THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 130 November 2015 Issue 4

THE HIGH-FREQUENCY TRADING ARMS RACE: FREQUENT BATCH AUCTIONS AS A MARKET DESIGN RESPONSE*

Eric Budish
Peter Cramton
John Shim

The high-frequency trading arms race is a symptom of flawed market design. Instead of the continuous limit order book market design that is currently predominant, we argue that financial exchanges should use frequent batch

*First version: July 2013. Project start date: October 2010. For helpful discussions we are grateful to numerous industry practitioners; seminar audiences at the University of Chicago, Chicago Fed, Université Libre de Bruxelles, University of Oxford, Wharton, NASDAQ, IEX Group, Berkeley, NBER Market Design, NYU, MIT, Harvard, Columbia, Spot Trading, CFTC, Goldman Sachs, Toronto, AQR, FINRA, SEC First Annual Conference on Financial Market Regulation, NBER IO, UK FCA, Northwestern, Stanford, Netherlands AFM, Paris Market Microstructure, Utah WFC, NY Fed, Cornell, FRB; and Susan Athey, Larry Ausubel, Eduardo Azevedo, Simcha Barkai, Ben Charoenwong, Adam Clark-Joseph, John Cochrane, Doug Diamond, Darrell Duffie, Gene Fama, Doyne Farmer, Thierry Foucault, Alex Frankel, Matt Gentzkow, Larry Glosten, Terry Hendershott, Ali Hortacsu, Laszlo Jakab, Emir Kamenica, Brian Kelly, Pete Kyle, Jon Levin, Donald MacKenzie, Gregor Matvos, Albert Menkveld, Paul Milgrom, Toby Moskowitz, Matt Notowidigdo, Mike Ostrovsky, David Parkes, Canice Prendergast, Al Roth, Gideon Saar, Jesse Shapiro, Spyros Skouras, Andy Skrzypacz, Chester Spatt, Lars Stole, Geoff Swerdlin, Richard Thaler, Brian Weller, Michael Wellman, and Bob Wilson. We thank Daniel Davidson, Michael Wong, Ron Yang, and especially Geoff Robinson for outstanding research assistance. We thank the editor, Andrei Shleifer, and four anonymous referees for detailed and insightful comments that improved the article in many ways. Budish gratefully acknowledges financial support from the National Science Foundation (ICES-1216083), the Fama-Miller Center for Research in Finance at the University of Chicago Booth School of Business, and the Initiative on Global Markets at the University of Chicago Booth School of Business. Corresponding author, Eric Budish: Chicago Booth, 5807 S Woodlawn Avenue, Chicago IL 60637; phone: 773-702-8453; email: eric.budish@chicagobooth.edu.

The Quarterly Journal of Economics (2015), 1547–1621. doi:10.1093/qje/qjv027.
Advance Access publication on July 23, 2015.

auctions: uniform price double auctions conducted, for example, every tenth of a second. That is, time should be treated as discrete instead of continuous, and orders should be processed in a batch auction instead of serially. Our argument has three parts. First, we use millisecond-level direct-feed data from exchanges to document a series of stylized facts about how the continuous market works at high-frequency time horizons: (i) correlations completely break down; which (ii) leads to obvious mechanical arbitrage opportunities; and (iii) competition has not affected the size or frequency of the arbitrage opportunities, it has only raised the bar for how fast one has to be to capture them. Second, we introduce a simple theory model which is motivated by and helps explain the empirical facts. The key insight is that obvious mechanical arbitrage opportunities, like those observed in the data, are built into the market design—continuous-time serial-processing implies that even symmetrically observed public information creates arbitrage rents. These rents harm liquidity provision and induce a never-ending socially wasteful arms race for speed. Last, we show that frequent batch auctions directly address the flaws of the continuous limit order book. Discrete time reduces the value of tiny speed advantages, and the auction transforms competition on speed into competition on price. Consequently, frequent batch auctions eliminate the mechanical arbitrage rents, enhance liquidity for investors, and stop the high-frequency trading arms race. *JEL* Codes: D47, D44, D82, G10, G14, G20.

I. Introduction

In 2010, Spread Networks completed construction of a new high-speed fiber optic cable connecting financial markets in New York and Chicago. Whereas previous connections between the two financial centers zigzagged along railroad tracks, around mountains, etc., Spread Networks' cable was dug in a nearly straight line. Construction costs were estimated at $300 million. The result of this investment? Round-trip communication time between New York and Chicago was reduced . . . from 16 milliseconds to 13 milliseconds. Three milliseconds may not seem like much, especially relative to the speed at which fundamental information about companies and the economy evolves. (The blink of a human eye lasts 400 milliseconds; reading this parenthetical took roughly 3,000 milliseconds.) But industry observers remarked that 3 milliseconds is an "eternity" to high-frequency trading (HFT) firms, and that "anybody pinging both markets has to be on this line, or they're dead." One observer joked at the time that the next innovation will be to dig a tunnel, speeding up transmission time even further by "avoiding the planet's pesky curvature." Spread Networks may not find this joke funny anymore, as its cable is already obsolete. While tunnels have yet to materialize, a different way to get a straighter line from New York to Chicago is to use microwaves rather than fiber optic cable, since light travels faster through air than through glass. Since its emergence in around 2011, microwave technology has reduced round-trip transmission time first to around 10 milliseconds, then 9 milliseconds, then 8.5 milliseconds, and most recently to 8.1 milliseconds. Analogous speed races are occurring throughout the financial system, sometimes measured at the level of microseconds (millionths of a second) and even nanoseconds (billionths of a second).[1]

We argue that the high-frequency trading arms race is a symptom of a basic flaw in the design of modern financial exchanges: *continuous-time trading*. That is, under the continuous limit order book market design that is currently predominant, it is possible to buy or sell stocks or other exchange-traded financial instruments at any instant during the trading day.[2] We propose a simple alternative: *discrete-time trading*. More precisely, we propose a market design in which the trading day is divided into extremely frequent but discrete time intervals; to fix ideas, say, 100 milliseconds. All trade requests received during the same interval are treated as having arrived at the same (discrete) time. Then, at the end of each interval, all outstanding orders are processed in batch, using a uniform-price auction, as opposed to the serial processing that occurs in the continuous market. We call this market design *frequent batch auctions*. Our argument against continuous limit order books and in favor of frequent batch auctions has three parts.

The first part uses millisecond-level direct-feed data from exchanges to document a series of stylized facts about continuous limit order book markets. Together, the facts suggest that continuous limit order book markets violate basic asset pricing principles at high-frequency time horizons—that is, the continuous market does not actually "work" in continuous time. Consider Figure I. The figure depicts the price paths of the two largest financial instruments that track the S&P 500 index, the SPDR S&P 500 exchange traded fund (ticker SPY) and the S&P 500 E-mini futures contract (ticker ES), on a trading day in 2011. In

[1] Sources for this paragraph: Steiner (2010); Najarian (2010); Conway (2011); Troianovski (2012); Adler (2012); Bunge (2013); Laughlin, Aguirre, and Grundfest (2014); McKay Brothers Microwave Latencies Table, January 20, 2015 (http://www.mckay-brothers.com/product-page/#latencies), Aurora-Carteret route.

[2] Computers do not literally operate in continuous time; they operate in discrete time in increments of about 0.3 nanosecond. More precisely, what we mean by continuous time is as-fast-as-possible discrete time plus random serial processing of orders that reach the exchange at the exact same discrete time.

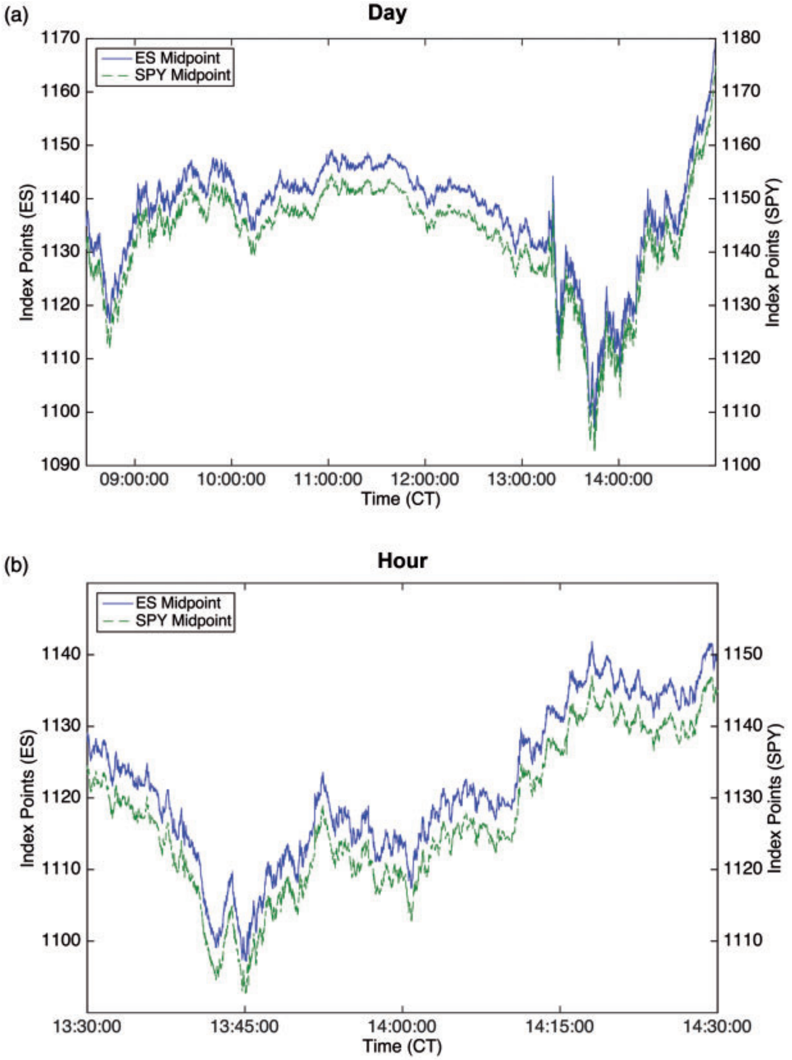

FIGURE I

ES and SPY Time Series at Human-Scale and High-Frequency Time Horizons

This figure illustrates the time series of the E-mini S&P 500 future (ES) and SPDR S&P 500 ETF (SPY) bid-ask midpoints over the course of a trading day (August 9, 2011) at different time resolutions: the full day (a), an hour (b), a minute (c), and 250 milliseconds (d). SPY prices are multiplied by 10 to reflect that SPY tracks $\frac{1}{10}$ the S&P 500 Index. Note that there is a difference in levels between the two financial instruments due to differences in cost-of-carry, dividend exposure, and ETF tracking error; for details see Section V.B. For details regarding the data, see Section IV.

(continued)

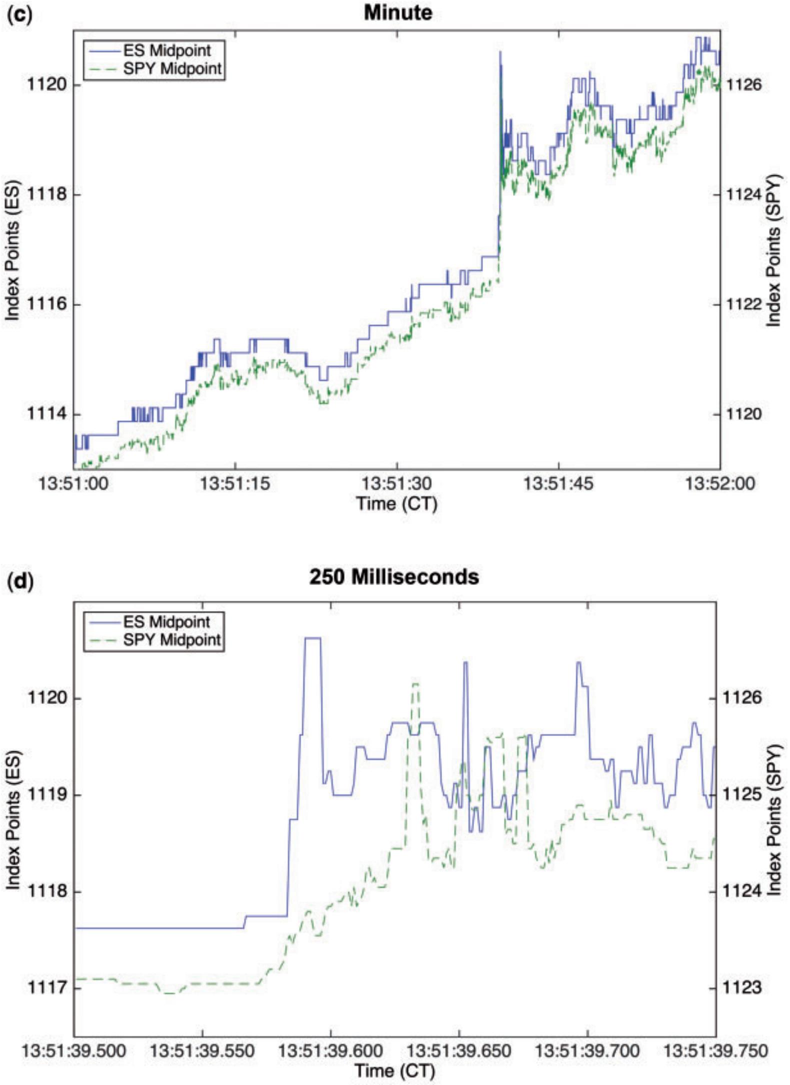

FIGURE I

Continued

Panel A, we see that the two instruments are nearly perfectly correlated over the course of the trading day, as we would expect given the near-arbitrage relationship between them. Similarly, the instruments are nearly perfectly correlated over the course of an hour (Panel B) or a minute (Panel C). However, when we zoom in to high-frequency time scales, in Panel D, we see that the correlation breaks down. Over all trading days in 2011, the median return correlation is just 0.1016 at 10 milliseconds and 0.0080 at 1 millisecond.[3] This correlation breakdown in turn leads to obvious mechanical arbitrage opportunities, available to whoever is fastest. For instance, at 1:51:39.590 PM, after the price of ES has just jumped roughly 2.5 index points, the arbitrage opportunity is to buy SPY and sell ES.

The usual economic intuition about obvious arbitrage opportunities is that once discovered, competitive forces eliminate the inefficiency. But that is not what we find here. Over the time period of our data, 2005–2011, we find that the duration of ES-SPY arbitrage opportunities declines dramatically, from a median of 97 milliseconds in 2005 to a median of 7 milliseconds in 2011. This reflects the substantial investments by HFT firms in speed during this time period. But we also find that the profitability of ES-SPY arbitrage opportunities is remarkably constant throughout this period, at a median of about 0.08 index points per unit traded. The frequency of arbitrage opportunities varies considerably over time, but its variation is driven almost entirely by variation in market volatility. These findings suggest that while there is an arms race in speed, the arms race does not actually affect the size of the arbitrage prize; rather, it just continually raises the bar for how fast one has to be to capture a piece of the prize. A complementary finding is that the number of milliseconds necessary for economically meaningful correlations to emerge has been steadily decreasing over the time period 2005–2011; but in all years, correlations are essentially zero at high-

[3] There are some subtleties involved in calculating the 1 millisecond correlation between ES and SPY, since it takes light roughly 4 milliseconds to travel between Chicago (where ES trades) and New York (where SPY trades), and this represents a lower bound on the amount of time it takes information to travel between the two markets (Einstein 1905). Whether we compute the correlation based on New York time (treating Chicago events as occurring 4 milliseconds later in New York than they do in Chicago), based on Chicago time, or ignore the theory of special relativity and use SPY prices in New York time and ES prices in Chicago time, the correlation remains essentially zero. See Section V and Online Appendix A.1 for further details.

enough frequency. Overall, our analysis suggests that the mechanical arbitrage opportunities and resulting arms race should be thought of as a constant of the market design, rather than as an inefficiency that is competed away over time.

We compute that the total prize at stake in the ES-SPY race averages $75 million per year. And, of course, ES-SPY is just a single pair of financial instruments—there are hundreds if not thousands of other pairs of highly correlated instruments, and, in fragmented equity markets, arbitrage trades that are even simpler, since the same stock trades on multiple venues. Although we hesitate to put a precise estimate on the total size of the prize in the speed race, commonsense extrapolation from our ES-SPY estimates suggests that the sums are substantial.

The second part of the article presents a simple new theory model which is motivated by and helps explain and interpret these empirical facts. The model serves as a critique of the continuous limit order book market design, and it articulates the economics of the HFT arms race. In the model, there is a security, x, that trades on a continuous limit order book, and a public signal of x's value, y. We make a purposefully strong assumption about the relationship between x and y: the fundamental value of x is perfectly correlated to the public signal y. Moreover, we assume that x can always be costlessly liquidated at its fundamental value, and initially assume away all latency for trading firms (aka HFTs, market makers, algorithmic traders). This setup can be interpreted as a "best-case" scenario for price discovery and liquidity provision in a continuous limit order book, assuming away asymmetric information, inventory costs, etc.

Given that we have eliminated the traditional sources of costly liquidity provision, one might expect that Bertrand competition among trading firms leads to costless liquidity for investors and zero rents for trading firms. But consider the mechanics of what happens in the market for x when the public signal y jumps—the moment at which the correlation between x and y temporarily breaks down. For instance, imagine that x represents SPY and y represents ES, and consider what happens at 1:51:39.590 PM in Figure I, Panel D, when the price of ES has just jumped. At this moment, trading firms providing liquidity in the market for x will send a message to the exchange to adjust their quotes—cancel their stale quotes and replace them with updated quotes based on the new value of y. At the exact same time, however, other trading firms will try to "snipe" the stale quotes—send a message to the exchange attempting to buy x at the old ask, before the liquidity providers can adjust. Since the continuous limit order book processes message requests in serial (i.e., one at a time in order of receipt), it is effectively random whose request is processed first. And, to avoid being sniped, each one of the liquidity provider's request to cancel has to get processed before all of the other trading firms' requests to snipe her stale quotes; hence, if there are N trading firms, each liquidity provider is sniped with probability $\frac{N-1}{N}$. This shows that trading firms providing liquidity, even in an environment with only symmetric information and with no latency, still get sniped with high probability because of the rules of the continuous limit order book. The obvious mechanical arbitrage opportunities we observed in the data are in a sense "built in" to the market design: continuous-time serial-processing creates arbitrage rents from symmetrically observed public information.

These arbitrage rents increase the cost of liquidity provision. In a competitive market, trading firms providing liquidity incorporate the cost of getting sniped into the bid-ask spread that they charge; so there is a positive bid-ask spread even without asymmetric information about fundamentals. Similarly, sniping causes the continuous limit order book market to be thin; that is, it is especially expensive for investors to trade large quantities of stock. The reason is that sniping costs scale linearly with the quantity liquidity providers offer in the book—if quotes are stale, they will get sniped for the whole amount. Whereas the benefits of providing a deep book scale less than linearly—since only some investors wish to trade large amounts.[4,5]

[4] Our source of costly liquidity provision should be viewed as incremental to the usual sources of costly liquidity provision: inventory costs (Demsetz 1968; Stoll 1978), asymmetric information (Copeland and Galai 1983; Glosten and Milgrom 1985; Kyle 1985), and search costs (Duffie, Garleanu, and Pedersen 2005). Mechanically, our source of costly liquidity provision is most similar to that in Copeland and Galai (1983) and Glosten and Milgrom (1985)—we discuss the relationship in detail in Section VI.C. Note too that while our model is extremely stylized, one thing we do not abstract from is the rules of the continuous limit order book itself, whereas Glosten and Milgrom (1985) and subsequent market microstructure analyses of limit order book markets use a discrete-time sequential-move modeling abstraction of the continuous limit order book. This abstraction is innocuous in the context of these prior works, but it precludes a race to respond to symmetrically observed public information as in our model.

