a selective delay and frequent batch auctions depend upon the parameters of the model. Appendix B.3 discusses another alternative trading mechanism that has been proposed: implementing a minimum resting time for quotes. We find that this generates equilibrium outcomes identical to those prevailing under a limit order book. Appendix B.4 discusses the implications of consolidating data centers of all exchanges to a single location. We find that this generates equilibrium outcomes identical to those prevailing under frequent batch auctions.

B.1 Selective Delay Comparative Statics

This section uses the characterization of equilibrium outcomes under a selective delay given in Theorem 3 to study how these outcomes vary with the parameters of the model. Formally, let S^*_{SD} and R^*_{SD} denote the set of equilibrium spreads and research intensities that occur in equilibria of the form described in Theorem 3. The comparative statics of these sets with respect to the parameters of the model are given by the following theorem, and for additional convenience are also summarized in Table 3.

Theorem 10. *Within the set of parameters that satisfy Assumptions 1 and 2, the selective delay equilibrium sets of bid-ask spreads, S^*_{SD}, and research intensities, R^*_{SD}, have the following comparative statics (in the strong set order):*

(i) S^*_{SD} is nondecreasing in δ_E, nondecreasing in δ_H, nonincreasing in μ_A, nonincreasing in λ_{invest}, nondecreasing in λ_{jump}, and nondecreasing in X.

(ii) R^*_{SD} is nondecreasing in δ_E, nondecreasing in δ_H, nonincreasing in μ_A, nondecreasing in λ_{invest}, nondecreasing in λ_{jump}, and nondecreasing in X.

Table 3: Summary of predictions of Theorem 10

	δ_E	δ_H	μ_A	λ_{invest}	λ_{jump}	X
S^*_{SD}	+	+	−	−	+	+
R^*_{SD}	+	+	−	+	+	+

Proof of Theorem 10. As is evident from equation (4), the bid-ask spread is *ceteris paribus*, (*i*) nondecreasing in research intensity, (*ii*) nondecreasing in δ_E, (*iii*) nondecreasing in δ_H, (*iv*) nonincreasing in μ_A, (*v*) nonincreasing in λ_{invest}, (*vi*) nondecreasing in λ_{jump}, and (*vii*) nondecreasing in X. Applying Topkis' Theorem to equation (5), we see that research intensity is *ceteris paribus*, (*i*) nonincreasing in the spread, (*ii*) nondecreasing in δ_E, (*iii*) nondecreasing in δ_H, (*iv*) nonincreasing in μ_A, (*v*) constant in λ_{invest}, (*vi*) nondecreasing in λ_{jump}, and (*vii*) nondecreasing in X. By combining these observations, we establish all claimed comparative statics except those of R^*_{SD} with respect to δ_E, δ_H, μ_A, λ_{jump}, and X.

Recall that the set of equilibrium research intensities, R^*_{SD}, is the intersection of $R_{SD}(\hat{r})$, given below, with the forty-five degree line:

$$R_{SD}(\hat{r}) = \underset{r\in[0,1]}{\arg\max} \left\{ \frac{r\lambda_{jump}\lambda_{invest}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]}{\lambda_{invest} + \hat{r}\lambda_{jump}\left[X - \frac{\mu_A}{\mu_A+\mu_H}(X-1)e^{-(\delta_H+\delta_E)/\mu_A}\right]} - c(r) \right\}$$

By Topkis' theorem, $R_{SD}(\hat{r})$ is nonincreasing in $\hat{r}$, nondecreasing in δ_E, nondecreasing in δ_H, nonincreasing in μ_A, nondecreasing in λ_{jump}, and nondecreasing in X. We conclude that R^*_{SD} has the claimed comparative statics with respect to these parameters. □

Intuition for the comparative statics with respect to δ_E, δ_H, μ_A, λ_{invest}, and λ_{jump} is analogous to that under the limit order book discussed in Section 4.2. Intuition for the comparative statics with respect to X is as follows. The addition of another exchange

increases the number of venues at which the analyst may trade after observing a jump, 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 in response.[37]

B.2 Frequent Batch Auctions Comparative Statics

This section uses the characterization of equilibrium outcomes with frequent batch auctions given in Theorem 4 to study how these outcomes vary with the parameters of the model. Formally, let S^*_{FBA} and R^*_{FBA} denote the set of equilibrium spreads and research intensities that occur in equilibria of the form described in Theorem 4. The comparative statics of these sets with respect to the parameters of the model are given by the following theorem, and for additional convenience are also summarized in Table 4.