[5] A point of clarification: our claim is *not* that markets are less liquid today than before the rise of electronic trading and HFT; the empirical record is clear that

These arbitrage rents also induce a never-ending speed race. We modify our model to allow trading firms to invest in a simple speed technology, which allows them to observe innovations in *y* faster than trading firms who do not invest. With this modification, the arbitrage rents lead to a classic prisoner's dilemma: snipers invest in speed to try to win the race to snipe stale quotes; liquidity providers invest in speed to try to get out of the way of the snipers; and all trading firms would be better off if they could collectively commit not to invest in speed, but it is in each firm's private interest to invest. Notably, competition in speed does not fix the underlying problem of mechanical arbitrages from symmetrically observed public information. The size of the arbitrage opportunity, and hence the harm to investors via reduced liquidity, depends neither on the magnitude of the speed improvements (be they milliseconds, microseconds, nanoseconds, etc.), nor on the cost of cutting-edge speed technology (if speed costs get smaller over time there is simply more entry). The arms race is thus an equilibrium constant of the market design—a result that ties in closely with our empirical findings.

The third and final part of our article shows that frequent batch auctions directly address the problems we have identified with the continuous limit order book. Frequent batch auctions may sound like a very different market design from the continuous limit order book, but there are really just two differences. First, time is treated as a discrete variable instead of a continuous variable.[6] Second, orders are processed in batch instead of serial—since multiple orders can arrive at the same (discrete) time—using a standard uniform-price auction. All other design details are similar. For instance, orders consist of a price, quantity, and direction and can be canceled or modified at any time; priority is price then (discrete) time; there is a well-defined bid-ask spread; and information policy is analogous: orders are

trading costs are lower today than in the pre-HFT era, though most of the benefits appear to have been realized in the late 1990s and early 2000s (see Virtu 2014, p. 103; Angel, Harris, and Spatt 2015, p. 23; Frazzini, Israel, and Moskowitz 2012, table IV). Rather, our claim is that markets are less liquid today than they would be under an alternative market design that eliminated sniping. For further discussion see Section VI.E.

[6] This article does not characterize a specific optimal batch interval. See Section VII.D and Online Appendix B.3 for a discussion of what the present paper's analysis does and does not teach us about the optimal batch interval.

received by the exchange, processed by the exchange (at the end of the discrete time interval, as opposed to continuously), and only then announced publicly.

Together, the two key design differences—discrete time, and batch processing using a uniform-price auction—have two beneficial effects. First, discrete time substantially reduces the value of a tiny speed advantage, which eliminates the arms race. In the continuous-time market, if one trader is even 100 microseconds faster than the next, then any time there is a public price shock the faster trader wins the race to respond. In the discrete-time market, such a small speed advantage almost never matters. Formally, if the batch interval is τ, then a δ speed advantage is only $\frac{\delta}{\tau}$ as likely to matter as in the continuous-time market. So, if the batch interval is 100 milliseconds, then a 100 microsecond speed advantage is $\frac{1}{1000}$ as important. Second, and more subtly, the auction eliminates sniping by transforming the nature of competition. In the continuous market, it is possible to earn a rent based on a piece of information that many traders observe at basically the same time (e.g., a jump in ES), because orders are processed in serial and someone is always first. In the frequent batch auction market, by contrast, if multiple traders observe the same information at the same time, they are forced to compete on price instead of speed. It is no longer possible to earn a rent from symmetrically observed public information.

For both of these reasons, frequent batch auctions eliminate the cost of liquidity provision in continuous limit order book markets associated with stale quotes getting sniped. Intuitively, discrete time reduces the likelihood that a tiny speed advantage yields asymmetric information, and the auction ensures that symmetric information does not generate arbitrage rents. Batching also resolves the prisoner's dilemma caused by the continuous market, and in a manner that allocates the welfare savings to investors. In equilibrium, relative to the continuous limit order book, frequent batch auctions eliminate sniping, enhance liquidity, and stop the HFT arms race.

We emphasize that the market design perspective we take in this article sidesteps the "is HFT good or evil?" debate which seems to animate much of the current discussion about HFT among policy makers, the press, and market microstructure researchers. The market design perspective assumes that market participants optimize with respect to market rules as given, but takes seriously the possibility that the rules themselves are flawed. Many of the negative aspects of HFT that have garnered so much public attention are best understood as symptoms of flawed market design. However, the policy ideas that have been most prominent in response to concerns about HFT—for example, Tobin taxes, minimum resting times, message limits—attack symptoms rather than address the root market design flaw: continuous-time, serial-process trading. Frequent batch auctions directly address the root flaw.

The rest of the article is organized as follows. Section II discusses related literature. Section III briefly reviews the rules of the continuous limit order book. Section IV describes our direct-feed data from NYSE and the CME. Section V presents the empirical results on correlation breakdown and mechanical arbitrage. Section VI presents the model and solves for and discusses the equilibrium of the continuous limit order book. Section VII analyzes frequent batch auctions, shows why they directly address the problems with the continuous market, and discusses their equilibrium properties. Section VIII uses our model to discuss alternative proposed responses to the HFT arms race. Section IX discusses computational advantages of discrete-time trading over continuous-time trading. Section X concludes. Online Appendix A provides backup materials for the empirical analysis. Online Appendix B provides proofs and other backup materials for the theoretical analysis.

II. Related Literature

First, there is a well-known older academic literature on infrequent batch auctions, for instance, three times per day (open, midday, and close). Important contributions to this literature include Cohen and Schwartz (1989), Madhavan (1992), and Economides and Schwartz (1995); see also Schwartz (2001) for a book treatment. We emphasize that the arguments for infrequent batch auctions in this earlier literature are completely distinct from the arguments we make for frequent batch auctions. Our argument focuses on eliminating sniping, encouraging competition on price rather than speed, and stopping the arms race. The earlier literature focused on enhancing the accuracy of price discovery by aggregating the dispersed information of investors into a single price,[7] and reducing intermediation costs by enabling investors to trade with each other directly. Perhaps the simplest way to think about the relationship between our work and this earlier literature is as follows. Our work shows that there is a discontinuous welfare and liquidity benefit from moving from continuous time to discrete time—more precisely, from the continuous-time serial-process limit order book market to discrete-time batch-process auctions. The earlier literature suggests that there might be additional further benefits to greatly lengthening the batch interval that are outside our model. But there are also likely to be important costs to such lengthening that are outside our model and outside the models of this earlier literature as well. Developing a richer understanding of the costs of lengthening the time between auctions is an important topic for future research.

We also note that our specific market design details differ from those in this earlier literature, beyond simply the frequency with which the auctions are conducted. Differences include information policy, the treatment of unexecuted orders, and time priority rules; see Section VII.A for a full description.

Second, there are two recent papers, developed independently and contemporaneously[8] from ours and coming from different methodological perspectives, that also make cases for frequent batch auctions. Closest in spirit is Farmer and Skouras (2012), a policy paper commissioned by the UK Government's Foresight Report. They, too, argue that continuous trading leads to an arms race for speed, and that frequent batch auctions stop the arms race. There are three substantive differences between our arguments. First, two conceptually important ideas that come out of our formal model are that arbitrage rents are built in to the continuous limit order book market design, in the sense that even symmetrically observed public information creates arbitrage opportunities due to serial processing, and that the auction eliminates these rents by transforming competition on speed into competition on price. These two ideas are not identified in Farmer and Skouras (2012). Second, the details of our proposed market designs are substantively different. Our theory

[7] In Economides and Schwartz (1995), the aggregation is achieved by conducting the auction at three significant points during the trading day (open, midday, and close). In Madhavan (1992) the aggregation is achieved by waiting for a large number of investors with both private- and common-value information to arrive to market.

[8] We began work on this project in October 2010.

identifies the key flaws of the continuous limit order book and shows that these flaws can be corrected by modifying only two things: time is treated as discrete instead of continuous, and orders are processed in batch using an auction instead of serially. Farmer and Skouras (2012) depart more dramatically from the continuous limit order book, demarcating time using an exponential random variable and entirely eliminating time-based priority.[9] Last, a primary concern of Farmer and Skouras (2012) is market stability, a topic we touch on only briefly in Section IX. Wah and Wellman (2013) make a case for frequent batch auctions using a zero-intelligence (i.e., non–game theoretic) agent-based simulation model. In their simulation model, investors have heterogeneous private values (costs) for buying (selling) a unit of a security, and use a mechanical strategy of bidding their value (or offering at their cost). Batch auctions enhance efficiency in their setup by aggregating supply and demand and executing trades at the market-clearing price. Note that this is a similar argument in favor of frequent batch auctions as that associated with the older literature on infrequent batch auctions referenced previously. The reason that this force pushes towards frequent batch auctions in Wah and Wellman (2013) is that their simulations utilize an extremely high discount rate of 6 basis points per millisecond.

Third, our paper relates to the burgeoning academic literature on high-frequency trading; see Jones (2013), Biais and Foucault (2014), and O'Hara (2015) for recent surveys. One focus of this literature has been on the empirical study of the effect of high-frequency trading on market quality, within the context of the current market design. Examples include Hendershott, Jones, and Menkveld (2011); Hasbrouck and Saar (2013); Brogaard, Hendershott, and Riordan (2014a,b); Foucault, Kozhan, and Tham (2014); and Menkveld and Zoican (2014). We discuss the relationship between our results and aspects of this literature in Section VI.E. Biais, Foucault, and Moinas (2015) study the equilibrium level of investment in speed technology in the context of a Grossman-Stiglitz style rational expectations

[9] There have been several other white papers and essays making cases for frequent batch auctions, which to our knowledge were developed independently and roughly contemporaneously: Cinnober (2010); Sparrow (2012); McPartland (2015); ISN (2013); Schwartz and Wu (2013). In each case, either the proposed market design departs more dramatically from the continuous limit order book than ours or important design details are omitted.

model. They find that investment in speed can be socially excessive, as we do in our model, and argue for a Pigovian tax on speed technology as a policy response; see Section VIII.A for discussion of the traditional Tobin tax and the Biais, Foucault, and Moinas (2015) tax. The Nasdaq “SOES bandits” were an early incarnation of stale-quote snipers, in the context of a part-human part-electronic market design that had an unusual feature that was exploited by the bandits—the prohibition of automated quote updates, which necessitated costly and imperfect human monitoring. See Foucault, Roell, and Sandas (2003) for a theoretical analysis and Harris and Schultz (1998) for empirical facts. Further discussion of other related work from this literature is incorporated into the body of the article.

Fourth, this article is in the tradition of the academic literature on market design. This literature focuses on designing the “rules of the game” in real-world markets to achieve objectives such as economic efficiency. Examples of markets that have been designed by economists include auction markets for wireless spectrum licenses and the market for matching medical school graduates to residency positions. See Klemperer (2004) and Milgrom (2004, 2011) for surveys of the market design literature on auction markets and Roth (2002, 2008) for surveys on the market design literature on matching markets. Some papers in this literature that are conceptually related to ours, albeit focused on different market settings, are Roth and Xing (1994) on the timing of transactions, Roth and Xing (1997) on serial versus batch processing, and Roth and Ockenfels (2002) on bid sniping.

Last, several of the ideas in our critique of the continuous limit order book are new versions of classical ideas. Correlation breakdown is an extreme version of a phenomenon first documented by Epps (1979); see Section V for further discussion. Sniping, and its negative effect on liquidity, is closely related to Glosten and Milgrom (1985) adverse selection; see Section VI.C, which discusses this relationship in detail. The idea that financial markets can induce inefficient speed competition traces at least to Hirshleifer (1971); in fact, our model clarifies that in the continuous market fast traders can earn a rent even from information that they observe at exactly the same time as other fast traders, which can be viewed as the logical extreme of what Hirshleifer (1971) called “foreknowledge” rents.

III. Brief Description of the Continuous Limit Order Book

In this section we summarize the rules of the continuous limit order book market design. Readers familiar with these rules can skip this section. Readers interested in further details should consult Harris (2002).

The basic building block of this market design is the limit order. A limit order specifies a price, a quantity, and whether the order is to buy or sell, for example, "buy 100 shares of XYZ at $100.00." Traders may submit limit orders to the market at any time during the trading day, and they may fully or partially withdraw their outstanding limit orders at any time.

The set of limit orders outstanding at any particular moment is known as the limit order book. Outstanding orders to buy are called bids and outstanding orders to sell are called asks. The difference between the best (highest) bid and the best (lowest) ask is known as the bid-ask spread.

Trade occurs whenever a new limit order is submitted that is either a buy order with a price weakly greater than the current best ask or a sell order with a price weakly smaller than the current best bid. In this case, the new limit order is interpreted as either fully or partially accepting one or more outstanding asks. Orders are accepted in order of the attractiveness of their price, with ties broken based on which order has been in the book the longest; this is known as price-time priority. For example, if there are outstanding asks to sell 1,000 shares at $100.01 and 1,000 shares at $100.02, a limit order to buy 1,500 shares at $100.02 (or greater) would get filled by trading all 1,000 shares at $100.01, and then by trading the 500 shares at $100.02 that have been in the book the longest. A limit order to buy 1,500 shares at $100.01 would get partially filled, by trading 1,000 shares at $100.01, with the remainder of the order remaining outstanding in the limit order book (500 shares at $100.01).

Observe that order submissions and order withdrawals are processed by the exchange in serial, that is, one at a time in order of their receipt. This serial-processing feature of the continuous limit order book plays an important role in the theoretical analysis in Section VI.

In practice, there are many other order types that traders can use in addition to limit orders. These include market orders, stop-loss orders, immediate-or-cancel, and dozens of others that are considerably more obscure (e.g., Patterson and Strasburg 2012; Nanex 2012). These alternative order types are ultimately just proxy instructions to the exchange for the generation of limit orders. For instance, a market order is an instruction to the exchange to place a limit order whose price is such that it executes immediately, given the state of the limit order book at the time the message is processed.

IV. Data

We use "direct-feed" data from the Chicago Mercantile Exchange (CME) and New York Stock Exchange (NYSE). Direct-feed data record all activity that occurs in an exchange's limit order book, message by message, with millisecond resolution timestamps assigned to each message by the exchange at the time the message is processed.[10] Practitioners who demand the lowest latency data (e.g., high-frequency traders) use this direct-feed data in real time to construct the limit order book.

The CME data set is called CME Globex DataMine Market Depth. Our data cover all limit order book activity for the E-mini S&P 500 Futures Contract (ticker ES) over the period of January 1, 2005–December 31, 2011. The NYSE data set is called TAQ NYSE ArcaBook. While this data covers all U.S. equities traded on NYSE, we focus most of our attention on the SPDR S&P 500 exchange traded fund (ticker SPY). Our data cover the period of January 1, 2005–December 31, 2011, with the exception of a three-month gap from 5/30/2007 to 8/28/2007 resulting from data issues acknowledged to us by the NYSE data team. We also drop, from both data sets, the Thursday and Friday from the week prior to expiration for every ES expiration month (March, June, September, December) due to the rolling over of the front month contract, half days (e.g., day after Thanksgiving), and a small number of days in which either data set's zip file is corrupted or truncated. We are left with 1,560 trading days in total.

[10] Prior to November 2008, the CME datafeed product did not populate the millisecond field for time stamps, so the resolution was actually centisecond not millisecond. CME recently announced that the next iteration of its datafeed product will be at microsecond resolution.

Each message in direct-feed data represents a change in the order book at that moment in time. It is the subscriber's responsibility to construct the limit order book from this feed, maintain the status of every order in the book, and update the internal limit order book based on incoming messages. In order to interpret raw data messages reported from each feed, we write a feed parser for each raw data format and update the state of the order book after every new message.

We emphasize that direct feed data are distinct from the consolidated feeds that aggregate data from individual exchanges. In particular, the TAQ NYSE ArcaBook data set is distinct from the more familiar TAQ NYSE Daily data set (sometimes simply referred to as TAQ), which is an aggregation of orders and trades from all Consolidated Tape Association exchanges. The TAQ data is comprehensive in regard to trades and quotes listed at all participant exchanges, which includes the major electronic exchanges BATS, NASDAQ, and NYSE and also small exchanges such as the Chicago Stock Exchange. However, practitioners estimate that the TAQ's timestamps are substantially delayed relative to the direct-feed data that comes directly from the exchanges (our own informal comparisons confirm this; see also Ding, Hanna, and Hendershott 2014). One source of delay is that the TAQ's timestamps do not come directly from the exchanges' order matching engines. A second source of delay is the aggregation of data from several different exchanges, with the smaller exchanges considered especially likely to be a source of delay. The key advantage of our direct-feed data is that the time stamps are as accurate as possible. In particular, these are the same data that HFT firms subscribe to and process in real time to make trading decisions.

V. Correlation Breakdown and Mechanical Arbitrage

In this section we report two sets of stylized facts about how continuous limit order book markets behave at high-frequency time horizons. First, we show that correlations completely break down at high-enough frequency. Second, we show that there are frequent mechanical arbitrage opportunities associated with this correlation breakdown, which are available to whichever trader acts fastest.

For each result we first present summary statistics and then explore how the phenomenon has evolved over the time period of our data, 2005–2011. The summary statistics give a sense of magnitudes for what we depicted anecdotally in Figure I. The time-series evidence suggests that correlation breakdown and mechanical arbitrage are intrinsic features of the continuous limit order book market, rather than market failures that are competed away over time.

Before proceeding, we emphasize that the finding that correlations break down at high-enough frequency—which is an extreme version of a phenomenon discovered by Epps (1979)[11]—is obvious from introspection alone, at least ex post. There is nothing in current market architecture—in which each financial instrument trades in continuous time on its own separate limit-order book, rather than in a single combinatorial auction market—that would allow different instruments' prices to move at exactly the same time.

V.A. Correlation Breakdown

1. Summary Statistics. Figure II displays the median, min, and max daily return correlation between ES and SPY for time intervals ranging from 1 millisecond to 60 seconds, for our 2011 data, under our main specification for computing correlation. In this main specification, we compute the correlation of percentage changes in the equal-weighted midpoint of the ES and SPY bid and ask, and ignore speed-of-light issues. As can be seen from the figure, the correlation between ES and SPY is nearly 1 at long-enough intervals,[12] but breaks down at high-frequency time intervals. The 10 millisecond correlation is just 0.1016, and the 1 millisecond correlation is just 0.0080.

We consider several other specifications for computing the ES-SPY correlation in Online Appendix A.1.1. We also examine correlations for pairs of related equity securities in Online

[11] Epps (1979) found that equity market correlations among stocks in the same industry (e.g., Ford-GM) were much lower over short time intervals than over longer time intervals; in that era, "very short" meant 10 minutes, and long meant a few days.

[12] It may seem surprising at first that the ES-SPY correlation does not approach 1 even faster. An important issue to keep in mind, however, is that ES and SPY trade on discrete price grids with different tick sizes: ES tick sizes are 0.25 index points, whereas SPY tick sizes are 0.10 index points. As a result, small changes in the fundamental value of the S&P 500 index manifest differently in the two markets, due to what are essentially rounding issues. At long time horizons these rounding issues are negligible relative to changes in fundamentals, but at shorter frequencies these rounding issues are important, and keep correlations away from 1.

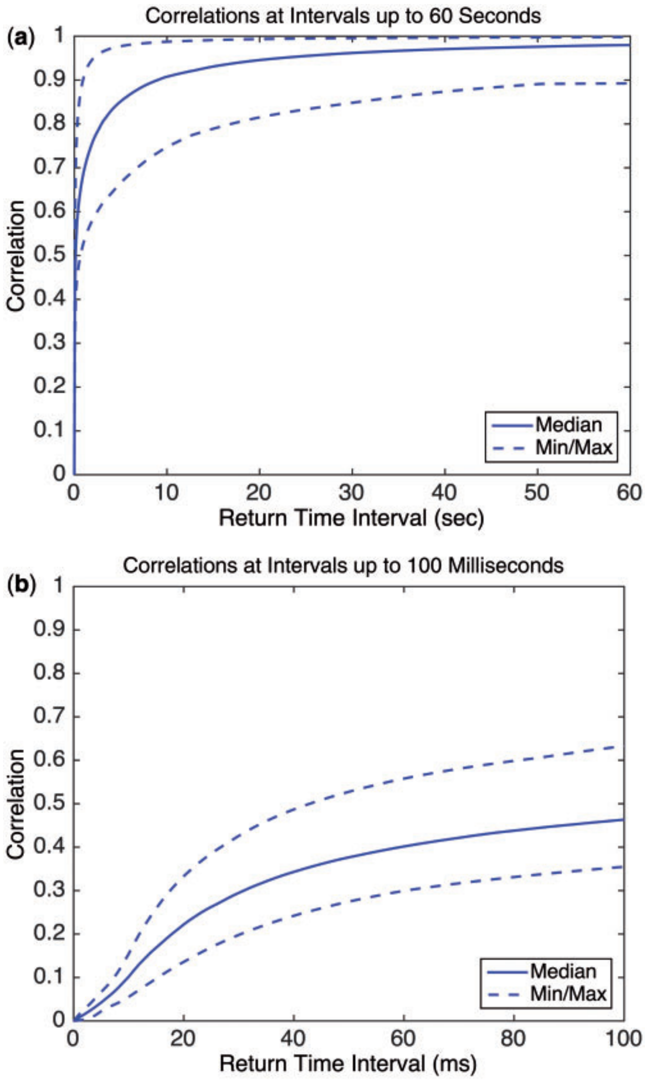

FIGURE II

ES and SPY Correlation by Return Interval: 2011

This figure depicts the correlation between the return of the E-mini S&P 500 future (ES) and the SPDR S&P 500 ETF (SPY) bid-ask midpoints as a function of the return time interval in 2011. The solid line is the median correlation over all trading days in 2011 for that particular return time interval. The dotted lines represent the minimum and maximum correlations over all trading days in 2011 for that particular return time interval. Panel A shows a range of time intervals from 1 to 60,000 milliseconds (ms) or 60 seconds. Panel B shows that same picture but zoomed in on the interval from 1 to 100 ms. For more details on the data, refer to Section IV.