Theorem 11. *Within the set of parameters that satisfy Assumptions 1 and 2, the frequent batch auction equilibrium sets of bid-ask spreads, S^*_{FBA}, and research intensities, R^*_{FBA}, have the following comparative statics (in the strong set order):*

(i) *S^*_{FBA} is nonincreasing in λ_{invest}, nondecreasing in λ_{jump}, and nondecreasing in X.*

(ii) *R^*_{FBA} is nondecreasing in λ_{invest}, nondecreasing in λ_{jump}, and nondecreasing in X.*

Table 4: Summary of predictions of Theorem 11

	λ_{invest}	λ_{jump}	X
S^*_{FBA}	$-$	$+$	$+$
R^*_{FBA}	$+$	$+$	$+$

Proof of Theorem 11. As is evident from equation (7), the bid-ask spread is *ceteris paribus*, (*i*) nondecreasing in research intensity, (*ii*) nonincreasing in λ_{invest}, (*iii*) nondecreasing in λ_{jump}, and (*iv*) nondecreasing in X. Applying Topkis' Theorem to equation (8), we see that research intensity is *ceteris paribus*, (*i*) nonincreasing in the spread, (*ii*) constant in λ_{invest}, (*iii*) nondecreasing in λ_{jump}, and (*iv*) nondecreasing in X. By combining these observations, we establish all claimed comparative statics except those of R^*_{FBA} with respect to X and λ_{jump}.

Recall that the set of equilibrium research intensities, R^*_{FBA}, is the intersection of $R_{FBA}(\hat{r})$, given below, with the forty-five degree line:

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

[37] This is in contrast to under the limit order book, in which there is an ambiguous relationship between research intensity and X. The crucial difference is that with a selective delay, increased adverse selection by the analyst is the only channel through which the costs of the liquidity provider rise. On the other hand, in the limit order book, the costs of the liquidity provider also rise as a result of increased adverse selection from stale-quote snipers. These additional costs push the spread even higher, possibly to the point that the returns to research—and therefore equilibrium research intensity—would actually decline. With a selective delay, this additional cost is absent, and research intensity is unambiguously nondecreasing in X.

By Topkis' theorem, $R_{FBA}(\hat{r})$ is nonincreasing in $\hat{r}$, nondecreasing in λ_{jump}, and nondecreasing in X. We conclude that R^*_{FBA} has the claimed comparative statics with respect to λ_{jump} and X. □

Intuition for the comparative statics with respect to λ_{invest} and λ_{jump} is analogous to that under the limit order book discussed in Section 4.2. Intuition for the comparative statics with respect to X is analogous to that under a selective delay discussed in Appendix B.1. However, in contrast to before, there are no comparative statics with respect to δ_E, δ_H, and μ_A. This is because these parameters control the fraction of quoted orders that the analyst expects to convert into fills. With frequent batch auctions, the analyst converts all orders, and therefore changes in these parameters have no effect.

B.3 Minimum Resting Time

Another proposal that has received significant attention from policymakers is the implementation of a minimum resting time for quotes, or a requirement that limit orders cannot be cancelled until, at the earliest, some fixed amount of time after submission. A half-second minimum resting time received strong consideration from the European Parliament for inclusion in MiFID II, before eventually being dropped from the legislation (Stafford, 2013). A similar proposal was also considered by the Australian Securities & Investments Commission (ASIC), but ultimately rejected (Medcraft, 2013). Minimum resting times have received mention from policymakers in the United States as well (SEC, 2010).[38]

Those who support this type of policy intervention seem to do so because they view it as a way of "slowing down" high-frequency traders, which they believe would improve outcomes for ordinary traders. On the other hand, many oppose this policy because it "slows down" the "good" high-frequency traders (i.e. liquidity providers) without doing the same to the "bad" high-frequency traders (i.e. stale-quote snipers), which they believe would worsen outcomes by raising spreads.

However, in this model, a minimum resting time would have no effect whatsoever on equilibrium behavior, relative to that under a limit order book. In the limit order book, the liquidity provider never successfully cancels her mispriced quotes before they are hit by an analyst or a stale-quote sniper. Therefore, a prohibition against canceling orders soon after submission would not change the incentives of any traders in any way. Formally, the equilibrium of the limit order book identified in Theorem 1 remains an equilibrium under a minimum resting time.