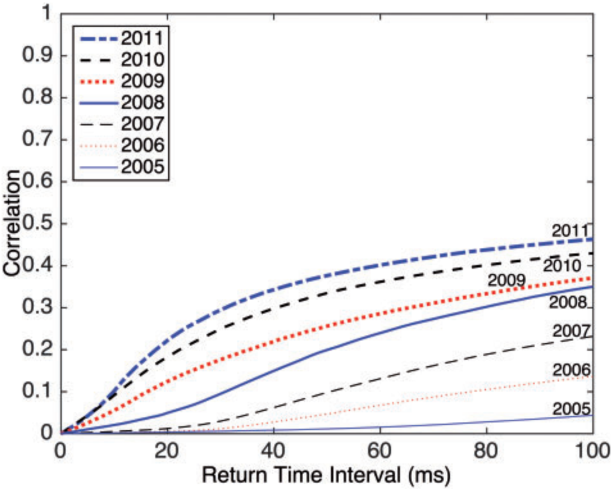

Figure III

ES and SPY Correlation Breakdown over Time: 2005–2011

This figure depicts the correlation between the return of the E-mini S&P 500 future (ES) and the SPDR S&P 500 ETF (SPY) bid-ask midpoints as a function of the return time interval for every year from 2005 to 2011. Each line depicts the median correlation over all trading days in a particular year, taken over each return time interval from 1 to 100 milliseconds. For 2005–2008 the CME data is only at 10 milliseconds resolution, so we compute the median correlation for each multiple of 10 milliseconds and then fit a cubic spline. For more details on the data, refer to Section IV.

Appendix A.1.2, for which speed-of-light issues do not arise since they trade in the same physical location. In all cases, correlations break down at high frequency.

2. Correlation Breakdown over Time. Figure III displays the ES-SPY correlation versus time interval curve that we depicted above as II, Panel b, but separately for each year in the time period 2005–2011 that is covered in our data. As can be seen in the figure, the market has gotten faster over time in the sense that economically meaningful correlations emerge more quickly in the later years of our data than in the earlier years. For instance, in 2011 the ES-SPY correlation reaches 0.50 at a 142-millisecond interval, whereas in 2005 the ES-SPY correlation only reaches 0.50 at a 2.6-second interval. However, in all years correlations are essentially zero at high enough frequency.

V.B. Mechanical Arbitrage

1. Computing the ES-SPY Arbitrage. Conceptually, our goal is to identify all of the ES-SPY arbitrage opportunities in our data in the spirit of the example shown in Figure I, Panel D—buy cheap and sell expensive when one instrument has jumped and the other has yet to react—and for each such opportunity measure its profitability and duration. The full details of our method for doing this are in Online Appendix A.2.1. Here, we mention the most important points.

First, there is a difference in levels between the two instruments, called the spread. The spread arises from three sources: ES is larger than SPY by a term that represents the carrying cost of the S&P 500 index until the ES contract's expiration date; SPY is larger than ES by a term that represents S&P 500 dividends, which SPY holders receive and ES holders do not; and the basket of stocks in the ETF typically differs slightly from the basket of stocks in the S&P 500 index, called ETF tracking error. Our arbitrage computation assumes that at high-frequency time horizons, changes in the ES-SPY spread are mostly driven not by changes in these persistent factors but instead by temporary noise, that is, by correlation breakdown. We then assess the validity of this assumption empirically by classifying as "bad arbs" anything that looks like an arbitrage opportunity to our computational procedure but turns out to be a persistent change in the level of the ES-SPY spread, for example, due to a change in short-term interest rates.

Second, while Figure I depicts bid-ask midpoints, in computing the arbitrage opportunity we assume that the trader buys the cheaper instrument at its ask while selling the more expensive instrument at its bid (with cheap and expensive defined relative to the difference in levels). That is, the trader pays bid-ask spread costs in both markets.[13] Our arbitrageur only initiates a trade when the expected profit from doing so, accounting for bid-ask

[13] This is a simple and transparent estimate of transactions costs. A richer estimate would account for the fact that the trader might not need to pay half the bid-ask spread in both ES and SPY, which would lower costs, and would account for exchange fees and rebates, which on net would increase costs. As an example, a high-frequency trader who detects a jump in the price of ES that makes the price of SPY stale might trade instantaneously in SPY at the stale prices, paying half the bid-ask spread plus an exchange fee, but might seek to trade in ES at its new price as a liquidity provider, in which case he would earn rather than pay half the bid-ask spread.

spread costs, exceeds a modest profitability threshold of 0.05 index points (one-half of one penny in the market for SPY). If the jump in ES or SPY is sufficiently large that the arbitrageur can profitably trade through multiple levels of the book net of costs and the threshold, then he does so.

Third, we only count arbitrage opportunities that last at least 4 milliseconds, the one-way speed-of-light travel time between New York and Chicago. Arbitrage opportunities that last fewer than 4 milliseconds are not exploitable under any possible technological advances in speed (other than by a god-like arbitrageur who is not bound by special relativity). Therefore, such opportunities should not be counted as part of the prize that high-frequency trading firms are competing for, and we drop them from the analysis.

2. Summary Statistics. Table I reports summary statistics on the ES-SPY arbitrage opportunity over our full data set, 2005–2011.

An average day in our data set has about 800 arbitrage opportunities, while an average arbitrage opportunity has quantity of 14 ES lots (7,000 SPY shares) and profitability of 0.09 in index points (per unit traded) and $98.02 in dollars. The 99th percentile of arbitrage opportunities has a quantity of 145 ES lots (72,500 SPY shares) and profitability of 0.22 in index points and $927.07 in dollars.

Total daily profits in our data are on average $79,000 per day, with profits on a 99th percentile day of $554,000. Since our SPY data come from just one of the major equities exchanges, and depth in the SPY book is the limiting factor in terms of quantity traded for a given arbitrage in nearly all instances (typically the depths differ by an order of magnitude), we also include an estimate of what total ES-SPY profits would be if we had SPY data from all exchanges and not just NYSE. We do this by multiplying each day's total profits based on our NYSE data by a factor of (1 / NYSE's market share in SPY), with daily market share data sourced from Bloomberg.[14] This yields average profits of $306,000 a day, or roughly $75 million a year. We discuss the

[14] NYSE's daily market share in SPY has a mean of 25.9 percent over the time period of our data, with mean daily market share highest in 2007 (33.0 percent) and lowest in 2011 (20.4 percent). Most of the remainder of the volume is split between the other three largest exchanges, NASDAQ, BATS, and DirectEdge.

TABLE I

ES-SPY Arbitrage Summary Statistics, 2005–2011

		Percentile						
	Mean	1	5	25	50	75	95	99
# of arbs/day	801	118	173	285	439	876	2498	5353
Per arb quantity (ES lots)	13.83	0.20	0.20	1.25	4.20	11.99	52.00	145.00
Per arb profits (index pts)	0.09	0.05	0.05	0.06	0.08	0.11	0.15	0.22
Per arb profits ($)	98.02	0.59	1.08	5.34	17.05	50.37	258.07	927.07
Total daily profits, NYSE data ($)	79 k	5 k	9 k	18 k	33 k	57 k	204 k	554 k
Total daily profits, all exchanges ($)	306 k	27 k	39 k	75 k	128 k	218 k	756 k	2,333 k
% ES initiated	88.56							
% good arbs	99.99							
% buy vs. sell	49.77							

Notes. This table shows the mean and various percentiles of arbitrage variables from the mechanical trading strategy between the E-mini S&P 500 future (ES) and the SPDR S&P 500 ETF (SPY) described in Section V.B and Online Appendix A.2.1. The data, described in Section IV, cover January 2005 to December 2011. Variables are described in the text of Section V.B.

total size of the arbitrage opportunity in more detail below in Section V.C.

The majority (88.56 percent) of the arbitrage opportunities in our data set are initiated by a price change in ES, with the remaining 11.44 percent initiated by a price change in SPY. That the large majority of arbitrage opportunities are initiated by ES is consistent with the practitioner perception that the ES market is the center for price discovery in the S&P 500 index, as well as with our finding in Online Appendix Table A.1 that correlations are higher when we treat the New York market as lagging Chicago than when we treat the Chicago market as lagging New York or treat the two markets equally.

Nearly all (99.99 percent) of the arbitrage opportunities we identify are "good arbs," meaning that deviations of the ES-SPY spread from our estimate of fair value that are large enough to trigger an arbitrage nearly always reverse within a modest amount of time. This is one indication that our method of computing the ES-SPY arbitrage opportunity is sensible.

3. Mechanical Arbitrage over Time: 2005–2011. In this subsection we explore how the ES-SPY arbitrage opportunity has evolved over time.

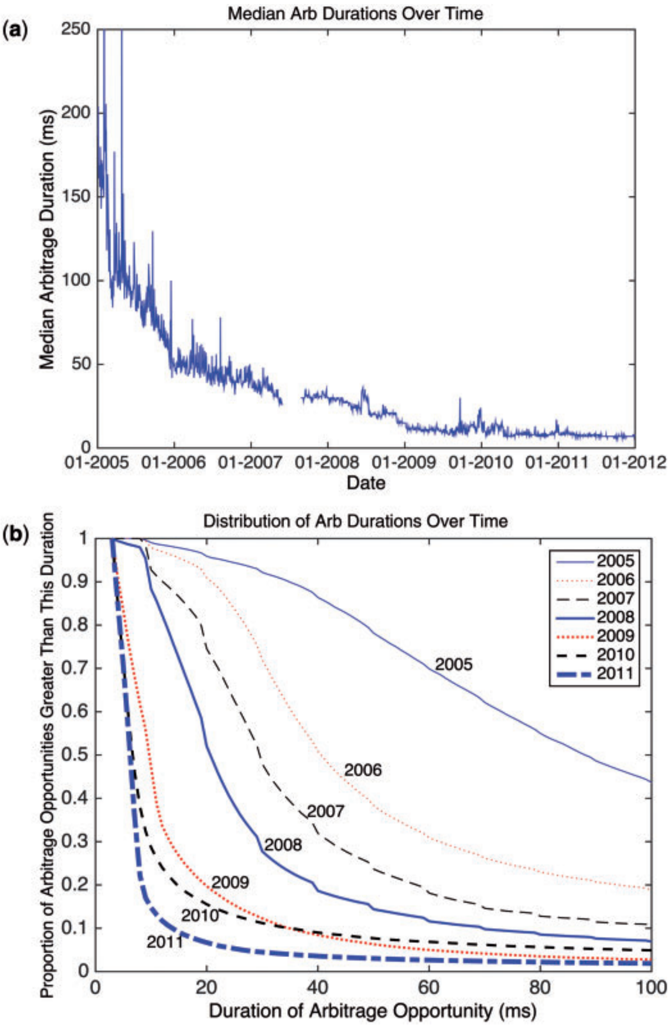

FIGURE IV

Duration of ES & SPY Arbitrage Opportunities over Time: 2005–2011

Panel A shows the median duration of ES-SPY arbitrage opportunities for each day in our data. Panel B plots arbitrage duration against the proportion of opportunities lasting at least that duration, for each year in our data. We drop opportunities that last fewer than 4 milliseconds, the speed-of-light travel time between New York and Chicago. Prior to November 24, 2008, we drop opportunities that last fewer than 9 milliseconds, the maximum combined effect of the speed-of-light travel time and the rounding of CME data to centiseconds. See Section V.B for details regarding the arbitrage. See Section IV for details regarding the data.

Figure IV explores the duration of ES-SPY arbitrage opportunities over the time of our data set, covering 2005–2011. As can be seen in Panel A, the median duration of arbitrage opportunities has declined dramatically over this time period, from a median of 97 milliseconds in 2005 to a median of 7 milliseconds in 2011. Panel B plots the distribution of arbitrage durations over time, asking what proportion of arbitrage opportunities last at least a certain amount of time, for each year in our data. The figure conveys how the speed race has steadily raised the bar for how fast one must be to capture arbitrage opportunities. For instance, in 2005 nearly all arbitrage opportunities lasted at least 10 milliseconds and most lasted at least 50 milliseconds, whereas by 2011 essentially none lasted 50 milliseconds and very few lasted even 10 milliseconds.

Figure V explores the per arbitrage profitability of ES-SPY arbitrage opportunities over the time of our data set. In contrast to arbitrage durations, arbitrage profits have remained remarkably constant over time. Panel A shows that the median profits per contract traded have remained steady at around 0.08 index points, with the exception of the 2008 financial crisis when they were a bit larger. Panel B shows that the distribution of profits has also remained relatively stable over time, again with the exception of the 2008 financial crisis where the right tail of profit opportunities is noticeably larger.

Figure VI explores the frequency of ES-SPY arbitrage opportunities over the time of our data set. Unlike per arb profitability, the frequency of arbitrage opportunities varies considerably over time. Figure VI, Panel A shows that the median arbitrage frequency seems to track the overall volatility of the market, with frequency especially high during the financial crisis in 2008, the Flash Crash on May 6, 2010, and the European crisis in summer 2011. This makes intuitive sense: when the market is more volatile, there are more arbitrage opportunities because there are more jumps in one market that leave prices temporarily stale in the other market. Panel B confirms this intuition formally. The figure plots the number of arbitrage opportunities on a given trading day against a simple proxy of that day's volatility we call distance traveled, defined as the sum of the absolute-value of changes in the ES midpoint price over the course of the trading day. This one simple variable explains nearly all of the variation in the number of arbitrage opportunities per day: the R^2 of the

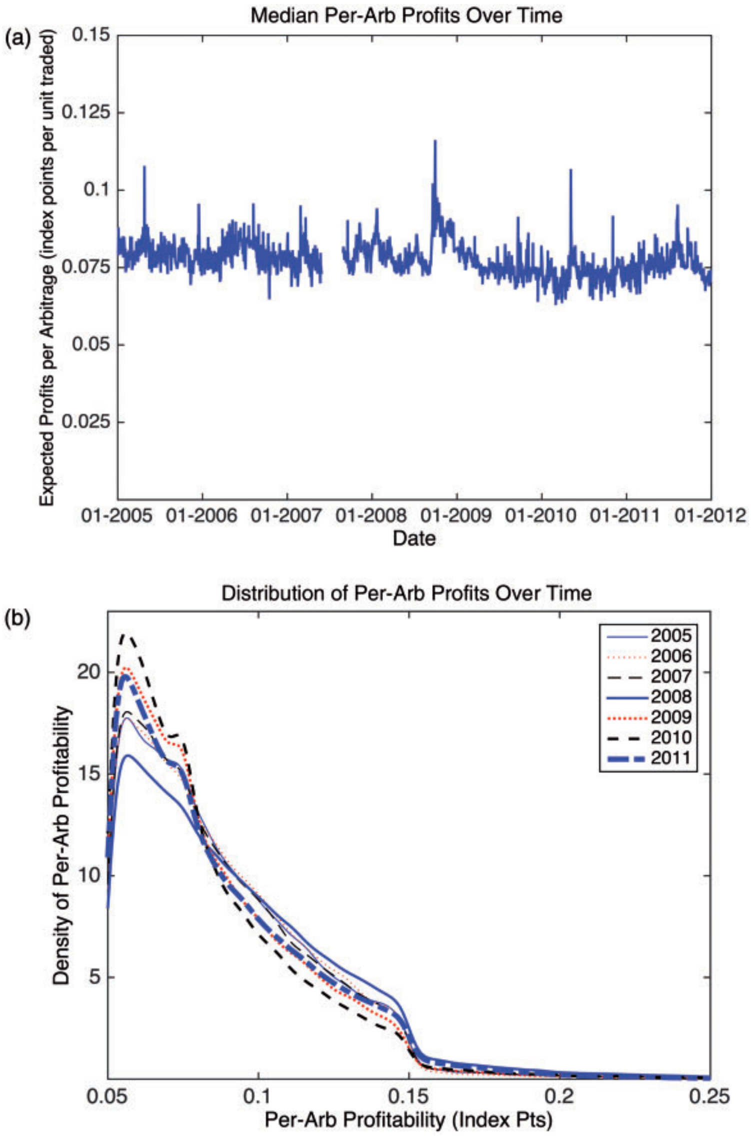

FIGURE V

Profitability of ES & SPY Arbitrage Opportunities over Time: 2005–2011

Panel A shows the median profitability of ES-SPY arbitrage opportunities, per unit traded, for each day in our data. Panel B plots the kernel density of the profitability of arbitrage opportunities, per unit traded, for each year in our data. See Section V.B for details regarding the ES-SPY arbitrage. See Section IV for details regarding the data.

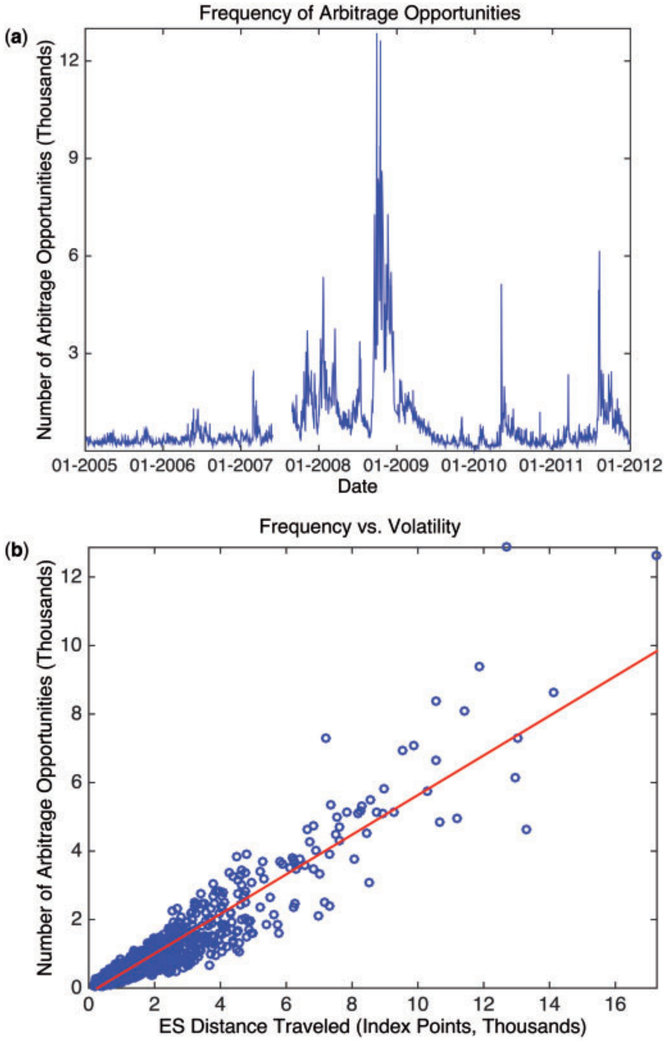

FIGURE VI

Frequency of ES & SPY Arbitrage Opportunities over Time: 2005–2011

Panel A shows the time series of the total number of ES-SPY arbitrage opportunities in each day in our data. Panel B depicts a scatter plot of the total number of arbitrage opportunities in a trading day against ES distance traveled, defined as the sum of the absolute value of changes in the ES midpoint price over the course of the trading day. The solid line represents the fitted values from a linear regression of arbitrage frequency on distance traveled. See Section V.B for details regarding the arbitrage. See Section IV for details regarding the data.

regression of daily arbitrage frequency on daily distance traveled is 0.87.

Together, the results depicted in Figures IV, V, and VI suggest that the ES-SPY arbitrage opportunity should be thought of more as a mechanical "constant" of the continuous limit order book market than as a profit opportunity that is competed away over time. Competition has clearly reduced the amount of time that arbitrage opportunities last (Figure IV), but the size of arbitrage opportunities has remained remarkably constant (Figure V), and the frequency of arbitrage opportunities seems to be driven mostly by market volatility (Figure VI). Figure III, on the time series of correlation breakdown, reinforces this story: competition has increased the speed with which information from Chicago prices is incorporated into New York prices and vice versa (the analogue of Figure IV), but competition has not fixed the root issue that correlations break down at high enough frequency (the analogue of Figure V).

These facts both inform and are explained by our model in Section VI.

V.C. Discussion

In this section, we make two remarks about the size of the prize in the speed race.

First, we suspect that our estimate of the annual value of the ES-SPY arbitrage opportunity—an average of around \$75 million per year, fluctuating as high as \$151 million in 2008 (the highest volatility year in our data) and as low as \$35 million in 2005 (the lowest volatility year in our data)—is an underestimate for at least three reasons. (i) Our trading strategy is extremely simplistic. This simplicity is useful for transparency of the exercise and for consistency when we examine how the arbitrage opportunity has evolved over time, but it is likely that there are more sophisticated trading strategies that produce higher profits. (ii) Our trading strategy involves transacting at market in both ES and SPY, which means paying half the bid-ask spread in both markets. An alternative approach that economizes on transactions costs is to transact at market only in the instrument that lags—for example, if ES jumps, transact at market in SPY but not in ES. Since 89 percent of our arbitrage opportunities are initiated by a jump in ES, and the minimum ES bid-ask spread is substantially larger than the minimum SPY bid-ask spread (0.25 index points versus 0.10 index points), the transactions cost savings from this approach can be meaningful. (iii) Our CME data consist of all of the messages that are transmitted publicly to CME data feed subscribers, but we do not have access to the trade notifications that are transmitted privately to the parties involved in a particular trade. It has recently been reported (Patterson, Strasburg, and Pleven 2013) that the public message feed lags private trade notifications by an average of several milliseconds, because of the way the CME processes message notifications. This lag could cause us to miss profitable trading opportunities; in particular, we worry that we are especially likely to miss some of the largest trading opportunities, since large jumps in ES triggered by large orders in ES also will trigger the most trade notifications, and hence the most lag.

Second, and more important, ES-SPY is just the tip of the iceberg in the race for speed. We are aware of at least five categories of speed races analogous to ES-SPY. (i) There are hundreds of trades substantially similar to ES-SPY, consisting of exchange-traded instruments that are highly correlated and with sufficient liquidity to yield meaningful profits from simple mechanical arbitrage strategies. Figure A.2 in the Online Appendix provides an illustrative partial list.[15] (ii) Because equity markets are fragmented (the same security trades on multiple exchanges) there are trades even simpler than ES-SPY. For instance, one can arbitrage SPY on NYSE against SPY on NASDAQ (or BATS, dark pools, etc.). We are unable to detect such trades because the latency between equities exchanges—all of whose servers are located in data centers in New Jersey—is measured in microseconds, which is finer than the current resolution of researcher-available exchange data. (iii) Instruments that are meaningfully correlated, but with correlation far from 1, can also be traded in a manner analogous to ES-SPY. For instance, even though the Goldman Sachs–Morgan Stanley correlation is far from 1, a large jump in GS may be sufficiently informative about the price of MS that it induces a race to react in the market for MS. As we show in Online Appendix A.1.2, the

[15] In equities data downloaded from Yahoo! Finance, we found 391 pairs of equity securities with daily returns correlation of at least 0.90 and average daily trading volume of at least $100 million per security (calendar year 2011). It has not yet been possible to perform a similar screen on the universe of all exchange-traded financial instruments, including, for example, index futures, commodities, bonds, currencies, etc., due to data limitations. Instead, we include illustrative examples across all instrument types in Online Appendix Figure A.2.

equities market correlation matrix breaks down at high frequency, suggesting that such trading opportunities—whether they involve pairs of stocks or simple statistical relationships among sets of stocks—may be important. (iv) There is a race to respond to public news events such as Fed announcements, the release of important government statistics, the posting of corporate SEC filings, and so on. In this race, the precise effect of the public news on asset prices is often hard to determine at high frequency, but the sign and rough magnitude of the news can be determined quickly (Rogers, Skinner, and Zechman 2014). (v) In addition to the race to snipe stale quotes, there is also a race among liquidity providers to the top of the book (see Moallemi 2014; Yao and Ye 2014). This last race is an artifact of the minimum tick increment imposed by regulators and/or exchanges.

While we hesitate, in the context of the present article, to put a precise estimate on the total prize at stake in the arms race, back-of-the-envelope extrapolation from our ES-SPY estimates suggests that the annual sums are substantial.

VI. Model: Critique of the Continuous Limit Order Book

We have established three empirical facts about continuous limit order book markets. First, correlations completely break down at high-enough frequency, even for financial instruments that are nearly perfectly correlated at longer frequencies. Second, this correlation breakdown is associated with frequent mechanical arbitrage opportunities, available to whoever wins the race to exploit them. Third, the prize in the arms race seems to be more like a constant than something that is competed away over time.

We now develop a purposefully simple model that is informed by and helps make sense of these empirical facts. The model ultimately serves two related purposes: it is a critique of the continuous limit order book market design, and it articulates the economics of the HFT arms race.

VI.A. Preliminaries

Security x with Perfect Public Signal y. There is a security x that trades on a continuous limit order book, the rules of which are described in Section III. There is a publicly observable signal y of the value of security x. We make the following purposefully strong assumption: the fundamental value of x is *perfectly* correlated to the public signal y, and, moreover, x can always be costlessly liquidated at this fundamental value. This is a best-case scenario for price discovery and liquidity provision in a continuous limit order book, abstracting from both asymmetric information and inventory costs.

We think of x and y as a metaphor for pairs or sets of exchange-traded financial instruments that are highly correlated. For instance, x is SPY and y is ES. Alternatively, y can be interpreted more abstractly as a publicly observable perfect signal about the value of security x.

The signal y, and hence the fundamental value of security x, evolves as a compound Poisson jump process with arrival rate λ_{jump} and jump distribution F_{jump}. The jump distribution has finite bounded support and is symmetric with mean zero. Let J denote the random variable formed by drawing randomly according to F_{jump}, and then taking the absolute value; we refer to J as the jump size distribution.

Investors and Trading Firms. There are two types of players, investors and trading firms. Both types of players are risk-neutral and there is no discounting.

The players we call investors we think of as the end users of financial markets: mutual funds, pension funds, hedge funds, individuals, etc. Since there is no asymmetric information about fundamentals in our model, our investors could equivalently be called "liquidity traders" as in Glosten and Milgrom (1985) or "noise traders" as in Kyle (1985). Investors arrive stochastically to the market with an inelastic need to either buy or sell a unit of x (we generalize to multiple units in Section VI.B). The arrival process is Poisson with rate λ_{invest}, and, conditional on arrival, it is equally likely that the investor needs to buy versus sell. We assume that all else equal, investors prefer to transact sooner rather than later. Formally, if an investor arrives to market at time t needing to buy one unit, and then buys a unit at time $t' \geq t$ for price p, her payoff is $v + (y_{t'} - p) - f_{delaycost}(t' - t)$, where v is a large positive constant that represents her inelastic need to complete the trade, $y_{t'}$ is the fundamental value of x at the time she trades, and the function $f_{delaycost}(\cdot)$, which is strictly increasing and continuous with $f_{delaycost}(0) = 0$, represents her preference to transact sooner rather than later. If the investor arrives needing to sell, and sells a unit at price p at time t', her payoff is

$v + (p - y_{t'}) - f_{delaycost}(t' - t)$. In the equilibrium of the continuous limit order book we derive in Section VI.B, investors choose to transact immediately. In the equilibria of frequent batch auctions, studied in Section VII, investors will choose to transact in the discrete-time analogue of immediately, namely, at the next available batch auction. Once investors transact, they exit the game.

Trading firms (equivalently HFTs, market makers, algorithmic traders) have no intrinsic demand to buy or sell x. Their goal in trading is simply to buy x at prices lower than y, and to sell x at prices higher than y. If a trading firm buys a share of x at price p at time t, they earn profits from that trade of $y_t - p$; similarly, if they sell a share of x at price p at time t they earn profits from that trade of $p - y_t$. Trading firms' objective is to maximize profits per unit time. We initially assume that the number of trading firms N is exogenous, and assume that $N \geq 2$. Later, we endogenize entry.

We assume that investors act only as "takers" of liquidity, whereas trading firms act as both "makers" and "takers" of liquidity. More concretely, we assume that investors only use marketable limit orders, which are limit orders with a bid price weakly greater than the best outstanding ask (if buying) or an ask price weakly lower than the best outstanding bid (if selling), whereas trading firms may use both marketable and nonmarketable limit orders.[16]

Latency. Initially, we assume away all latency for trading firms; again, our goal is to create a best-case environment for price discovery and liquidity provision in a continuous limit order book market. Trading firms observe innovations in the signal y with zero time delay, and there is zero latency in sending orders to the exchange and receiving updates from the exchange. If multiple messages reach the market at the same time, they are processed in serial in a random order. This random tie-breaking can be interpreted as messages being transmitted with small

[16] The assumption that investors (equivalently, liquidity traders or noise traders) are liquidity takers is standard in the market microstructure literature. Our treatment of trading firms as both makers and takers of liquidity is slightly nonstandard. This is because our trading firms will play a role that combines aspects of what the traditional market microstructure literature calls a market maker (who provides liquidity) and what the traditional literature calls an informed trader (who takes liquidity). This will become clearer when we describe the role trading firms play in equilibrium in Section VI.B.

random latency, and then processed serially in the order received.[17]

When we endogenize entry by trading firms, we add latency to the observation of innovations in y and the ability to invest resources to reduce this latency.

We assume that investors observe y with latency strictly greater than trading firms; it is unimportant by how much.

VI.B. Equilibrium, Exogenous Entry

In this section we describe the equilibrium of our model with exogenous entry by trading firms. The structure of this equilibrium is unique (as made precise below), but the assignment of trading firms to roles within this structure is not unique. Our solution concept is pure-strategy static Nash equilibrium.[18]

1. Investors. Investors trade immediately when their demand arises, buying or selling at the best available ask or bid, respectively. As we will see, the bid-ask spread is constant in equilibrium, so investors have no incentive to delay trade.

2. Behavior of Trading Firms. The N trading firms endogenously sort themselves into two roles: 1 plays a role we call “liquidity provider” and $N-1$ play a role we call “stale-quote sniper.” Trading firms will be indifferent between these two roles in equilibrium, and our equilibrium uniqueness claim does not specify the precise sorting of trading firms into roles. For simplicity, we assume that they sort themselves into the two roles in a coordinated manner, specifically, player 1 always plays the role of liquidity provider. However, there are economically equivalent equilibria in which who plays the role of liquidity

[17] Exchanges offer a service called colocation to HFT firms, whereby HFTs pay for the right to place their computers in the same location as the exchange’s computers. The exchanges are careful to ensure that each colocated computer is the same physical distance, measured by cord length, from the exchange computers. Hence, if multiple HFTs send orders to the exchange at the same time, it really is random which will be processed first. See Rogow (2012) for more details on colocation.

[18] Static Nash equilibrium means that investors’ and trading firms’ play constitutes a standard Nash equilibrium in each instant of the trading day. This rules out, for instance, the possibility of equilibria in which trading firms collude.

provider is stochastic, or rotates, etc.[19] In practice, some HFT firms primarily play the role of liquidity provider, some primarily play the role of sniper, and some perform both roles.

Liquidity Provider: The liquidity provider behaves as follows. At the start of trading, which we denote by time 0, the liquidity provider submits two limit orders, the first to buy 1 unit of x at price $y_0 - \frac{s}{2}$, the other to sell 1 unit of x at price $y_0 + \frac{s}{2}$. These quotes constitute the opening bid and ask, respectively, and $s \geq 0$ is the bid-ask spread.[20] We derive the equilibrium value of s below. The bid-ask spread will be constant throughout the trading day.

If the signal y jumps at time t, from y_{t^-} to y_t (we use the notation $y_{t^-} = lim_{\; t' \to t^-} y_{t'}$), per the Poisson arrival process described above, the liquidity provider immediately adjusts her quotes. Specifically, at time t she sends a message to the exchange to cancel her previous quotes, of $y_{t^-} - \frac{s}{2}$ and $y_{t^-} + \frac{s}{2}$, and also sends a message with a new bid and ask of $y_t - \frac{s}{2}$ and $y_t + \frac{s}{2}$.

If an investor arrives to the market at time t, per the Poisson arrival process described above, and buys at the current ask of $y_t + \frac{s}{2}$, the liquidity provider immediately replaces the accepted ask with a new ask at this same value of $y_t + \frac{s}{2}$. Similarly, if an investor arrives at time t and sells at the current bid of $y_t - \frac{s}{2}$, the liquidity provider immediately replaces the accepted bid with a new bid at this same value of $y_t - \frac{s}{2}$. In either case, the liquidity provider books profits of $\frac{s}{2}$. Note that the liquidity provider does not directly observe that her trading partner is an investor as opposed to another trading firm, though she can infer this in equilibrium from the fact that trade has occurred at a time t when there is not a jump in the signal y.

If in some time interval there is neither a jump in the signal y, nor the arrival of a new investor, the liquidity provider does not take any action.

[19] In practice tick sizes are discrete (penny increments), whereas we allow for bids and asks to be any real value. If we used a discrete price grid, then the role of liquidity provider would generically be strictly preferred to the role of stale-quote sniper at the equilibrium bid-ask spread. In this case, the N trading firms would race to play the role of liquidity provider, and then the $N-1$ losers of the race would play the role of stale-quote sniper. For a large enough tick size there would also be greater than unit depth in the book.

[20] We adopt the convention that it is possible for a liquidity provider to quote a zero bid-ask spread. Formally, this can be interpreted as the limit as $\epsilon \to 0_+$ of a bid-ask spread of $s = \epsilon$.

Stale-Quote Snipers: Suppose that at time t the signal y jumps from y_{t^-} to y_t, and the jump size $|y_t - y_{t^-}|$ exceeds $\frac{s}{2}$. As described above, the liquidity provider will send a message at time t to cancel her old quotes at $y_{t^-} - \frac{s}{2}$ and $y_{t^-} + \frac{s}{2}$ and replace them with new quotes based on the new value of y. At the exact same time, the $N-1$ other trading firms respond to the change in y by sending a message attempting to "snipe" the stale quotes. That is, they attempt to trade at the old quotes based on y_{t^-} before those quotes are canceled. Since the continuous limit order book processes messages in serial, it is possible that a message to snipe a stale quote will get processed before the liquidity provider's message to cancel the stale quote, creating a rent for the sniper and a cost for the liquidity provider. In fact it is not only possible but probable, because there are $N-1$ snipers against 1 liquidity provider, and it is random whose message is processed first.[21]

Formally, if $y_t > y_{t^-} + \frac{s}{2}$, each stale-quote sniper submits a limit order at t to buy a single unit at price $y_{t^-} + \frac{s}{2}$; symmetrically, if $y_t < y_{t^-} - \frac{s}{2}$, each stale-quote sniper submits a limit order at t to sell a single unit at price $y_{t^-} - \frac{s}{2}$. If the sniper's order executes against the stale quote she books profits of $|y_t - y_{t^-}| - \frac{s}{2}$. If the sniper's order does not execute against the stale quote, that is, if her order is not the first of the N to be processed, she immediately withdraws her order.[22]

If the jump at t is small, specifically, if $y_{t^-} - \frac{s}{2} < y_t < y_{t^-} + \frac{s}{2}$, then the sniper takes no action. Similarly, if in some time interval there is no jump in the signal y, the sniper takes no action.

[21] In our model, all trading firms are equally fast, so their messages reach the exchange at the exact same time, and then the exchange breaks the tie randomly. A more realistic model would add a small random latency to each trading firm's message transmission—for example, a uniform-random draw from $[0, \epsilon]$—and then whichever trading firm had the smallest draw from $[0, \epsilon]$ would win the race (see also note 17). This would yield exactly the same probability of winning the race of $\frac{1}{N}$. Note too that in a richer model with multiple liquidity providers this basic $\frac{1}{N}$ logic still obtains: for every one liquidity provider trying to cancel, all other trading firms—including firms that are also providing their own liquidity—attempt to snipe.

[22] By "immediately withdraws her order" we mean the following. As soon as the sniper receives confirmation from the exchange that her order was not executed, she sends a message to the exchange to remove the order. In our model, both the confirmation that the initial order was not executed and the message to remove the order occur instantaneously. Thus, for any time $t' > t$, the unsuccessful sniper's order is removed by the market by t'. In practice, exchanges automate this type of behavior with an order type called "immediate or cancel."

Equilibrium Bid-Ask Spread. In equilibrium, the bid-ask spread s leaves trading firms indifferent between liquidity provision and stale-quote sniping.

The liquidity provider earns profits of $\frac{s}{2}$ when investors arrive to market, which occurs at arrival rate λ_{invest}, and incurs losses whenever stale quotes are sniped. The losses from sniping arise if there is a jump, which occurs at rate λ_{jump}; the jump is larger than $\frac{s}{2}$; and the liquidity provider does not win the race to react (i.e., is not processed first), which occurs with probability $\frac{N-1}{N}$. In the event she loses the race, her expected loss is $\mathbb{E}(J-\frac{s}{2}|J>\frac{s}{2})$, that is, the conditional expectation of the jump size less half the bid-ask spread. Thus, the benefits less costs of providing liquidity, per unit time, are

$$(1) \qquad \lambda_{invest}\cdot\frac{s}{2}-\lambda_{jump}\cdot\Pr\left(J>\frac{s}{2}\right)\cdot\mathbb{E}\left(J-\frac{s}{2}|J>\frac{s}{2}\right)\cdot\frac{N-1}{N}.$$

Stale-quote snipers earn profits when they successfully exploit a stale quote after a jump larger in size than half the bid-ask spread. When such a jump occurs, each sniper wins the race to exploit with probability $\frac{1}{N}$. Hence each sniper's expected profits, per unit time, are

$$(2) \qquad \lambda_{jump}\cdot\Pr\left(J>\frac{s}{2}\right)\cdot\mathbb{E}\left(J-\frac{s}{2}|J>\frac{s}{2}\right)\cdot\frac{1}{N}.$$

Notice that, summed over all $N-1$ snipers, this equals the liquidity provider's cost of providing liquidity; this captures that trade among trading firms is zero sum.

Equating (1) and (2) yields the equilibrium indifference condition:

$$(3) \qquad \lambda_{invest}\cdot\frac{s}{2}=\lambda_{jump}\cdot\Pr\left(J>\frac{s}{2}\right)\cdot\mathbb{E}\left(J-\frac{s}{2}|J>\frac{s}{2}\right).$$

Equation (3) uniquely pins down the equilibrium bid-ask spread s^*, because the left-hand side is strictly increasing in s and has value 0 at $s=0$, whereas the right-hand side is strictly decreasing in s and is positive for $s=0$. The equation also has a natural economic interpretation. The left-hand side is the total revenue earned by trading firms from investors from the positive bid-ask spread. The right-hand side is the total rents to trading firms from sniping stale quotes. Notice that $\frac{N-1}{N}$ of these rents go to stale-quote snipers and the remaining $\frac{1}{N}$ of these rents goes to the liquidity provider, who is compensated for her opportunity cost of not being a sniper. Notice, too, that equation (3) does not depend on N; this foreshadows that endogenizing entry will have no effect on the bid-ask spread or arms-race prize.

We summarize the equilibrium with the following proposition.