B.4 Data Center Consolidation

A trend toward data center consolidation has been another recent trend affecting modern financial markets (TABB Group, 2014). A notable example of this phenomenon are the BATS and Direct Edge Exchanges who plan to consolidate into a single data centers in Secaucus, NJ in 2015 (BATS, 2014).

[38] Despite the attention that minimum resting times have received, there have been few studies of the topic. An exception is Brewer, Cvitanić, and Plott (2012).

The model also allows us to speak to the consequences of this development. Data center consolidation might manifest itself within this model as the introduction of correlation across exchanges of communication latency.

Formally, suppose that the correlation structure of communication latency presented in Section 3.2 were instead replaced with the following correlation structure. Let $L_{i,t}$ denote the amount of latency for an order submitted by trader i *to any exchange* at time t. We assume the following correlation structure: $L_{i,t} = L_{i',t'}$ if $i = i'$ and $|t - t'|$ is an infinitesimal; they are otherwise independent.

If the model were adjusted in this way, then equilibrium outcomes would be identical to those that prevail under frequent batch auctions, characterized in Theorem 4. The intuition is the same: an analyst becomes able to trade against all mispriced quotes before high-frequency traders react. Whereas batching makes this possible by synchronizing the time of trade across exchanges, data center consolidation makes this possible by synchronizing the time of order arrival across exchanges.

C Benefits of Price Efficiency

This appendix discusses rationales for the social value of price efficiency. First, price efficiency may be a positive externality for economic agents outside the financial system. The literature has identified a number of channels through which this may be the case, which are summarized in Appendix C.1.

Second, if investors are risk-averse, then they may prefer information to be incorporated into prices as soon as possible in order to reduce the amount of uncertainty that is resolved after they trade. In the baseline version of the model, this effect is not explicitly modeled because investors are risk-neutral. However, in Appendix C.2, we consider an extension in which risk-averse investors prefer higher price efficiency in addition to lower spreads.

C.1 Positive Externality

Price efficiency may be a positive externality for agents in the economy outside the financial system. By conveying information to real-world decision-makers, price efficiency can improve *economic efficiency*. The literature has identified several channels through which price efficiency may influence economic efficiency.[39] Spurred by Baumol (1965), the literature has tended to focus on two such channels: the *incentive channel* and the *learning channel*.

The incentive channel is that higher price efficiency assists a board of directors in gauging a manager's performance, thus enabling them to provide better incentives for the manager, and thereby raising the manager's effort. Diamond and Verrecchia (1982), Fishman and Hagerty (1989), and Holmström and Tirole (1993) provide theoretical models of this channel. Kang and Liu (2008) and Ferreira, Ferreira, and Raposo (2011) find empirical evidence consistent with the predictions of these models.

The learning channel is that higher price efficiency provides better feedback to firm managers, thus enabling them to make better decisions. Dow and Gorton (1997) and Sub-

[39] See Bond, Edmans, and Goldstein (2012) for a review of the literature on the effects of financial markets on the real economy.

rahmanyam and Titman (1999) provide theoretical models of this channel. Chen, Goldstein, and Jiang (2007), Bakke and Whited (2010), Kau, Linck, and Rubin (2008), and Luo (2005) find empirical evidence consistent with the operation of this channel.

In addition, higher degrees of price efficiency may be especially valuable for promoting efficient investment by "equity-dependent" firms (i.e. firms able to raise funds only through the issue of equity). Such a firm may be discouraged from undertaking an investment in the event that its stock price falls far below its fundamental value.[40] Higher degrees of price efficiency reduce the probability that this effect prevents a productive investment.

There may also be several other channels through which price efficiency might raise 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, credit-rating agencies, regulators, and blockholders, all of whom take actions that may influence the efficiency of resource allocation.[41]

C.2 Risk-Averse Investors

Risk averse investors may also prefer higher price efficiency. In this section, we present an alternate version of the model that features risk averse investors. In this version, the equilibrium utility of investors is determined not only by the spread, as in the baseline model, but also by price efficiency.

The only differences between the primitives of this version and those of the baseline model are in terms of the utility functions of investors. The first difference is that investors face a delay cost: an investor who arrives at time t with a need either to buy or to sell receives infinite disutility if this need is not satisfied by some time $t' \simeq t$.[42] The equivalence relation $t \simeq t'$ (read "t is infinitely close to t'") is defined to be the case if and only if $t - t'$ is an infinitesimal.