Proposition 1 (Equilibrium with Exogenous Entry). There is an equilibrium of the continuous limit order book market design with play as described above. The structure of this equilibrium is unique in the following sense. In any equilibrium:

(i) At almost all times t, there is exactly one unit offered in the limit order book at bid $y_t - \frac{s^*}{2}$ and exactly one unit offered at ask $y_t + \frac{s^*}{2}$, with the bid-ask spread s^* uniquely characterized by the solution to equation (3). These two quotes may belong to one trading firm or to two distinct trading firms. There are no other orders in the book, except possibly for orders that trade with probability zero.

(ii) Investors trade immediately when their demand arises.

(iii) If there is a jump in y_t that is strictly larger than $\frac{s^*}{2}$, the 1 trading firm with a snipe-able stale quote (i.e., the ask if the jump is positive, the bid if the jump is negative) immediately sends a message to cancel her stale quote, and the other $N - 1$ trading firms immediately send a message to snipe the stale quote. The liquidity provider is sniped with probability $\frac{N-1}{N}$.

(iv) Trading firms are indifferent between liquidity provision and stale-quote sniping, for both the bid and the ask.

(v) As per equation (3), the following two quantities are equivalent in any equilibrium and do not depend on N:

- The total rents to trading firms, $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$. That is, the sum of the value of all arbitrage opportunities that the snipers are racing to capture.
- The total revenue liquidity providers earn from investors via the positive bid-ask spread, $\lambda_{invest} \cdot \frac{s^*}{2}$.

See Online Appendix B for further details about this equilibrium, such as behavior off the equilibrium path, which complete the proof of Proposition 1.

4. Market Depth. Consider the model of Section VI.A but modified so that investors sometimes need to buy or sell multiple units. Specifically, investors arrive to market at rate λ_{invest} and are equally likely to need to buy or sell, as before, but now they need to transact a quantity $q \in \{1, \ldots, \overline{q}\}$, with $p_k > 0$ the probability that they need to transact k units, for $k = 1, \ldots, \overline{q}$. Before, we assumed that investors trade a single unit immediately at market. Here, we make a stronger assumption which is that investors transact their full quantity desired immediately at market. We emphasize that such behavior is not optimal: an investor with multi-unit demand will prefer to split his order into several smaller orders (analogously to Kyle (1985); Vayanos (1999); Sannikov and Skrzypacz (2014)). Instead, we view this assumption as allowing us to illustrate a mechanical point about continuous limit order book markets, which is that sniping makes it especially costly to provide a deep book.

There is an equilibrium of this model analogous to that in Section VI.B, in which the N trading firms serve both as liquidity providers and stale-quote snipers, and are indifferent between these two roles quote by quote. In equilibrium, the bid-ask spread for the kth unit of liquidity, s_k, is governed by indifference between liquidity provision (LHS) and stale-quote sniping (RHS) at the kth level of the book:

$$\lambda_{invest} \cdot \sum_{i=k}^{\overline{q}} p_i \cdot \frac{s_k}{2} - \lambda_{jump} \cdot \Pr\left(J > \frac{s_k}{2}\right) \cdot \mathbb{E}\left(J - \frac{s_k}{2} | J > \frac{s_k}{2}\right) \cdot \frac{N-1}{N}$$

$$= \lambda_{jump} \cdot \Pr\left(J > \frac{s_k}{2}\right) \cdot \mathbb{E}\left(J - \frac{s_k}{2} | J > \frac{s_k}{2}\right) \cdot \frac{1}{N}. \tag{4}$$

The LHS of equation (4) represents the benefits less costs of liquidity provision in the kth level of the book. Notice that the second term on the LHS of equation (4), which describes the costs of getting sniped, is the same as the second term of equation (1). This is because if a quote becomes stale, stale-quote snipers will attempt to pick off as much quantity as is available at an advantageous price. Similarly, the RHS of equation (4), which represents the benefits of sniping the kth level of the book, is the same as equation (2).

By contrast, except for the case of $k = 1$, the first term on the LHS of equation (4), which describes the benefits of providing liquidity, is strictly smaller than the first term of equation (1). This is because only proportion $\sum_{i=k}^{\overline{q}} p_i$ of investors trade the kth level of the order book.

Intuitively, the benefits of providing liquidity scale sublinearly with the quantity offered, because only some investors require a large quantity; whereas the costs of providing liquidity scale linearly with the quantity offered, because snipers will exploit stale quotes in the full quantity offered.[23] The result is that the equilibrium bid-ask spread is wider for the second unit than for the first unit, wider for the third unit than the second unit, etc. That is, the market is "thin" for large-quantity trades.

PROPOSITION 2 (Market Thinness). There exists an equilibrium of the multi-unit demand model with play as described above. The structure of this equilibrium is unique in the following sense. In any equilibrium:

(i) At almost all times t there is exactly one unit offered in the limit order book at bid $y_t - \frac{s_k^*}{2}$ and one unit offered at ask $y_t + \frac{s_k^*}{2}$, for each $k = 1, \ldots, \overline{q}$. The bid-ask spread s_k^* for the kth unit of liquidity is uniquely characterized by equation (4). These $2\overline{q}$ quotes may belong to one trading firm or to multiple distinct trading firms. There are no other orders in the book, except possibly for orders that trade with probability zero.

(ii) Spreads are strictly increasing,

$$s_1^* < s_2^* < \ldots < s_{\overline{q}}^*.$$

Hence, investors' per-unit cost of trading is strictly increasing in order size.

(iii) If there is a jump in y_t that is strictly larger than $\frac{s_k^*}{2}$ and weakly less than $\frac{s_{k+1}^*}{2}$, then there are k snipe-able stale quotes. For each of the k stale quotes, the trading firm with the stale quote immediately sends a message to cancel, and the $N-1$ other trading firms immediately send a message to snipe. Each stale quote is sniped with probability $\frac{N-1}{N}$.

(iv) The N trading firms are indifferent between liquidity provision and stale-quote sniping at all levels of the order book.

[23] A similar intuition is present in Glosten (1994), which derives bid-ask spreads that increase with quantity in a model with asymmetric information. Our market thinness result is to Glosten (1994) as our bid-ask spread result is to Glosten and Milgrom (1985).

(v) As per equation (4), the following two quantities are equivalent in any equilibrium and do not depend on N:
- The total rents to trading firms: $\sum_{k=1}^{\overline{q}} \lambda_{jump} \cdot \Pr(J > \frac{s_k^*}{2}) \cdot \mathbb{E}(J - \frac{s_k^*}{2} | J > \frac{s_k^*}{2})$. That is, the sum of the value of all sniping opportunities, across all levels of the book.
- The total revenue liquidity providers earn from investors via the positive bid-ask spreads, $\lambda_{invest} \sum_{k=1}^{\overline{q}} \cdot \sum_{i=k}^{\overline{q}} p_i \cdot \frac{s_k^*}{2}$.

VI.C. Discussion: Sniping Is "Built In" to the Market Design

Given the model setup, one might have conjectured that Bertrand competition among the N trading firms leads to infinite costless liquidity for investors and zero rents for trading firms. All of the usual channels of costly liquidity provision are turned off. There is no asymmetric information as in the models of Copeland and Galai (1983), Glosten and Milgrom (1985), or Kyle (1985); instead, all trading firms observe innovations in the signal y at exactly the same time, and this signal y is perfectly informative about the fundamental value of x. There are no inventory costs as in Stoll (1978) or search costs as in Duffie, Garleanu, and Pedersen (2005); instead, the security x can at all times be costlessly liquidated at its fundamental value y. So one should expect that competitive forces would drive the price for liquidity to zero.

Our analysis shows, however, that the continuous limit order book market design itself is a source of costly liquidity provision. The core issue is that even symmetrically observed public information creates arbitrage opportunities for trading firms, because trade requests are processed serially. As suggested by the empirics, obvious mechanical arbitrage opportunities are "built in" to the market design. Moreover, serial processing stacks the deck against liquidity providers in the race to respond to new public information. To avoid being sniped, the liquidity provider's request to cancel her stale quote must be processed before *all* of the other trading firms' requests to exploit her stale quote. Hence, liquidity providers get sniped with probability $\frac{N-1}{N}$ even though they learn their quotes are stale at exactly the same time as the other trading firms. In a competitive market, liquidity providers recover the expense of being sniped by charging more for liquidity, that is, sniping costs lead to wider bid-ask spreads and thinner markets.

REMARK 1 (Sniping Harms Liquidity). In our model there are no inventory costs, search costs, or information asymmetries. Nevertheless, in any equilibrium, the bid-ask spread s^* is strictly positive and investors' per-unit cost of trading is strictly increasing in order size.

Our source of costly liquidity provision is most similar to that in Copeland and Galai (1983) and Glosten and Milgrom (1985), namely, a liquidity provider sometimes gets exploited by another player who knows that the liquidity provider's quote is mispriced. The conceptual difference is that in Copeland and Galai (1983) and Glosten and Milgrom (1985) there is asymmetric information between the liquidity provider and this other player (the "informed trader"), whereas in our model the liquidity providers and these other players (stale-quote snipers) are symmetrically informed.[24] The mechanical reason that our source of costly liquidity provision does not arise in these prior works is a subtle difference in how the continuous limit order book is modeled. Our model uses the actual rules of the continuous limit order book (see Section III) in which the market runs in continuous time and players can submit orders whenever they like. Copeland and Galai (1983) and Glosten and Milgrom (1985), as well as other subsequent market microstructure analyses of limit order books such as Foucault (1999) and Goettler, Parlour, and Rajan (2005), use abstractions of the continuous limit order book in which play occurs in discrete time and players can only act when it is their exogenously specified turn to do so. This abstraction is innocuous in the context of their analyses, but it precludes the possibility of a race to respond to symmetrically observed public information as in our analysis.

A potentially useful way to summarize the relationship is that our model shows that the continuous limit order book

[24] One could argue that, in reality, information among HFT firms is always at least slightly asymmetric. Some firm detects the change in the signal y a tiny bit earlier than other firms, and during the interval is asymmetrically informed. Thus, one might argue, sniping is no different from traditional adverse selection due to asymmetric information. However, this argument implicitly assumes that there is no such thing as symmetrically observed information in financial markets (other than perhaps when the market is closed), whereas it is clearly implicit in financial market regulation that certain kinds of information—company news releases, government data announcements, order book activity—should be disseminated symmetrically. Frequent batch auctions restore the possibility of meaningfully symmetric information.

market design causes symmetrically observed public information to be processed by the market as if it were asymmetrically observed private information. As we will see, discrete-time batching eliminates this built-in adverse selection and restores the possibility of meaningfully symmetric information.

VI.D. Equilibrium with Endogenous Entry: The HFT Arms Race

The equilibrium analysis in Section VI.B shows that the continuous limit order book creates rents for trading firms, and that the sniping associated with these rents harms liquidity provision. In this section we incorporate latency into the model and endogenize entry by allowing trading firms to invest in a costly speed technology. This modification induces an arms race for speed. The arms race dissipates the rents created by the continuous market while doing nothing to fix the underlying liquidity problem associated with sniping.

1. Speed Technology. We model investment in speed in a simple way. Trading firms can costlessly observe the signal y with latency $\delta_{slow} > 0$, meaning that the value of signal y at time t is observed at time $t + \delta_{slow}$. In addition, trading firms can choose to invest in a speed technology, at a rental cost of c_{speed} per unit time, which reduces their latency to $\delta_{fast} < \delta_{slow}$. The cost c_{speed} is a metaphor for the cost of access to high-speed data connections (such as the Spread Networks cable, or the microwaves that replaced it), the cost of cutting-edge hardware, the cost of colocation facilities, the cost of the relevant human capital, and so on. We assume that the decision of whether to pay this speed cost is taken at the start of the game and is observable and irreversible.[25] Define $\delta = \delta_{slow} - \delta_{fast}$, the speed difference between fast and slow trading firms.

Before, the number of trading firms N was exogenously specified. Here, we assume that there is a large fringe of slow trading firms, of whom N endogenously decide to invest in speed. There will be no role for slow trading firms in equilibrium. We assume that the cost of speed satisfies a mild condition, described below after equation (7) in note 26, which ensures that in equilibrium

[25] Given the speed investment stage, our equilibrium concept becomes pure-strategy subgame perfect Nash equilibrium for the investment stage, and pure-strategy static Nash equilibrium throughout the trading day.

there are at least two fast trading firms. For simplicity we then allow N to take on any real value greater than or equal to 2, rather than requiring N to be an integer. This allows us to characterize the equilibrium level of N using a zero-profit condition. Alternatively we could require that N is an integer, in which case equilibrium N is characterized by weakly positive profits for trading firms with N entrants and strictly negative with $N+1$.

2. Equilibrium. For expositional simplicity we focus on the case where investors need to buy or sell a single unit; the generalization to multi-unit trading akin to Section VI.B follows naturally.

Equilibrium has a very similar structure to above. The N fast trading firms who endogenously enter then sort themselves into 1 liquidity provider and $N-1$ stale-quote snipers. Both the liquidity provider and the stale-quote snipers behave exactly as described above in Section VI.B, with the one modification that they now each react to jumps in y with latency δ_{fast}. Investors behave exactly as before, buying or selling immediately on arrival.

Notice that while a fast liquidity provider successfully avoids getting sniped $\frac{1}{N}$ of the time, a slow liquidity provider would always be sniped. Similarly, a fast stale-quote sniper is successful $\frac{1}{N}$ of the time, whereas a slow stale-quote sniper would never be successful. This is the intuition for why there is no role for slow-trading firms in equilibrium.

Equilibrium is characterized by two zero-profit conditions. First, we have the zero-profit condition for the liquidity provider, which says that revenues minus costs as written in equation (1) equal the costs of speed:

$$\lambda_{invest} \cdot \frac{s}{2} - \lambda_{jump} \cdot \Pr\Big(J > \frac{s}{2}\Big) \cdot \mathbb{E}\Big(J - \frac{s}{2}|J > \frac{s}{2}\Big) \cdot \frac{N-1}{N} = c_{speed}. \tag{5}$$

Second is the zero-profit condition for stale-quote snipers, which says that the rents from sniping as written in equation (2) equal the costs of speed:

$$\lambda_{jump} \cdot \Pr\Big(J > \frac{s}{2}\Big) \cdot \mathbb{E}\Big(J - \frac{s}{2}|J > \frac{s}{2}\Big) \cdot \frac{1}{N} = c_{speed}. \tag{6}$$

Together, equations (5) and (6) characterize the equilibrium bid-ask spread s^* and the equilibrium quantity of entry N^*. Notice that subtracting equation (6) from equation (5) yields exactly equation (3); hence, the equilibrium bid-ask spread is the same as in the exogenous entry case. We can then solve for the equilibrium entry quantity by adding equation (5) and $N-1$ times equation (6) to obtain

$$\lambda_{invest} \cdot \frac{s^*}{2} = N^* \cdot c_{speed}. \tag{7}$$

The economic interpretation of equation (7) is that all of the expenditure by trading firms on speed technology (RHS) is ultimately borne by investors via the cost of liquidity (LHS). Examining equation (3) as well, we have an equivalence between the total prize in the arms race, the total expenditures on speed in the arms race, and the cost to investors.[26] Hence, the rents created by the continuous limit order book are dissipated by the speed race.[27]

Proposition 3. There is an equilibrium of the continuous limit order book market design with endogenous entry with play as described above. The equilibrium number of fast trading firms N^* and the equilibrium bid-ask spread s^* are uniquely determined by the zero-profit conditions equations (5) and (6). The structure of play in this equilibrium is identical to that in the exogenous entry case, as characterized by Proposition 1, but replacing the exogenous N trading firms with the endogenous N^* fast trading firms. Slow trading firms play no role in equilibrium. The following three quantities are equivalent in any equilibrium:

[26] The assumption that $N \geq 2$ in equilibrium can be written as $c_{speed} < \frac{1}{2}\lambda_{invest} \cdot \frac{s^*}{2}$. This is mild since $\lambda_{invest} \cdot \frac{s^*}{2}$ is equal to the total prize in the arms race.

[27] We have assumed that all fast traders are equally fast and have the same cost of speed. A simple way to capture the fact that some trading firms may have a comparative advantage in speed technology is to allow the cost of speed to vary over firms. Under this modification, the marginal fast trading firm earns zero profits, while inframarginal trading firms earn strictly positive profits. With this modification, the total sniping rents to trading firms remain equal to the total revenue liquidity providers earn from investors via the positive bid-ask spread, and these two quantities each strictly exceed the total equilibrium expenditure by trading firms on speed technology.

- The total rents to trading firms, $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$. That is, the sum of the value of all arbitrage opportunities that the snipers are racing to capture.
- The total revenue liquidity providers earn from investors via the positive bid-ask spread, $\lambda_{invest} \cdot \frac{s^*}{2}$.
- The total equilibrium expenditure by trading firms on speed technology, $N^* \cdot c_{speed}$.

VI.E. Discussion of the Equilibrium

1. Welfare Costs of the Arms Race: A Prisoner's Dilemma among Trading Firms. The equilibrium derived above can be interpreted as the outcome of a prisoner's dilemma among trading firms. To see this, compare the equilibrium outcome with endogenous entry to the equilibrium outcome with exogenous entry if the exogenous number of trading firms is N^* and their latency is δ_{slow}. In both cases, the N^* trading firms sort themselves into 1 liquidity provider and $N^* - 1$ stale-quote snipers, and in both cases the bid-ask spread, s^*, is characterized by trading firms' indifference between liquidity provision and stale-quote sniping. The only difference is that now all trading firms—both the liquidity provider and the snipers—respond to changes in y with a delay of δ_{slow} instead of δ_{fast}. Investors still get to trade immediately and still pay the same bid-ask spread cost of $\frac{s^*}{2}$, so their welfare is unchanged. The welfare of the N^* trading firms is strictly greater though, since they no longer pay the cost of speed.

Proposition 4 (Prisoner's Dilemma). Consider the model of Section VI.D modified so that the number of trading firms is N^*. Social welfare would be higher by $N^* \cdot c_{speed}$ per unit time if the N^* trading firms could commit not to invest in the speed technology, with the gains shared equally among the N^* trading firms. But each individual trading firm has a dominant strategy incentive to deviate and invest in speed, so this is not an equilibrium. The situation constitutes a prisoner's dilemma with social costs equal to the total expenditure on speed.

As we will see below, frequent batch auctions resolve this prisoner's dilemma, and in a manner that allocates the welfare savings to investors instead of trading firms.

2. Connection to the Empirics: The Arms Race Is a "Constant".

PROPOSITION 5 (Comparative Statics of the Arms Race Prize). The size of the prize in the arms race, $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} \,|J > \frac{s^*}{2})$, has the following comparative statics:

(i) The size of the prize is increasing in the frequency of jumps, λ_{jump}.

(ii) If jump distribution F'_{jump} is a mean-preserving spread of F_{jump}, then the size of the prize is strictly larger under F'_{jump} than F_{jump}.

(iii) The size of the prize is invariant to the cost of speed, c_{speed}.

(iv) The size of the prize is invariant to the speed of fast trading firms, δ_{fast}.

(v) The size of the prize is invariant to the difference in speed between fast and slow trading firms, δ.

Proposition 5 suggests that the HFT arms race is best understood as an equilibrium constant of the continuous limit order book—and thus helps make sense of our empirical results. Specifically, suppose that speed technology improves each year, and reinterpret the model so that c_{speed} is the cost of being at the cutting edge of speed technology in the current year, δ_{fast} is the speed at the cutting edge, and δ is the speed differential between the cutting edge and other trading firms. Under this interpretation, in equilibrium of our model, the speed with which information (y) is incorporated into prices (x) grows faster and faster each year—as consistent with our findings in the correlation breakdown analysis (Figure III). And, arbitrage durations decline each year—as consistent with our findings on the duration of ES-SPY opportunities (Figure IV). However, the arms race prize itself is unaffected by these advances in speed, which is consistent with Figures V and VI because the total size of the prize can be decomposed as per arbitrage profitability $\mathbb{E}(J - \frac{s^*}{2} \,|J > \frac{s^*}{2})$ times arbitrage frequency $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2})$. What does affect the size of the prize are the market volatility parameters, again consistent with our findings in the arbitrage analysis.

3. Relationship to the Efficient Markets Hypothesis. It is interesting to interpret the equilibrium derived above as it relates to the efficient markets hypothesis.

On the one hand, the market is highly efficient in the sense of instantaneously incorporating news about y into the price of x. Formally, the midpoint of the bid-ask spread for x is equal to fast trading firms' information about x's fundamental value, $y_{t-\delta_{fast}}$, for proportion one of the trading day.

On the other hand, a strictly positive volume of trade is conducted at prices known by all trading firms to be stale. Formally, the proportion of trade that is conducted at quotes that do not contain $y_{t-\delta_{fast}}$ is

$$\frac{\lambda_{jump}\cdot \Pr(J>\frac{s^*}{2})\cdot\frac{N^*-1}{N^*}}{\lambda_{jump}\cdot \Pr(J>\frac{s^*}{2})\cdot\frac{N^*-1}{N^*}+\lambda_{invest}}.$$

Hence, the market is extremely efficient in *time space* but not in *volume space*: a lot of volume gets transacted at incorrect prices. This volume is in turn associated with rents from symmetrically observed public information about securities' prices, which is in violation of the weak-form efficient markets hypothesis (see Fama 1970).[28]

4. Role of HFTs. In equilibrium of our model, fast trading firms endogenously serve two roles: liquidity provision and stale-quote sniping. The liquidity provision role is useful to investors; the stale-quote sniping role is detrimental to investors because it increases the costs of liquidity provision.[29]

This distinction between roles is important to keep in mind when interpreting the historical evidence on the effect of HFT on liquidity. The rise of HFT over the past 15 years or so conflates two distinct phenomena: the increased role of information technology (IT) in financial markets, and the speed race. The empirical record is unambiguous that, overall, IT has improved liquidity—see especially Hendershott, Jones, and Menkveld

[28] The citation for the 2013 Nobel Prize in Economics asserted that asset prices are predictable in the long run but “next to impossible to predict in the short run” (Economic Sciences Prize Committee 2013). Our empirical and theoretical results show that in fact, prices are extremely easy to predict in the *extremely* short run.

[29] In practice, and in richer models, HFTs serve roles beyond these two. For instance, Clark-Joseph (2013) studies an HFT strategy that relates to the sophisticated use of information technology to detect patterns in others' trading activity and trade in advance of large orders. Clark-Joseph (2013) argues that this strategy, which he calls exploratory trading, is detrimental to investors, and it is clearly distinct from stale-quote sniping.

(2011), which uses a natural experiment to show that the transition from human-based liquidity provision to computer-based liquidity provision enhanced liquidity. This makes intuitive economic sense, as IT has lowered costs in numerous sectors throughout the economy. However, there is little support for the proposition that the speed race per se has improved liquidity. Moreover, in the time series of both bid-ask spreads over time (Virtu 2014, p. 103) and the cost of executing large trades over time (Angel, Harris and Spatt 2015, p. 23; Frazzini, Israel, and Moskowitz 2012, table IV), it appears that most of the improvements in liquidity associated with the rise of IT were realized in the late 1990s and early to mid-2000s, well before the millisecond- and microsecond-level speed race.

We emphasize that our results do not imply that on net HFT has been negative for liquidity or social welfare. Our results say that sniping is negative for liquidity and that the speed race is socially wasteful. Frequent batch auctions preserve (in a sense, enhance) the useful function served by HFTs—liquidity provision and price discovery—while eliminating sniping and the speed race.

VII. Frequent Batch Auctions as a Market Design Response

In this section we define the frequent batch auction market design and show that it directly addresses the problems we have identified with the continuous limit order book market design.

VII.A. Frequent Batch Auctions: Definition

Informally, frequent batch auctions are just like the continuous limit order book but with two departures: (i) time is treated as discrete, not continuous; and (ii) orders are processed in batch, using a uniform-price auction, instead of serially in order of receipt. The remainder of this subsection defines frequent batch auctions formally.[30]

The trading day is divided into equal-length discrete time intervals, each of length $\tau > 0$. We refer to the parameter τ as the *batch length* and to the intervals as *batch intervals*. We

[30] See also Budish, Cramton, and Shim (2014), which provides additional practical implementation details.

refer to a generic batch interval either using the interval, generically $(0, \tau]$, or using the ending time, generically t.

At any moment in time during a batch interval, traders (i.e., investors or trading firms) may submit offers to buy and sell shares of stock in the form of limit orders and market orders. Just as in the continuous market, a limit order is a price-quantity pair expressing an offer to buy or sell a specific quantity at a specific price, and a market order specifies a quantity but not a price.[31] A single trader may submit multiple orders, which can be interpreted as submitting a demand function or a supply function (or both). Just as in the continuous market, traders may freely modify or cancel their orders at any moment in time. Also, just as in the continuous market, orders remain outstanding until either executed or canceled; that is, if an order is not executed in the batch at time t, it automatically carries over for $t+\tau$, $t+2\tau$, $t+3\tau$, etc.

At the end of each batch interval, the exchange batches all outstanding orders—both new orders received during this interval, and orders outstanding from previous intervals—and computes the aggregate demand and supply functions out of all bids and asks, respectively. If demand and supply do not intersect, then there is no trade and all orders remain outstanding for the next batch auction. If demand and supply do intersect, then the market clears where supply equals demand, with all transactions occurring at the same price—that is, at a "uniform price." There are two cases to consider. If demand and supply intersect horizontally or at a point, this pins down a unique market-clearing price p^* and a unique maximum possible quantity q^*. In this case, offers to buy with bids strictly greater than p^* and offers to sell with asks strictly less than p^* transact their full quantity at price p^*, whereas for bids and asks of exactly p^* it may be necessary to ration one side of the market to enable market clearing.[32]

[31] We assume that there is a finite maximum allowable bid and minimum allowable ask. A market order to buy q shares is then interpreted as a limit order to buy q shares at the maximum allowable bid, and symmetrically for market orders to sell. In practice, price circuit breakers might determine what constitutes these maximum and minimum amounts (e.g., the most recently transacted price plus or minus some specified percentage).

[32] A potential reason to favor fine rather than coarse tick sizes is to reduce the likelihood of ties and hence the amount of rationing. We also note that one of the arguments against fine tick sizes in continuous markets—the explosion in message traffic associated with traders outbidding each other by economically negligible

For this rationing, we adopt a time-priority rule analogous to current practice under the continuous market but treating time as discrete: orders that have been left outstanding for a larger (integer) number of batch intervals have higher priority, whereas if two orders were submitted in the same batch interval they have the same priority irrespective of the precise time they were submitted within that batch interval. If necessary to break ties between orders submitted during the same batch interval the rationing is random (pro rata). If demand and supply intersect vertically, this pins down a unique quantity q^* and an interval of market-clearing prices, $[p_L^*, p_H^*]$. In this case, all offers to buy with bids weakly greater than p_H^* and all offers to sell with asks weakly lower than p_L^* transact their full quantity, and the price is $\frac{p_L^*+p_H^*}{2}$.

Information policy details are as follows. After each auction is computed all of the orders that were entered into the batch auction, both outstanding orders from previous batch intervals and new orders entered during the just-completed batch interval, are displayed publicly. Also displayed are details of the auction outcome: the supply and demand functions, and the market-clearing price and quantity (or “no trade”). Activity during the batch interval is not displayed publicly during the batch interval; that is, information is disseminated in discrete time. So, for the time t auction, participants see all of the orders and auction information from the auctions at time $t-\tau, t-2\tau, t-3\tau, \ldots$, but they do not see new activity for the time t auction until after the auction is completed. This information policy may sound different from current practice, but it is in fact closely analogous. In the continuous market, new order book activity is first economically processed by the exchange (e.g., a new order is entered in the book, or a new order trades against the book), and only then is the order announced publicly (along with the updated state of the book). Similarly, here, new order book activity is first economically processed by the exchange and only then announced publicly; the only difference is that the economic processing occurs in

amounts—is less of an issue in discrete time. For these reasons, we conjecture that the optimal tick size in a frequent batch auction is at least as fine as that in the continuous limit order book. This is an open question for future research.

discrete time, and hence the information dissemination occurs in discrete time as well.[33]

To further clarify the relationship to the continuous limit order book market design, it is helpful to discuss three scenarios. A first scenario is that there is no new activity during the batch interval; this case would be quite common if the batch interval is short. In this case, all outstanding orders simply carry over to the next batch interval, analogous to displayed liquidity in a continuous limit order book. A second scenario is that a single investor arrives during the batch interval and submits an order to buy (analogously, to sell) at the best outstanding offer from the previous batch interval, that is, the frequent batch auction version of the ask (analogously, bid). This scenario is also closely analogous to the continuous market. The investor trades at the bid or ask, and which order or orders get filled is based on our version of time priority. A third scenario is that there is a large amount of new activity in the batch interval; for example, there is a news event and many trading algorithms are reacting at once. In this scenario frequent batch auctions and the continuous limit order book are importantly different: frequent batch auctions process all of the new activity together in a uniform-price auction, at the end of the interval, whereas the continuous market processes the new activity serially in order of arrival.

VII.B. *Why and How Frequent Batch Auctions Address the Problems with Continuous Trading*

Frequent batch auctions directly address the problems we identified in Section VI with the continuous limit order book, for two reasons.

First, and most obviously, discrete time reduces the value of tiny speed advantages. To see this, consider a situation with two trading firms, one who pays the cost c_{speed} and hence has latency δ_{fast}, and one who does not pay the cost and hence has latency δ_{slow}. In the continuous market, whenever there is a jump in y the fast trading firm gets to act on it first. In the frequent batch

[33] Displaying new activity during the batch interval would create at least two problems. First, orders would be displayed that might never be intended to be economically binding. For instance, a fast trader could place a large order to buy early in the batch interval, to create the impression that there is a lot of demand to buy, only to withdraw the buy order right at the end of the batch interval and instead place a large order to sell. Second, an investor wishing to buy or sell at market could not do so without displaying his intention publicly before his trade is executed.

auction market, the fast trading firm's speed advantage is only relevant if the jump in y occurs at a very specific time in the batch interval. Any jumps in y that occur during the window $(0, \tau - \delta_{slow}]$ are observed by both the slow and fast trading firm in time to react for the batch auction at τ. Similarly, any jumps in y that occur during the window $(\tau - \delta_{fast}, \tau]$ are observed by neither the fast nor the slow trading firm in time for the auction at τ. It is only jumps that occur in a window of time of length $\delta = \delta_{slow} - \delta_{fast}$, taking place from $(\tau - \delta_{slow}, \tau - \delta_{fast}]$, that create meaningful asymmetric information between the fast and slow trader. Hence, the proportion of the trading day during which the fast trader's speed advantage is relevant is reduced from 1 to $\frac{\delta}{\tau}$. For instance, if the batch interval is 100 milliseconds and the speed difference is 100 microseconds, the likelihood that the fast trading firm's speed advantage results in economically relevant asymmetric information is reduced by a factor of $\frac{1}{1000}$. See Figure VII for an illustration.

Second, and more subtly, the use of batch auctions eliminates sniping. This is best explained with two examples. In the first example, consider the model of Section VI.B, with N trading firms exogenously in the market all equally fast. Consider a trading firm providing liquidity. In the continuous market, every time there is a jump in y, the liquidity provider is vulnerable to being sniped. He submits a message to cancel his stale quotes, but at the exact same time the other $N – 1$ trading firms submit a message to snipe the stale quotes, and it is random who gets processed first. So, if there is a large enough jump, he is sniped with probability $\frac{N-1}{N}$. In the batch auction market, by contrast, a symmetrically informed liquidity provider is never sniped.[34] If y jumps, say from $\underline{y}$ to $\overline{y} > \underline{y}$, then the liquidity provider cancels his old quotes based on $\underline{y}$ and submits new quotes

[34] That symmetrically informed liquidity providers are never sniped is an artifact of our stylized latency model. But consider as well the following more realistic latency model, which leads to a substantively similar conclusion. Trading firms observe each innovation in y with latency of δ_{fast} plus a uniform-random draw from $[0, \epsilon]$, where $\epsilon > 0$ represents the maximum difference in latency among trading firms in response to any particular signal. Now, a liquidity provider is vulnerable to being sniped if (i) a jump in y occurs during the interval $(\tau - \delta_{fast} - \epsilon, \tau - \delta_{fast})$, and (ii) this jump occurs later than the liquidity provider's own random draw from $[0, \epsilon]$. The proportion of a given batch interval during which (i) and (ii) obtain is $\frac{\epsilon}{2\tau}$. Whereas δ, the difference in speed between a fast and a slow trader in practice might be measured in milliseconds, the parameter ϵ would in practice be measured in microseconds. Hence, even for short batch intervals, the proportion $\frac{\epsilon}{2\tau}$ is very

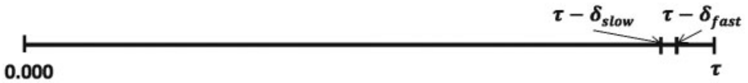

FIGURE VII

Illustration of How Discrete Time Reduces the Value of Tiny Speed Advantages

τ denotes the length of the batch interval, δ_{slow} denotes the latency with which slow traders observe information, and δ_{fast} denotes the latency with which fast traders observe information. Any events that occur between time 0 and time $\tau - \delta_{slow}$ are observed by both slow and fast traders in time for the next batch auction. Any events that occur between $\tau - \delta_{fast}$ and τ are observed by neither slow nor fast traders in time for the next batch auction. Only events that occur between $\tau - \delta_{slow}$ and $\tau - \delta_{fast}$ create an asymmetry between slow and fast traders, because fast traders observe them in time for the next batch auction but slow traders do not. This critical interval constitutes proportion $\frac{\delta}{\tau}$ of the trading day, where $\delta \equiv \delta_{slow} - \delta_{fast}$. For more details see the text of Section VII.B.

based on $\bar{y}$. The other trading firms may try to snipe the stale quotes, say, by submitting orders to buy at the old ask price based on $\underline{y}$. But, because the liquidity provider's canceled quotes are not even entered into the next batch auction, the snipers are irrelevant.[35]

In the second example, suppose as in the first that there are $N - 1$ fast trading firms who seek to snipe stale quotes, but that now the 1 trading firm providing liquidity is slow, not fast. Moreover, suppose that there is a jump in y during the critical interval $(\tau - \delta_{slow}, \tau - \delta_{fast}]$, say, from $\underline{y}$ to $\bar{y}$, where the fast trading firms see the jump but the slow trading firm does not. In the continuous market, if multiple fast traders attempt to exploit a stale quote at essentially the same time, the exchange processes whichever trader's order reaches the exchange the fastest. In a batch auction, by contrast, if multiple fast traders attempt to exploit a stale quote at essentially the same time—meaning in the same batch interval—the trade goes to whichever trader offers the best price. Serial processing implies speed-based competition, whereas batch processing using a uniform-price auction allows

small. For example, if ϵ is 10 microseconds and τ is 100 milliseconds, then $\frac{\epsilon}{2\tau} = 0.00005$.

[35] Observe that it is the combination of discrete-time and batch auctions that eliminates sniping. Discrete time alone is insufficient: if new messages received during the batch interval are processed serially at the end of the interval, for instance in a random order, then a sniper's request to buy at $\underline{y}$ may get serially processed before the liquidity provider's request to cancel his quotes at $\underline{y}$.

for price competition. Equilibrium price competition among fast traders then drives the price of x up to its new correct level, namely, $\bar{y}$. At any hypothetical market-clearing price $p < \bar{y}$, each fast trader strictly prefers to deviate and bid a tiny amount more, so in any equilibrium the market-clearing price in the auction is $\bar{y}$.

Thus, frequent batch auctions eliminate sniping and the arms race by transforming the nature of competition among trading firms: from competition on speed to competition on price. We summarize this discussion as follows:

Proposition 6 (Batching Eliminates Sniping and the Arms Race). Consider the frequent batch auction market design in the model of Section VI.D.

(i) The proportion of the trading day during which jumps in y leave a slow liquidity provider vulnerable to being sniped by a fast trader is $\frac{\delta}{\tau}$.
(ii) The proportion of the trading day during which jumps in y leave a fast liquidity provider vulnerable to being sniped is 0.
(iii) If there are $N \geq 2$ fast traders exogenously in the market, and there is a slow liquidity provider with a vulnerable stale quote—that is, there is a jump in y during $(\tau - \delta_{slow}, \tau - \delta_{fast}]$ such that $y_{\tau-\delta_{fast}}$ is either greater than the slow liquidity provider's ask or less than the bid—then Bertrand competition among the fast traders drives the batch auction price of x to $y_{\tau-\delta_{fast}}$. The slow liquidity provider transacts at $y_{\tau-\delta_{fast}}$.

By contrast, in the continuous limit order book:

(i) The proportion of the trading day during which jumps in y leave a slow liquidity provider vulnerable to being sniped by a fast trader is 1.
(ii) A fast liquidity provider is sniped for proportion $\frac{N-1}{N}$ of sufficiently large jumps in y, where N is the number of fast traders present in the market. This is the case even though she observes jumps in y at exactly the same time as the other $N - 1$ fast traders.
(iii) If there are $N \geq 2$ fast traders present in the market, and there is a slow liquidity provider with a vulnerable stale quote—that is, there is a jump in y at time t such that y_t is either greater than the slow liquidity provider's ask or less than the bid—then whichever of the N fast traders' orders

is processed by the exchange first transacts at the stale quote. The slow liquidity provider transacts at the stale quote.

VII.C. Equilibrium of Frequent Batch Auctions

Section VII.B described why frequent batch auctions eliminate sniping and the HFT arms race, by reducing the value of tiny speed advantages and transforming competition on speed into competition on price. In this section we study how this in turn translates into equilibrium effects on bid-ask spreads, market depth, and investment in speed. Following the analysis of Section VI, we first consider the case of exogenous entry and then consider endogenous entry.

1. Model. We study the equilibria of frequent batch auctions using the model of Section VI.A that we used to study the continuous limit order book, with one modification. In the model of Section VI.A, investors arrive according to a Poisson process with arrival rate λ_{invest}. In the context of the continuous market, the Poisson process makes an implicit finiteness assumption, because the probability that more than one investor arrives at any instant is zero. Here, we need to make an explicit finiteness assumption. Specifically, we assume that investors continue to arrive according to the Poisson process, and continue to be equally likely to need to buy or sell a unit, but we assume that the net demand of investors in any batch interval—number who need to buy minus number who need to sell—is bounded. Formally, let $A(\tau)$ denote the random variable describing the number of investors who arrive in a batch interval of length τ, and let $D(\tau)$ denote the random variable describing their net demand. We assume that $D(\tau)$ is symmetric about zero and that there exists a $\overline{Q} < \infty$ such that the absolute value of $D(\tau)$ is bounded by $\overline{Q} - 1$. We view this assumption as innocuous so long as $\overline{Q}$ is large relative to the standard deviation of the Poisson arrival process, $\sqrt{\tau\lambda_{invest}}$.

2. Exogenous Entry

We begin by considering the setting of Section VI.B in which the number of trading firms is exogenously set to $N \geq 2$ and there is no latency.

Since all trading firms are equally fast, there is no sniping, per the discussion in Section VII.B. Since all of the other sources of costly liquidity provision are turned off, there exists an equilibrium in which fast trading firms offer at least the maximum necessary depth, $\overline{Q}$, at zero bid-ask spread,[36] and investors trade at market in the batch auction immediately following their arrival. This equilibrium is essentially unique and obtains for any batch interval $\tau > 0$.

Proposition 7 (Equilibrium of Frequent Batch Auctions with Exogenous Entry). Fix any batch interval $\tau > 0$ and any number of trading firms $N \geq 2$. In any equilibrium of the frequent batch auction market design with exogenous entry, investors trade at market in the next batch auction after their arrival, and the N trading firms collectively offer at least the maximum necessary depth to satisfy investor demand at zero bid-ask spread. As compared to the equilibrium of the continuous limit order book market, the equilibrium effects of frequent batch auctions are as follows:

(i) The bid-ask spread for the first-quoted unit is narrower: it is 0 instead of the spread characterized by equation (3).
(ii) The market is deeper: the order book has the maximum depth necessary to serve all investors at zero bid-ask spread, whereas in the continuous limit order book, as per the model considered in Section VI.B.4, the bid-ask spread grows wider with the quantity traded.

This equilibrium highlights the central differences between frequent batch auctions and the continuous limit order book. There are no longer rents from symmetrically observed public information; in equilibrium, trading firms earn zero rents. Liquidity providers are no longer vulnerable to sniping; discrete time affords liquidity providers an opportunity after each jump in y to adjust their quotes to reflect the new public information. Bertrand competition competes the bid-ask spread to zero, and generates effectively infinite market depth, as we would have expected given the model setup.