The second difference is that investors who satisfy their need to transact receive the utility $u(v_T - p)$ if they buy at price p, and $u(p - v_T)$ if they sell at price p, where u is an increasing and concave function.

Suppose a particular trading mechanism gives rise to an equilibrium in which (i) there is a constant spread s^*, (ii) there is a constant research intensity r^*, (iii) all investors fill their demand within an infinitesimal interval after arrival, and (iv) jumps are observed by the analyst are incorporated into prices within an infinitesimal interval after arrival, and the midprice does not otherwise change. We demonstrate that in such an equilibrium, the utility of investors is not only a nonincreasing function of the equilibrium spread, but also a nondecreasing function of the equilibrium level of research intensity (which, as we have seen, is directly tied to price efficiency).

[40] Baker, Stein, and Wurgler (2003) and Chen, Goldstein, and Jiang (2007) find empirical evidence of this effect.

[41] For example, Faure-Grimaud and Gromb (2004) argue that higher levels of price efficiency increase the incentives of blockholders to take actions that increase the value of the company. Additionally, several papers have documented a relationship between price efficiency and economic efficiency without identifying a particular channel. Examples include Durnev, Morck, and Yeung (2004) and Wurgler (2000).

[42] Rather than infinite disutility, we only need the delay cost to be sufficiently large relative to the amount of risk aversion that delaying trade is never optimal. However, we make this assumption in order to circumvent a tedious discussion of what would constitute a "sufficiently large" delay cost.

In such an equilibrium, the expected utility of an investor who arrives at time t with a need to buy is $U_t^b(s^*, r^*) = \mathbb{E}\left[u\left(v_T - \tilde{v}_t - \frac{s^*}{2}\right) \;\middle|\; r^* \right]$. Similarly, the expected utility of an investor who arrives at time t with a need to sell is $U_t^s(s^*, r^*) = \mathbb{E}\left[u\left(\tilde{v}_t - v_T - \frac{s^*}{2}\right) \;\middle|\; r^* \right]$. By symmetry of the jump process, $U_t^b(s^*, r^*) = U_t^s(s^*, r^*) = U_t(s^*, r^*)$.

Theorem 12. *$U_t(s^*, r^*)$ is nonincreasing in s^* and nondecreasing in r^*.*

Proof of Theorem 12. The effect of s^* follows immediately from u being nondecreasing. As we argue below, the effect of r^* follows from the concavity of u. In what follows, J represents the random variable that denotes the number of unobserved jumps between time zero and time t. Furthermore, f_{jump} represents the probability mass function that places probability $\frac{1}{2}$ on both -1 and $+1$.

$$\begin{aligned} U_t(s^*, r^*) &= \mathbb{E}\left[u\left(v_T - \tilde{v}_t - \frac{s^*}{2}\right) \;\middle|\; r^* \right] \\ &= \mathbb{E}\left[u\left(v_T - v_t + v_t - \tilde{v}_t - \frac{s^*}{2}\right) \;\middle|\; r^* \right] \\ &= \sum_{d=-\infty}^{\infty} \left[\sum_{j=0}^{\infty} \left[\sum_{k=-j}^{j} u\left(d + k - \frac{s^*}{2}\right) \mathbb{P}(v_t - \tilde{v}_t = k \mid J = j) \right] \mathbb{P}(J = j \mid r^*) \right] \mathbb{P}(v_T - v_t = d) \end{aligned}$$

In the above, d indexes $v_T - v_t$, the net change in value between t and T; j indexes the number of unobserved jumps between 0 and t; and k indexes $v_t - \tilde{v}_t$, the net change in value due to unobserved jumps between 0 and t. We use the fact that $v_T - v_t$ and $v_t - \tilde{v}_t$ are independent random variables.

We first note that $\mathbb{P}(v_t - \tilde{v}_t = k \mid J = j) = f_{jump}^{*j}(k)$, where f_{jump}^{*j} denotes the jth convolution power of f_{jump}. For any j, f_{jump}^{*j} is a mean-preserving spread of $f_{jump}^{*(j-1)}$. Therefore, f_{jump}^{*j} is nonincreasing in j in the sense of second-order stochastic dominance. Thus, by the concavity of u, for any fixed d, $\sum_{k=-j}^{j} u\left(d + k - \frac{s^*}{2}\right) \mathbb{P}(v_t - \tilde{v}_t = k \mid J = j)$ is nonincreasing in j.

For a given r^*, J follows a Poisson distribution with mean $t(1 - r^*)\lambda_{jump}$. This distribution is nonincreasing in r^* in the sense of first-order stochastic dominance. Combining this observation with the above paragraph allows us to conclude that the following expression is nondecreasing in r^* for any fixed d

$$\sum_{j=0}^{\infty} \left[\sum_{k=-j}^{j} u\left(d + k - \frac{s^*}{2}\right) \mathbb{P}(v_t - \tilde{v}_t = k \mid J = j) \right] \mathbb{P}(J = j \mid r^*).$$

Then, finally, we conclude that $U_t(s^*, r^*)$ is nondecreasing in r^*. □

D Implementation of Repeated Principal-Agent Problem

In this section, we demonstrate that (IC–1) is equivalent to the condition that $s(J_T, I_T)$ can be written in the form $s(J_T, I_T) = s_0(I_T) + \sum_{t \in J_T^+ \cup J_T^-} s(t)$, where for all $t \in [0, T]$,

$r(t) \in \arg\max_{\hat{r}\in[0,1]}\{s(t)\lambda_{jump}\hat{r} - c(\hat{r})\}$. The analysis relies heavily on Holmström and Milgrom (1987) and closely follows arguments made in Breuer (1995).

D.1 Implementation in the Single Period Problem

Consider a risk-neutral principal who interacts with a risk-neutral agent over the time interval $[0, \frac{1}{N}]$. At each time $\tau \in [0, \frac{1}{N}]$, the agent chooses the arrival rate of a Poisson process as $r_1(\tau) \in [0, \frac{1}{N}]$ at the cost $\int_0^{1/N} c(r_1(\tau))\, d\tau$. Let x^1 be the random variable that is the number of arrivals. Suppose that no information about x^1 is available to either party until time $\frac{1}{N}$. The function $r_1(\tau)$ is implementable by the sharing rule $s(x^1)$ if

$$r_1(\tau) \in \arg\max_{\hat{r}_1} \left\{ \mathbb{E}_{\hat{r}_1}\left[s(x^1)\right] - \int_0^{1/N} c(\hat{r}_1(\tau))\, d\tau \right\}$$

Notice that

$$\begin{aligned}
\mathbb{E}_{r_1}\left[s(x^1)\right] &= e^{-\frac{1}{N}\int_0^{1/N} r_1(\tau)\, d\tau} \sum_{k=0}^{\infty} \frac{(\frac{1}{N}\int_0^{1/N} r_1(\tau)\, d\tau)^k}{k!} s(k) \\
&= s(0) + e^{-\frac{1}{N}\int_0^{1/N} r_1(\tau)\, d\tau} \sum_{k=1}^{\infty} \frac{(\frac{1}{N}\int_0^{1/N} r_1(\tau)\, d\tau)^k}{k!} [s(k) - s(0)]
\end{aligned}$$

so the implementation condition can be rewritten as

$$r_1(\tau) \in \arg\max_{\hat{r}_1} \left\{ N e^{-\frac{1}{N}\int_0^{1/N} \hat{r}_1(\tau)\, d\tau} \sum_{k=1}^{\infty} \frac{(\frac{1}{N}\int_0^{1/N} \hat{r}_1(\tau)\, d\tau)^k}{k!} [s(k) - s(0)] - \int_0^{1/N} c(\hat{r}_1(\tau))\, d\tau \right\}$$

D.2 Implementation in the Multiple Period Problem

Now suppose that the principal and agent interact over the time interval $[0, 1]$. At each time $\tau \in [0, 1]$, the agent chooses the arrival rate of a Poisson process as $r(\tau) \in [0, 1]$ at the cost $\int_0^1 c(r(\tau))\, d\tau$. Let x^n denote the random variable that is the number of arrivals in the interval $\left[\frac{(n-1)}{N}, \frac{n}{N}\right]$. Let $X^n = (x^1, \ldots, x^n)$. Suppose that no information about x^n is released to either party until the time $\frac{n}{N}$. Let $r_n(\tau)$ be the restriction of $r(\tau)$ to the domain $\left[\frac{(n-1)}{N}, \frac{n}{N}\right]$.