[36] We maintain the convention from Section VI that it is possible to offer a zero bid-ask spread. Formally, fast trading firms can be interpreted as offering to buy at least $\overline{Q}$ units at price $y_\tau - \epsilon$ and sell at least $\overline{Q}$ units at price $y_\tau + \epsilon$, in the limit as $\epsilon \rightarrow 0_+$.

3. Endogenous Entry. In this section we consider the equilibrium of frequent batch auctions with endogenous entry. We show that if the batch interval τ is sufficiently large relative to δ there is an essentially unique equilibrium in which no trading firms pay the cost c_{speed} to have latency δ_{fast} rather than δ_{slow}. Liquidity is provided to investors by slow trading firms. As in the equilibrium with exogenous entry, competition leads to a bid-ask spread of zero and effectively infinite depth.

Suppose that slow trading firms in aggregate provide $\overline{Q}$ of depth for x at zero bid-ask spread. That is, in the auction ending at τ, slow trading firms collectively offer to buy and sell $\overline{Q}$ units at price $y_{\tau-\delta_{slow}}$, where $y_{\tau-\delta_{slow}}$ represents the best available information for a slow trader about the value of security x in the auction ending at τ.

A potential entrant considers whether to invest c_{speed} to be fast, with the aim of sniping this $\overline{Q}$ of depth in the event that there is a jump in y in the time interval $(\tau - \delta_{slow}, \tau - \delta_{fast}]$, which the entrant would observe and the slow trading firms providing liquidity would not. If there are $\overline{Q}$ units of depth in the limit order book, and there is, say, a positive jump, the entrant will wish to buy all $\overline{Q}$ units at the stale ask prices. If the imbalance D of investors—number of orders to buy minus orders to sell—is positive, then the amount that the fast trader can transact will be smaller than $\overline{Q}$ by the amount D, because the investors will outbid him for D of the $\overline{Q}$ units. On the other hand, if the imbalance D is negative, the fast trader can transact not just the $\overline{Q}$ units offered by the slow trading firms but can also satisfy the imbalance. He can achieve this by submitting a large limit order to buy at a price slightly larger than $y_{\tau-\delta_{slow}}$, so that he purchases all $\overline{Q}$ units at the ask of $y_{\tau-\delta_{slow}}$ as well as satisfies the D net market orders to sell. Hence, the fast trader transacts an expected quantity of $\overline{Q}$ units in any batch interval where there is an exploitable jump.

Let p_{jump} denote the probability that there are one or more jumps in y in the δ interval, and let J' denote the random variable describing the total jump amount in a δ interval, conditional on there being at least one jump. Since the probability of multiple jumps in a δ interval is small, $p_{jump} \approx \delta\lambda_{jump}$ and $\mathbb{E}(J') \approx \mathbb{E}(J)$. The fast trader's expected profits from exploiting the slow liquidity providers, on a per unit time basis, are thus $\frac{p_{jump}}{\tau}\mathbb{E}(J') \cdot \overline{Q} \approx \frac{\delta}{\tau} \cdot \lambda_{jump}\mathbb{E}(J) \cdot \overline{Q}$. Note that a difference versus the analogous expression in equation (2) is that the bid-ask spread is now zero, so any jump can be profitably exploited, in the full jump size amount. The fast trading firm's costs per unit time are c_{speed}. Hence, entry as a fast trading firm sniping the slow trading firms is not optimal if, using the approximations above,

$$\frac{\delta}{\tau} \cdot \lambda_{jump} \cdot \mathbb{E}(J) \cdot \overline{Q} < c_{speed}. \tag{8}$$

The fraction $\frac{\delta}{\tau}$ is the proportion of time that the fast trader sees jumps in y that the slow traders do not see in time (see Figure VII), and these jumps occur at rate λ_{jump}. The LHS of equation (8) is thus increasing in δ, the fast trader's speed advantage, but decreasing in τ, the batch interval. Intuitively, in a long batch interval, most jumps occur at times where both the fast and slow traders are able to react in time.

For any finite $\overline{Q}$, equation (8) is satisfied for sufficiently large τ. Hence, any desired market depth can be provided by slow trading firms at zero cost if the batch interval τ is sufficiently large. Moreover, the maximum depth $\overline{Q}$ consistent with equation (8) grows linearly with τ, whereas the expected imbalance of investor demand in a batch interval grows at rate $\sqrt{\tau}$.

We summarize the derived equilibrium as follows.

Proposition 8 (Equilibrium of Frequent Batch Auctions with Endogenous Entry). Fix any batch interval τ satisfying $\frac{p_{jump}}{\tau}\mathbb{E}(J') \cdot \overline{Q} < c_{speed}$, the exact version of equation (8). In any equilibrium of the frequent batch auction market design with endogenous entry, investors trade at market in the next batch auction after their arrival, and slow trading firms collectively offer at least the maximum necessary depth to satisfy investor demand at zero bid-ask spread. As compared to the equilibrium of the continuous limit order book market, the equilibrium effects of frequent batch auctions are as follows:

(i) The bid-ask spread for the first-quoted unit is narrower: it is 0 instead of $\frac{N^* \cdot c_{speed}}{\lambda_{invest}}$.

(ii) The market is deeper: the order book has the maximum depth necessary to serve all investors at zero bid-ask spread, whereas in the continuous limit order book, as per the extended model considered in Section VI.B, the bid-ask spread grows wider with the quantity traded.

(iii) Social welfare:

- Benefits: there is no more arms race. Trading firms choose latency δ_{slow} rather than paying c_{speed} to have latency δ_{fast}. This generates a welfare savings of $N^{*}\cdot c_{speed}$ per unit time, where N^{*} is the number of fast trading firms in equilibrium of the continuous limit order book.
- Costs: investors have to wait a positive amount of time to complete their trade. Expected delay costs are $\frac{1}{\tau}\int_{0}^{\tau} f_{delaycost}(x)\lambda_{invest}dx$ per unit time.

Notice that with respect to social welfare, frequent batch auctions have both benefits and costs. The benefit is that they stop the arms race in speed, which we showed in Proposition 4 can be understood as a socially wasteful prisoner's dilemma.[37] The cost is that investors have to wait a positive amount of time to trade, so they incur delay costs. Intuition suggests that these delay costs are likely to be negligible for the kinds of time intervals we are discussing in this article, but since we lack a theoretical foundation for where these delay costs come from, we do not reach a definitive conclusion about social welfare in the proposition.[38]

In Online Appendix B.3.1, we use a combination of our ES-SPY analysis and information from HFT public documents to calibrate equation (8). The goal of this exercise is not to determine the optimal batch interval, but rather to get an extremely rough

[37] Frequent batch auctions enhance liquidity, but since investors' demand to trade is inelastic in our model this enhanced liquidity does not translate into a welfare gain. In a richer model with elastic demand to trade this would be an additional welfare benefit of frequent batch auctions.

[38] The working paper version of this article considers the case where τ fails equation (8). In this case, it is no longer an equilibrium for liquidity to be provided by trading firms who do not pay c_{speed}. Such trading firms would be too vulnerable to sniping. Instead, liquidity is provided by a fast trading firm who pays c_{speed}, as in the equilibrium of the continuous limit order book with endogenous entry. The key difference is that the *fast* trading firm is no longer vulnerable to sniping, as per Proposition 6. As a result, the equilibrium bid-ask spread is narrower and depth is greater than in the continuous market with endogenous entry, though the spread is wider and the market is less deep than in the case of τ satisfying equation (8). Equilibrium expenditure on speed also lies between the continuous market with endogenous entry and frequent batch auctions with τ satisfying equation (8). In the limiting case of $\tau\rightarrow 0_{+}$ we can reach a definitive welfare conclusion, because there are benefits of frequent batch auctions—though not as large as in the case where τ satisfies equation (8)—and zero costs, because investor delay costs vanish as the delay goes to zero.

sense of magnitudes for how long a batch interval is long enough to stop the HFT speed race. The parameter δ is open to two potential interpretations. One interpretation is that δ represents the year-on-year speed improvements of state-of-the-art HFTs; in New York–Chicago trades like ES-SPY, the difference in one-way latency between state-of-the-art in 2014 versus 2013 was less than 100 microseconds. A second interpretation is that δ represents the speed difference between HFTs and sophisticated algorithmic trading firms that are not at the cutting edge of speed; in New York–Chicago trades, this difference might be a few milliseconds. Under the first interpretation of δ, when we plug in estimates for the other parameters in equation (8), we obtain a lower bound for τ on the order of 10–100 milliseconds. Under the second interpretation of δ we obtain a lower bound for τ on the order of 100 milliseconds to 1 second. Again, we make the caveat that the exercise is rough and at best gives a sense of magnitudes.

Online Appendix B.3.2 discusses a modification of the model in which, under frequent batch auctions, information arrives in discrete time rather than continuous time. The idea of this modification is that to the extent that information y about the value of security x is information about other security prices, then the use of frequent batch auctions would cause information to arrive in discrete time at frequency τ. Under this modification we obtain an equilibrium analogous to one we just got but with a simpler and less stringent sufficient condition under which frequent batch auctions stop the speed race: $\tau > \delta_{slow}$. Under this condition, any time there is a jump in y both slow and fast traders observe the jump in time for the next batch auction. This condition would point to a lower bound on τ on the order of 1–10 milliseconds.

VII.D. *Discussion of the Equilibria*

In this section, we make two sets of remarks concerning the equilibria of frequent batch auctions.

First, we discuss how the various cases we studied correspond to various potential implementations of frequent batch auctions. The exogenous entry case, studied in Section VII.C, is the right modeling device for scenarios in which the implementation of frequent batch auctions does not have a significant effect on the overall level of investment in speed. This could correspond to either a small-scale implementation of frequent batch auctions (e.g., a pilot test on a small number of stocks), which affects only a small proportion of the prize in the speed race, or a larger-scale implementation but in the short run during which speed investments are somewhat fixed. The endogenous entry case, studied in the previous section, is more appropriate for scenarios in which the implementation of frequent batch auctions would have a significant impact on trading firms' speed investment decisions. This would correspond to a larger-scale implementation of frequent batch auctions, in the medium to long run during which speed investments are flexible.

Second, we discuss what our analysis does and does not tell us about the choice of the batch interval. Both the discussion in Section VII.B and the equilibrium analysis for the exogenous entry case clarify that frequent batch auctions have important benefits over continuous limit order books even for exceptionally short τ. In the model, these benefits—the elimination of sniping, which in turn enhances liquidity—manifest for *any* $\tau > 0$. That is, there is a discontinuous benefit from switching from continuous time to discrete time. More practically, we think of this analysis as pertaining to any τ long enough to enable genuine batch processing of orders by traders responding to the same stimulus with essentially the same speed technology at essentially the same time. A batch interval of 1 nanosecond technically constitutes discrete time but would fail this practical test, because of randomness in computer response time, communications latencies, etc.

The discussion in Section VII.B and the equilibrium analysis for the endogenous entry case then clarify that a longer batch interval has an additional benefit over continuous limit order books, namely that it stops the arms race. In the model of Section VII.C, in which information arrives in continuous time, the batch interval τ should be long in relative terms compared to the increments at stake in the speed race δ. That is, the ratio $\frac{\delta}{\tau}$ should be sufficiently small, as per equation (8). In a modification of the model in which information arrives in discrete time, as discussed in Online Appendix B.3.2, the batch interval should be long in absolute terms compared to the speed of slow traders, δ_{slow}. The calibration exercise in Online Appendix B.3.1, although extremely rough, suggests that a batch interval on the order of 10 milliseconds or 100 milliseconds may be sufficient to stop the arms race by either measure.

Lengthening the batch interval may also have real costs, which we capture in a stylized way as investor delay costs.

Intuition suggests that such costs are small if the batch interval is small, and vanish to zero as the batch interval goes to zero.[39]

VIII. Alternative Responses to the HFT Arms Race

Policy discussions about the HFT arms race have suggested several alternative responses, most prominently Tobin taxes, minimum resting times, message-to-trade ratios, and random delays. In this section we briefly discuss each of these proposals. We also discuss a recent private sector market design innovation that is an alternative way to mitigate sniping.

VIII.A. Tobin Taxes

Tobin taxes (financial transactions taxes) were originally proposed as “sand in the gears” to curb perceived excessive speculation and excessive volatility in foreign exchange markets (see Tobin 1978; Summers and Summers 1989). More recently, Tobin taxes have been proposed as a response to the HFT arms race by Stiglitz (2014), among others,[40] and adopted in fall 2013 by Italy.

Tobin taxes can be formally modeled in our framework as follows. Introduce a Tobin tax of $\theta > 0$ per unit traded to the endogenous entry model of Section VI.D. For expositional simplicity assume that the tax is paid by the liquidity-taking side of the trade. This tax has two effects. The direct effect is that it simply increases the cost of trading by θ. The indirect effect is that by increasing the cost of trading the Tobin tax reduces the attractiveness of sniping opportunities. This in turn reduces entry by stale-quote snipers, which serves to reduce the equilibrium bid-ask spread and reduce equilibrium expenditure on speed. In Online Appendix B.2.1 we show formally that the Tobin tax (i) reduces investment in speed; (ii) reduces the sniping-cost component of transactions costs; and (iii) increases

[39] For example, suppose both options and stocks traded on frequent batch auction markets. Then liquidity providers in the option market, who, if traded against, seek to hedge in the underlying stock, would be exposed to delta risk for the length of the batch interval.

[40] The European Commission proposed a financial transactions tax in 2011. A Frequently Asked Questions document available on the EC website has a question “Who is most irritated by these taxation plans?”, the answer to which begins: “The taxation plans are, of course, most irritating for high-frequency traders and for fund and hedge fund managers whose business model is based on quick successions of financial transactions. . .” See European Commission (2013).

investors' all in trading costs, that is, from the perspective of investors the cost of the tax outweighs the benefit from less sniping. All three effects are monotonically increasing in θ. For intuition, consider the extreme case of a tax larger than the largest possible sniping opportunity; in this case there is no sniping, no investment in speed, and the equilibrium cost to investors to trade is simply the (very high) tax.

Hence, while the Tobin tax does address sniping and the HFT arms race, it achieves these benefits at the expense of making investors worse off.[41] A second caveat is that the Tobin tax is a relatively blunt instrument: to fully eliminate the incentive to invest in speed, the Tobin tax needs to be larger than the maximum possible sniping opportunity. In our ES-SPY data, reducing arms race profits by 90 percent would require a Tobin tax on the order of 10 basis points, or roughly 10 times the average SPY bid-ask spread in our data.[42]

Biais, Foucault, and Moinas (2015) argue for a tax directly on speed technology as opposed to trading. Such a tax can be modeled as increasing the cost of speed from c_{speed} to $c_{speed} + \theta_{speed}$, with θ_{speed} the level of the tax. Equations (5)–(6) imply that such a tax has no effect on investors' trading costs, while equation (7) implies that such a tax does reduce investment in speed. Hence, in our model, the Biais, Foucault, and Moinas (2015) tax is conceptually superior to the traditional Tobin tax. Again, though, the magnitudes necessary to meaningfully impact the arms race are large. To reduce arms race expenditures by 90 percent would require a tax of $\theta_{speed} = 9c_{speed}$, that is, a 900 percent tax on speed expenditures.

[41] Whether the Tobin tax enhances social welfare in our model depends on the interpretation of the social value of the tax revenue that the tax generates. If one assumes that a dollar of government revenue is as socially valuable as a dollar of investor profit, then the Tobin tax increases welfare. If the government uses the revenue from the Tobin tax to reduce other taxes that are distortionary, then the social welfare benefit would be higher; if the government wastes the money, then the net social welfare effect would be negative.

[42] This assumes that the arbitrageur pays the tax twice per arbitrage opportunity, once in ES and once in SPY. The total tax paid by the arbitrageur is 20 basis points.

VIII.B. "Bans" on HFT: Message Ratios and Minimum Resting Times

Two common characteristics of high-frequency trading strategies are (i) that HFTs often cancel their orders soon after placing them, and (ii) a high ratio of messages to completed trades.[43] Not coincidentally, two of the most widely discussed policy responses to the HFT arms race are minimum resting times and message-to-trade ratios. Minimum resting times prohibit canceling an order too soon after initial submission; orders must rest in the book for some minimum quantity of time, such as 500 milliseconds. Message-to-trade ratios prohibit having a ratio of messages to completed trades that is above some maximum threshold.

We wish to make two points about these proposals. First, these proposals seem to misunderstand cause and effect. Our model shows that both of these characteristics of HFT trading strategies are part of equilibrium behavior under the continuous limit order book. Liquidity providers cancel their orders and replace them with new orders every time there is a jump in y. Stale-quote snipers cancel their orders whenever their attempt to snipe does not win the race. See also Baruch and Glosten (2013) who analyze additional reasons why the continuous market may lead to what they call "flickering quotes."

Second, minimum resting times seem likely to exacerbate rather than reduce sniping. Specifically, if there is a jump in y that is within the resting time of the previous jump, then liquidity providers with stale quotes in the book are simply prohibited from attempting to cancel their stale quotes, ensuring that they will be sniped with probability 1.

VIII.C. Random Message Delays

Random message delays are described by Harris (2012) as follows: "Regulatory authorities could require that all exchanges delay the processing of every posting, canceling, and taking instruction they receive by a random period of between 0 and 10 milliseconds." The idea is that millisecond-level randomness dwarfs any microsecond-level differences in speed among trading

[43] The SEC's Concept Release on Equity Market Structure listed these as two of five common characteristics of HFT strategies, along with the use of speed technology, the use of colocation, and ending the trading day close to flat (Securities and Exchange Commission 2010).

firms responding to the same stimulus, which in turn reduces the incentive to invest in tiny speed improvements. Although intuitively appealing, there are two important concerns with random message delays. First, random message delays do not address sniping. If a liquidity provider attempts to cancel a stale quote, and other trading firms attempt to snipe a stale quote, the random message delay just adds an additional source of randomness regarding whose request is processed first. If there are N firms approximately equally fast, each of whom send one message, and the random message delay is large relative to any differences in speed among the N firms, then the liquidity provider will get sniped with probability of approximately $\frac{N-1}{N}$, just as in our model without random delay.

Second, random message delays incentivize trading firms to submit redundant messages. Consider our model of Section VI.B, modified to include a random message delay that is a uniform random draw from $[0, \epsilon]$. Suppose there is a jump in y that causes a liquidity provider's quotes to become stale. Then each of the other $N-1$ trading firms has incentive to submit not just one message to snipe, but many, because each message to snipe is like a lottery ticket hoping to get a short random delay. Similarly, the liquidity provider has incentive to submit not just one but many messages to cancel their stale quote, again in the hopes that one of these messages will get processed with delay 0.[44]

We show both of these points formally in Online Appendix B.2.2. Adding a random message delay to our model of the continuous limit order book with endogenous entry in Section VI.D has no effect on sniping, equilibrium expenditure on speed, or the bid-ask spread; the only effect is to encourage redundant message traffic.

VIII.D. *Asymmetric Delay to Immediately Executable Orders*

The previous section discussed why random message delays do not address sniping and encourage redundant message traffic. Consider, however, the following alternative, which captures the

[44] A natural idea in response to this concern is to place a cap on the number of redundant messages any one firm can send, for instance, a cap of one message per firm. However, such a cap would at best have no effect on sniping and could actually exacerbate sniping. The reason is that the cap would certainly bind for liquidity providers, whose message to cancel is tied to a specific quote of theirs in the book, whereas stale-quote snipers could circumvent a message cap by using multiple trading accounts.

key idea of recent market design innovations by TMX and IEX: apply a deterministic but asymmetric delay of $\Delta > 0$ only to immediately executable orders.[45,46] If immediately executable orders are delayed but posting and canceling messages are not, then, when there is a jump in y, liquidity providers have a head start over stale-quote snipers in the race to react. In the model of Section VI.D, it is straightforward to see that if the delay Δ exceeds the difference in speed δ between fast and slow trading firms, then slow trading firms can provide liquidity without risk of being sniped by fast trading firms. Recent work by Baldauf and Mollner (2014) shows this formally.