Theorem 4 of Holmström and Milgrom (1987) implies that the sharing rule $s(X^N)$ implements $r(\tau)$ if and only if it can be written in the form

$$s(X^N) = \sum_{n=1}^{N} s_n(x^n),$$

where each $r_n(\tau)$ is implementable by $s_n(x^n)$ in the single-period problem.

D.3 Implementation in the Continuous Time Limit

We treat continuous time as the discrete time limit (as $N \to \infty$) of the multiple period problem discussed above. As $N \to \infty$, the condition for single-period implementation converges to

$$r \in \arg\max_{\hat{r} \in [0,1]} \left\{ [s(1) - s(0)]\hat{r} - c(\hat{r}) \right\}.$$

Let X^* denote the set of times in $[0,1]$ at which an arrival occurred. As $N \to \infty$, the information contained in X^N converges to X^*. Thus, taking the limit of the multiple period problem as $N \to \infty$, the sharing rule $s(X^*)$ implements $r(\tau)$ if and only if it can be written in the form

$$s(X^*) = s_0 + \sum_{\tau \in X^*} \hat{s}(\tau), \tag{13}$$

where for all $\tau \in [0,1]$,

$$r(\tau) \in \arg\max_{\hat{r} \in [0,1]} \hat{s}(\tau)\hat{r} - c(\hat{r}). \tag{14}$$

Of course, s_0 and $\hat{s}(\tau)$ could possibly be functions of exogenous random variables (e.g. in the context of Section 6, the investor arrival history) rather than constants. However, it will not be optimal to condition on these random variables, and so we ignore this possibility.

References

Adler, J. (2012): "Raging Bulls: How Wall Street Got Addicted to Light-Speed Trading," *Wired Magazine*.

Aequitas (2013): "Application for Recognition of Aequitas Innovations Inc. and Aequitas Neo Exchange Inc. as an Exchange," Notice and Request for Comment.

Aggarwal, R. K. and G. Wu (2006): "Stock Market Manipulations," *The Journal of Business*, 79, 1915–1953.

Alpha Exchange Inc. (2014): "Notice of Proposed Rule Amendments and Request for Comments," Notice of Proposed Rule Amendments.

Back, K. and S. Baruch (2004): "Information in Securities Markets: Kyle Meets Glosten and Milgrom," *Econometrica*, 72, 433–465.

Baker, M., J. C. Stein, and J. Wurgler (2003): "When Does the Market Matter? Stock Prices and the Investment of Equity-Dependent Firms," *The Quarterly Journal of Economics*, 118, 969–1005.

Bakke, T.-E. and T. M. Whited (2010): "Which Firms Follow the Market? An Analysis of Corporate Investment Decisions," *Review of Financial Studies*, 23, 1941–1980.

Baldauf, M. and J. Mollner (2015): "Trading in Fragmented Markets," *Working Paper*.

Barclays Capital Inc. (2014): "Why the Smartness of Order Routing Matters in Options Trading: How Order Routing Performance Impacts the Bottom Line," .

Baron, M., J. Brogaard, and A. Kirilenko (2012): "The Trading Profits of High Frequency Traders," *Working Paper*.

BATS (2014): "Integration FAQ," http://www.batsglobalmarkets.com/us/equities/edge_integration/faq.

Baumol, W. J. (1965): *The Stock Market and Economic Efficiency*, Fordham University Press.

Berge, C. (1963): *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*, Courier Dover Publications.

Biais, B., T. Foucault, and S. Moinas (2013): "Equilibrium Fast Trading," *Working Paper*.

Boehmer, B. and E. Boehmer (2003): "Trading Your Neighbor's ETFs: Competition or Fragmentation?" *Journal of Banking & Finance*, 27, 1667–1703.

Boehmer, E., K. Y. Fong, and J. J. Wu (2014): "International Evidence on Algorithmic Trading," *Working Paper*.

Bond, P., A. Edmans, and I. Goldstein (2012): "The Real Effects of Financial Markets," *Annual Review of Financial Economics*, 4, 339–360.

Breuer, W. (1995): "Lineare Anreizverträge und Poisson-Prozesse in ökonomischen Agency-Modellen," *Operations-Research-Spektrum*, 17, 245–251.

Brewer, P., J. Cvitanić, and C. R. Plott (2012): "Minimum Resting Times vs. Call

Markets and Circuit Breakers," Foresight, Government Office for Science.

Brogaard, J. (2010): "High Frequency Trading and Its Impact on Market Quality," *Working Paper*.

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan (2013): "Trading Fast and Slow: Colocation and Market Quality," *Working Paper*.

Brogaard, J., T. Hendershott, and R. Riordan (2014): "High-Frequency Trading and Price Discovery," *Review of Financial Studies*, 27, 2267–2306.

Budish, E., P. Cramton, and J. Shim (2013): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Working Paper*.

——— (2014): "Implementation Details for Frequent Batch Auctions: Slowing Down Markets to the Blink of an Eye," *American Economic Review: Papers & Proceedings*, 104, 418–424.

Carrion, A. (2013): "Very Fast Money: High-Frequency Trading on the NASDAQ," *Journal of Financial Markets*, 16, 680–711.

Chaboud, A., B. Chiquoine, E. Hjalmarsson, and C. Vega (2013): "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market," *Journal of Finance, Forthcoming*.

Chen, Q., I. Goldstein, and W. Jiang (2007): "Price Informativeness and Investment Sensitivity to Stock Price," *Review of Financial Studies*, 20, 619–650.

Copeland, T. E. and D. Galai (1983): "Information Effects on the Bid-Ask Spread," *The Journal of Finance*, 38, 1457–1469.

Diamond, D. W. and R. E. Verrecchia (1982): "Optimal Managerial Contracts and Equilibrium Security Prices," *The Journal of Finance*, 37, 275–287.

Dow, J. and G. Gorton (1997): "Stock Market Efficiency and Economic Efficiency: Is There a Connection?" *The Journal of Finance*, 52, 1087–1129.

Durnev, A., R. Morck, and B. Yeung (2004): "Value-Enhancing Capital Budgeting and Firm-specific Stock Return Variation," *The Journal of Finance*, 59, 65–105.

Easley, D., T. Hendershott, and T. Ramadorai (2014): "Leveling the trading field," *Journal of Financial Markets*, 17, 65–93.

Fama, E. F. (1970): "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, 25, 383–417.

Faure-Grimaud, A. and D. Gromb (2004): "Public Trading and Private Incentives," *Review of Financial Studies*, 17, 985–1014.

Ferreira, D., M. A. Ferreira, and C. C. Raposo (2011): "Board Structure and Price Informativeness," *Journal of Financial Economics*, 99, 523–545.

Fink, J., K. E. Fink, and J. P. Weston (2006): "Competition on the Nasdaq and the Growth of Electronic Communication Networks," *Journal of Banking & Finance*, 30, 2537–2559.

Fishman, M. J. and K. M. Hagerty (1989): "Disclosure Decisions by Firms and the Competition for Price Efficiency," *The Journal of Finance*, 44, 633–646.

Foresight (2012): "The Future of Computer Trading in Financial Markets," The Government Office for Science, London.

Foucault, T., J. Hombert, and I. Roşu (2013): "News Trading and Speed," *Working Paper*.

Foucault, T. and A. J. Menkveld (2008): "Competition for Order Flow and Smart Order Routing Systems," *The Journal of Finance*, 63, 119–158.

Glosten, L. R. (1994): "Is the Electronic Open Limit Order Book Inevitable?" *The Journal of Finance*, 49, 1127–1161.

Glosten, L. R. and P. R. Milgrom (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 14, 71–100.

Goldblatt, R. (1998): *Lectures on the Hyperreals: An Introduction to Nonstandard Analysis*, Springer.

Grossman, S. J. and J. E. Stiglitz (1980): "On the Impossibility of Informationally Efficient Markets," *The American Economic Review*, 393–408.

Hasbrouck, J. and G. Saar (2013): "Low-Latency Trading," *Journal of Financial Markets*, 16, 646–679.

Hendershott, T., C. M. Jones, and A. J. Menkveld (2011): "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance*, 66, 1–33.

Hendershott, T. and P. C. Moulton (2011): "Automation, Speed, and Stock Market Quality: The NYSE's Hybrid," *Journal of Financial Markets*, 14, 568–604.

Holmström, B. and P. Milgrom (1987): "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica*, 55, 303–328.

Holmström, B. and J. Tirole (1993): "Market Liquidity and Performance Monitoring," *Journal of Political Economy*, 101, 678–709.

IEX Group (2014): "About Us: A Market that Works for Investors," http://www.iextrading.com/about.

Jensen, J. (1906): "Sur les Fonctions Convexes et les Inégalités Entre les Valeurs Moyennes," *Acta Mathematica*, 30, 175–193.

Kang, Q. and Q. Liu (2008): "Stock Trading, Information Production, and Executive Incentives," *Journal of Corporate Finance*, 14, 484–498.

Kau, J. B., J. S. Linck, and P. H. Rubin (2008): "Do Managers Listen to the Market?" *Journal of Corporate Finance*, 14, 347–362.

KCG Holdings, Inc. (2013): "Form 8-K," http://www.sec.gov/Archives/edgar/data/1569391/000119312513438413/d625930d8k.htm.

Knight Holdco, Inc. (2013): "Form S-4 Registration Statement," http://www.sec.gov/Archives/edgar/data/1569391/000119312513053260/d484578ds4.htm#toc.

Kyle, A. S. (1985): "Continuous Auctions and Insider Trading," *Econometrica*, 53, 1315–1335.

Luo, Y. (2005): "Do Insiders Learn from Outsiders? Evidence from Mergers and Acquisitions," *The Journal of Finance*, 60, 1951–1982.

Madhavan, A. (1992): "Trading Mechanisms in Securities Markets," *the Journal of Finance*, 47, 607–641.

Malinova, K., A. Park, and R. Riordan (2013): "Do Retail Traders Suffer From High Frequency Traders?" *Working Paper*.

Martinez, V. H. and I. Roşu (2013): "High Frequency Traders, News and Volatility," *Working Paper*.

McKay Brothers (2014): "Faster Than Others," http://www.mckay-brothers.com/faster-than-others.

Medcraft, G. (2013): "Building and Innovating Towards Stronger, Cleaner Markets," Australian Securities and Investments Commission.

Menkveld, A. J. (2013): "High Frequency Trading and the *New-Market* Makers," *Journal of Financial Markets*, 16, 712–740.

Nelemans, M. (2008): "Redefining Trade-Based Market Manipulation," *Valparaiso University Law Review*, 42, 1169–1219.

NYSE (2004): "NYSE Approves Expansion of Automatic Trading; Exchange Will Propose to Broaden Access to Speed and Certainty of NYSE Direct+®," http://www1.nyse.com/press/1075990965691.html.

——— (2009): "NYSE Cuts Order-Execution Time to 5 Milliseconds from 105," http://www1.nyse.com/press/1246442836537.html.

Peterffy, T. (2014): "Interactive Brokers Group Proposal to Address High Frequency Trading," Open Letter to SEC.

Riordan, R. and A. Storkenmaier (2012): "Latency, Liquidity and Price Discovery," *Journal of Financial Markets*, 15, 416–437.

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

Sannikov, Y. and A. Skrzypacz (2014): "Dynamic Trading: Price Inertia, Front-Running and Relationship Banking," *Working Paper*.

Schneiderman, E. (2014): "Remarks on High-Frequency Trading & Insider Trading 2.0," http://www.ag.ny.gov/pdfs/HFT_and_market_structure.pdf.

Stafford, P. (2013): "Europe Agrees on High-Speed Trading Regulation," *Financial Times*, http://on.ft.com/17GTEa5.

Subrahmanyam, A. and S. Titman (1999): "The Going-Public Decision and the Development of Financial Markets," *The Journal of Finance*, 54, 1045–1082.

TABB Group (2014): "TABB Group Says Significant Portion of Global Capital Markets Trading Activity Now Found in Only 10 Datacenters," http://www.tabbgroup.com/PageDetail.aspx?PageID=16&ItemID=1366.

Topkis, D. M. (1978): "Minimizing a Submodular Function on a Lattice," *Operations Research*, 26, 305–321.

United States Code (1934): "Securities Exchange Act of 1934," 15 U.S.C. §78a.

U.S. Securities and Exchange Commission (2010): "Concept Release on Equity Market Structure," *Release No. 34-61358*.

——— (2013): "The Investor's Advocate: How the SEC Protects Investors, Maintains Market Integrity, and Facilitates Capital Formation," http://www.sec.gov/about/whatwedo.shtml.

Wah, E. and M. P. Wellman (2013): "Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model," in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, ACM, 855–872.

White, M. J. (2014): "Enhancing Our Equity Market Structure," http://www.sec.gov/News/Speech/Detail/Speech/1370542004312, remarks at Sandler O'Neill & Partners, L.P. Global Exchange and Brokerage Conference, New York, N.Y.

Wurgler, J. (2000): "Financial Markets and the Allocation of Capital," *Journal of Financial Economics*, 58, 187–214.