Hence, in our model of Section VI, the asymmetric delay eliminates sniping and stops the arms race. However, there are two important disadvantages of the asymmetric delay relative to frequent batch auctions, each of which can be captured with simple extensions of our model. Both disadvantages stem from the fact that the continuous limit order book with asymmetric delay is still a continuous-time serial-process market design, and as a result cannot eliminate the incentive to be a tiny bit faster than the competition.

[45] The key details of the TMX Group's proposed TSX Alpha Exchange are as follows. There is an order type called Post Only that can be entered and canceled without delay. The two requirements on Post Only orders are (i) that they be nonexecutable at the time of submission, and (ii) that their quantity exceed a minimum threshold. All other orders and cancels are subject to a delay, called a speed bump. The length of the delay is random, which our analysis in Section VIII.C suggests may not be wise. For more details on the proposed rules see http://www.osc.gov.on.ca/documents/en/Marketplaces/alpha-exchange_20141106_amd-request-for-comments.pdf.

[46] The key details of the IEX Alternative Trading System are as follows (see IEX Group 2014). There is a 350-microsecond delay applied symmetrically to all orders and cancels. In addition, there is price-sliding logic that adjusts stale quotes in the order book based on updates to the National Best Bid and Offer (NBBO) coming from otherUSequity exchanges. The rule is that any order present in the IEX limit order book that is priced more aggressively than the NBBO midpoint slides to the NBBO midpoint. Since IEX receives updates to the NBBO faster than the 350-microsecond delay (latency in the NBBO is on the order of 200 microseconds, given the geographical distances between the different exchanges' data centers in New Jersey), the effect of the combination of the symmetric speed bump and price-sliding logic is economically similar to the effect of an asymmetric delay. Formally, if innovations in y are interpreted as innovations in the NBBO, then the IEX market design eliminates sniping in our model just as does the asymmetric delay.Apotential concern about this market design is that it only mitigates sniping to the extent that prices are discovered via the NBBO exchanges rather than IEX.

First, the asymmetric delay does not address the race to the top of the book; see Yao and Ye (2014) and Moallemi (2014) for analyses of this component of the speed race. Formally, consider a modification to the model of Section VI in which trades can only occur at prices on a discrete price grid, with the increment denoted $\iota > 0$. Suppose that ι is large relative to the bid-ask spread that would obtain in the absence of a price constraint; this is a common case in practice (Yao and Ye 2014). In this case, in the continuous market, trading firms strictly prefer the role of liquidity provider to the role of stale-quote sniper (see note 19). In equilibrium, after jumps in y, there are races both to snipe stale quotes and to be at the top of the queue to provide liquidity at the new price level. Frequent batch auctions address both races. The advantage a fast trading firm has over a slow trading firm with respect to obtaining priority in the order book is proportional to $\frac{\delta}{\tau}$,[47] just as is the advantage a fast trading firm has over a slow trading firm with respect to sniping stale quotes. By contrast, the asymmetric delay has zero effect on the race to the top of the book.

Second, the asymmetric delay does not transform competition on speed into competition on price if there are quotes in the book that become stale based on public information and are not updated within Δ; for example, nonmarketable limit orders submitted by participants more than Δ slower than the cutting edge. Formally, consider a modification of our model in Section VI in which some investors have latency $\delta_{slower} > \Delta$ and attempt to satisfy their demand to trade by buying at the bid or selling at the ask. This behavioral type captures the idea that some investors attempt to trade without paying the bid-ask spread even though their monitoring technology is meaningfully slower than the cutting edge. Suppose a behavioral type's quote becomes stale based on a sufficiently large jump in the public signal y. In the continuous limit order book with asymmetric delay, the stale quote induces a race to snipe; whichever trading firm reacts first gets to trade at the stale price. In the frequent batch auction market the stale quote induces competition on price and will get filled at a price determined by the batch auction based on the new public

[47] The fast trading firm obtains time priority in the book over the slow trading firm only if their order reaches the order book in an earlier batch interval. Hence, just as depicted in Figure VII, it is only jumps in y that occur during a $\frac{\delta}{\tau}$ proportion of the batch interval that give a time priority advantage to the fast trading firm.

information rather than at the stale price. Put differently, non-HFTs can provide liquidity in frequent batch auctions without getting sniped, even if they are more than Δ slower than the cutting edge, whereas they would get sniped by HFTs in the continuous limit order book with asymmetric delay.

IX. Computational Advantages of Discrete-Time Trading

Our theoretical argument for frequent batch auctions as a response to the HFT arms race focuses on sniping, liquidity, and socially wasteful expenditure on speed. Practitioners and policy makers have argued that another important cost of the HFT arms race is that it is destabilizing for financial markets, making the market more vulnerable to extreme events such as the Flash Crash.[48] Although an analysis of the effect of frequent batch auctions on market stability is beyond the scope of the present article, here we discuss several computational simplicity advantages of discrete-time trading over continuous-time trading. As we note in the conclusion, we think that market stability is an important topic for future research.

First, frequent batch auctions are computationally simple for exchanges. Uniform-price auctions are fast to compute,[49] and

[48] Duncan Niederauer, former CEO of NYSE Euronext, testified to Congress on market structure issues including the Flash Crash that “there is reason for Congress and the SEC to be concerned that without action, we leave ourselves open to a greater loss of investor confidence and market stability. To solve the problem, policymakers should focus on establishing fairer and more transparent equity markets, as well as a more level playing field among trading centers and investors” (Niederauer 2012). See also the report on the regulatory response to the Flash Crash prepared by the Joint CFTC-SEC Advisory Committee on Emerging Regulatory Issues (SEC and CFTC 2010), the CFTC Concept Release on Risk Controls and System Safeguards for Automated Trading (Commodity Futures Trading Commission 2013), and policy papers by Haldane (2011) and Farmer and Skouras (2012).

[49] Formally, the processing time of the uniform-price auction is $O(n \log n)$, where n is the number of orders. Sorting bids and asks to compute the demand and supply curve is $O(n \log n)$ (Cormen et al. 2009), and then walking down the demand curve and up the supply curve to compute the market clearing price is $O(n)$. We also ran some simple computational simulations of uniform-price auctions, using randomly generated bids and asks, on an ordinary laptop using C++. We found that a uniform-price auction with 250,000 orders—the rate of messages per second during the flash crash according to a Nanex analysis (2011)—cleared in about 10 milliseconds in this simple computational environment.

exchange computers can be allocated a discrete block of time during which to perform this computation.[50] By contrast, in the continuous limit order book market design, exchange computers are not allocated a block of time during which to perform order processing, but instead process orders and other messages serially on arrival. While processing any single order is computationally trivial, even a trivial operation takes strictly positive computational time, which implies that during surges of activity there will be backlog and processing delay. This backlog can lead to confusion for trading algorithms, which are temporarily left uncertain about the state of their own orders. Moreover, backlog is most severe at times of especially high market activity, when reliance on low-latency information is also at its highest; Facebook's initial public offering on NASDAQ and the Flash Crash are salient examples (Nanex 2011; Jones 2013; Strasburg and Bunge 2013).

A second computational simplicity benefit of frequent batching is that it gives algorithmic traders a discrete block of time between when they receive a message—for example, a trade notification or an order book update—and by when they must make a decision, for example, submit a new order. In the continuous market, by contrast, trading algorithms are incentivized to react as fast as possible whenever they receive a new piece of information. This means, first, that trading algorithms are incentivized to trade off "smarts" for speed, that is, to make trading decisions based on only partial information and with only simple economic logic, since incorporating additional information and using more complicated economic logic each take time. And, second, that trading algorithms are incentivized to trade off error and risk checking for speed, because error and risk checking each take time and even tiny speed advantages can matter.[51] While discrete time certainly will not prevent trading firms from making programming errors (e.g., the Knight Capital incident of August 2012, see Strasburg and Bunge 2012), it does reduce the incentive to sacrifice robustness for speed.

[50] For instance, with a 100 millisecond batch interval, the first 10 milliseconds of each batch interval could be allocated to the exchange computers for computing and reporting outcomes from the previous batch interval.

[51] The sociologist Donald MacKenzie (2014) provides several detailed examples of this trade off between code robustness and speed described to him in interviews with high-frequency traders. For example, one trader is quoted "There are rules you need to follow to write fast code. Don't touch the kernel. Don't touch main memory . . . . Don't branch."

Third, discrete time simplifies the market paper trail for regulators and other market observers. In the continuous-time market, figuring out the precise sequence of market events is difficult to impossible. Exchange timestamps are always somewhat noisy, due to various processing delays including backlog, which means that the sequence of time stamps across exchanges may not reflect the actual sequence of events. Further complicating the paper trail is the need to adjust for relativity—even perfect time stamps do not reveal the sequence of events because the sequence of events depends on the location of the observer. In the discrete-time market the paper trail is much simpler: the regulatory authorities observe everything that happens at time t, $t+\tau$, $t+2\tau$, etc. As long as the batch interval τ is long relative to the imprecision in time stamps and latency across exchanges, the complexities that affect the continuous market's paper trail become nonissues. As evidence for the potential importance of a simple paper trail, consider that it took months of analysis for regulators to understand the basic sequence of events that caused the Flash Crash (SEC and CFTC 2010), and even today our understanding of that day's events remains incomplete.

Last, discrete time makes it technologically possible to disseminate public information symmetrically. In the continuous-time market, it is technologically difficult to disseminate information in such a way that all market participants who wish to receive it do so at the same time. Two recent examples that have attracted considerable attention are the SEC's difficulty with symmetric dissemination of corporate filings (Rogers, Skinner, and Zechman 2014) and the discrepancy in latency between direct exchange feeds and the SIP feed (Ding, Hanna, and Hendershott 2014; see also Lewis 2014). Moreover, even if information could be disseminated in such a way that all market participants who wish to receive it do so at exactly the same time, our model in Section VI shows that, economically, the continuous-time market processes their responses to the information as if the information were asymmetric. In contrast, in the discrete-time market it is technologically simple both to disseminate information to many participants at the same time and to process their responses at the same time.

In a sense, continuous-time markets implicitly assume that computers and communications are infinitely fast. Computers are fast but not infinitely so. Discrete time respects the limits of computers.

X. Conclusion

This article argues that the continuous limit order book is a flawed market design and proposes a new market design, frequent batch auctions, which directly addresses the flaw. To recap, our basic argument is as follows. First, we show empirically that the continuous limit order book market design does not really "work" in continuous time: correlations completely break down at high-frequency time scales, which leads to obvious mechanical arbitrage opportunities. The time series evidence suggests that the arms race profits should be thought of more as a constant of the market design, rather than as a prize that is competed away over time. Next, we build a simple theoretical model guided by these empirical facts. We show that the mechanical arbitrage opportunities we observed in the data are in a sense "built in" to the market design: even symmetrically observed public information creates arbitrage rents. These rents come at the expense of liquidity provision, as measured by both bid-ask spreads and market depth, and induce a never-ending arms race for speed. Last, we show that frequent batch auctions eliminate the mechanical arbitrages and the HFT arms race, which in turn enhances liquidity and, unless investors are extremely impatient, improves social welfare. Discrete time makes tiny speed advantages orders of magnitude less valuable, and the auction transforms competition on speed into competition on price.

There are several important directions for future research. First, our model is extremely stylized. This level of abstraction is appropriate both for making stark the key design flaw of the continuous limit order book and for articulating why frequent batch auctions directly address the flaw. However, future analysis of frequent batch auctions should be conducted in a richer modeling environment, ideally including features such as asymmetric information, inventory management considerations, multileg trades, and investors needing to trade large quantities over time. Among other things, such a model would help shed light on the optimal batch interval.

A second area for future research is the nature of competition among exchanges. Suppose that one or more exchanges adopt frequent batch auctions while other exchanges continue to use continuous trading: what is the equilibrium? Can an entrant exchange that adopts frequent batch auctions attract market share? We note that these questions may also be related to the analysis of the optimal batch interval. They may have implications for regulatory policy as well.

A third topic for future research is the effect of frequent batch auctions on market stability. In Section IX we discussed several computational advantages of discrete-time trading over continuous-time trading. For example, the market paper trail becomes simpler because issues that complicate the paper trail in continuous time—exchange and communication latency, clock synchronization, the discrepancy between direct feeds and the SIP feed, relativity—are nonissues in discrete time. However, we caution that this discussion was necessarily informal and speculative. Further research is needed, especially to understand whether and to what extent computational simplicity reduces the market's vulnerability to the kinds of extreme events at the center of the debate on the effect of HFT on market stability.

University Of Chicago
University Of Maryland
University Of Chicago

Supplementary Material

An Online Appendix for this article can be found at QJE online (qje.oxfordjournal.org).

References

Adler, Jerry, "Raging Bulls: How Wall Street Got Addicted to Light-Speed Trading," *Wired Magazine*, August 2012, available at http://www.wired.com/business/2012/08/ff_wallstreet_trading/.

Angel, James J., Lawrence E. Harris, and Chester S. Spatt, "Equity Trading in the 21st Century: An Update," *Quarterly Journal of Finance*, 5 (2015), 1–39.

Baldauf, Markus, and Joshua Mollner, "High-Frequency Trade and Market Performance," Working Paper, 2014.

Baruch, Shmuel, and Lawrence R. Glosten, "Fleeting Orders," Working Paper, 2013.

Biais, Bruno, and Thierry Foucault, "HFT and Market Quality," *Bankers, Markets & Investors*, 128 (2014), 5–19.

Biais, Bruno, Theirry Foucault, and Sophie Moinas, "Equilibrium Fast Trading," *Journal of Financial Economics*, 116 (2015), 292–313.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, "High Frequency Trading and Price Discovery," *Review of Financial Studies*, 27 (2014a), 2267–2306.

———, "High Frequency Trading and the 2008 Short Sale Ban," Working Paper, 2014b.

Budish, Eric, Peter Cramton, and John Shim, "Implementation Details for Frequent Batch Auctions: Slowing Down Markets to the Blink of an Eye," *American Economic Review: Papers and Proceedings*, 104 (2014), 418–424.

Bunge, Jacob, “CME, Nasdaq Plan High-Speed Network Venture,” *Wall Street Journal*, March 28, 2013, available at http://online.wsj.com/article/SB10001424127887324685104578388343221575294.html.

Cinnober, “Using Adaptive Micro Auctions to Provide Efficient Price Discovery When Access in Terms of Latency Is Differentiated among Market Participants,” White Paper, 2010.

Clark-Joseph, Adam, “Exploratory Trading,” Working Paper, 2013.

Cohen, Kalman J., and Robert A. Schwartz, “The Challenge of Information Technology for the Securities Markets: Liquidity, Volatility and Global Trading,” in *An Electronic Call Market: Its Design and Desirability*, Henry Lucas and Robert Schwartz, eds. (Homewood, IL: Dow Jones-Irwin, 1989).

Commodity Futures Trading Commission, “Concept Release on Risk Controls and System Safeguards for Automated Trading Environments,” 2013.

Conway, Brendan, “Wall Street’s Need for Trading Speed: The Nanosecond Age,” *Wall Street Journal*, June 14, 2011, available at http://blogs.wsj.com/marketbeat/2011/06/14/wall-streets-need-for-trading-speed-the-nanosecond-age/.

Copeland, Thomas E., and Dan Galai, “Information Effects on the Bid-Ask Spread,” *Journal of Finance*, 38 (1983), 1457–1469.

Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms*, 3rd ed. (Cambridge, MA: MIT Press, 2009).

Demsetz, Harold, “The Cost of Transacting,” *Quarterly Journal of Economics*, 82 (1968), 33–53.

Ding, Shengwei, John Hanna, and Terrence Hendershott, “How Slow Is the NBBO? A Comparison with Direct Exchange Feeds,” *Financial Review*, 49 (2014), 313–332.

Duffie, Darrell, Nicolae Garleanu, and Lasse Heje Pedersen, “Over-the-Counter Markets,” *Econometrica*, 73 (2005), 1815–1847.

Economic Sciences Prize Committee, “Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2013: Understanding Asset Prices,” Royal Swedish Academy of Sciences, 2013.

Economides, Nicholas, and Robert A. Schwartz, “Electronic Call Market Trading: Let Competition Increase Efficiency,” *Journal of Portfolio Management*, 21 (1995), 10–18.

Einstein, Albert, “Zur Elektrodynamik bewegter Körper (On the Electrodynamics of Moving Bodies),” *Annalen der Physik*, 322 (1905), 891–921.

Epps, Thomas, “Comovements in Stock Prices in the Very Short Run,” *Journal of the American Statistical Association*, 74 (1979), 291–298.

European Commission, “FTT—Non-technical Answers to Some Questions on Core Features and Potential Effects,” 2013.

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

Farmer, J. Doyne, and Spyros Skouras, “Review of the Benefits of a Continuous Market vs. Randomised Stop Auctions and of Alternative Priority Rules (Policy Options 7 and 12),” UK Government’s Foresight Project, The Future of Computer Trading in Financial Markets, Economic Impact Assessment EIA11, 2012.

Foucault, Thierry, “Order Flow Composition and Trading Costs in a Dynamic Limit Order Market,” *Journal of Financial Markets*, 2 (1999), 99–134.

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham, “Toxic Arbitrage,” CEPR Discussion Papers 9925, 2014.

Foucault, Thierry, Ailsa Roell, and Patrik Sandas, “Market Making with Costly Monitoring: An Analysis of the SOES Controversy,” *Review of Financial Studies*, 16 (2003), 345–384.

Frazzini, Andrea, Ronen Israel, and Tobias J. Moskowitz, “Trading Costs of Asset Pricing Anomalies,” Fama-Miller Working Paper, Chicago Booth Research Paper No. 14-05, 2012.

Glosten, Lawrence R., “Is the Electronic Open Limit Order Book Inevitable?,” *Journal of Finance*, 49 (1994), 1127–1161.

Glosten, Lawrence R., and Paul Milgrom, “Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders,” *Journal of Financial Economics*, 14 (1985), 71–100.

Goettler, Ronald L., Christine A. Parlour, and Uday Rajan, “Equilibrium in a Dynamic Limit Order Market,” *Journal of Finance*, 60 (2005), 2149–2192.

Haldane, Andrew, “The Race To Zero,” *Speech at the International Economic Association Sixteenth World Congress*, Beijing, July 8, 2011.

Harris, Jeffrey, and Paul Schultz, “The Trading Profits of SOES Bandits,” *Journal of Financial Economics*, 50 (1998), 39–62.

Harris, Larry, *Trading and Exchanges: Market Microstructure for Practitioners* (Oxford: Oxford University Press, 2002).

———, “Stop the High-Frequency Trader Arms Race,” *Financial Times*, December 27, 2012.

Hasbrouck, Joel, and Gideon Saar, “Low-Latency Trading,” *Journal of Financial Markets*, 16 (2013), 646–679.

Hendershott, Terrence, Charles Jones, and Albert Menkveld, “Does Algorithmic Trading Improve Liquidity?,” *Journal of Finance*, 66 (2011), 1–33.

Hirshleifer, Jack, “The Private and Social Value of Information and the Reward to Inventive Activity,” *American Economic Review*, 61 (1971), 561–574.

IEX Group, “Form ATS: Initial Operation Report, Amendment to Initial Operation Report and Cessation of Operations Report for Alternative Trading Systems,” 2014.

ISN, “Toward a Fairer and More Efficient Market,” ISN Research Report, 2013.

Jones, Charles, “What Do We Know About High-Frequency Trading?,” Columbia University Working Paper, 2013.

Klemperer, Paul, *Auctions: Theory and Practice* (Princeton, NJ: Princeton University Press, 2004).

Kyle, Albert S., “Continuous Auctions and Insider Trading,” *Econometrica*, 53 (1985), 1315–1335.

Laughlin, Gregory, Anthony Agiurre, and Joseph Grundfest, “Information Transmission between Financial Markets in Chicago and New York,” *Financial Review*, 49 (2014), 283–312.

Lewis, Michael, *Flash Boys: A Wall Street Revolt* (New York: Norton, 2014).

MacKenzie, Donald, “Be Grateful for Drizzle,” *London Review of Books*, 36 (2014), 27–30.

Madhavan, Ananth, “Trading Mechanisms in Securities Markets,” *Journal of Finance*, 47 (1992), 607–641.

McPartland, John, “Recommendations for Equitable Allocation of Trades in High Frequency Trading Environments,” *Journal of Trading*, 10 (2015), 81–100.

Menkveld, Albert J., and Marius A. Zoican, “Need for Speed? Exchange Latency and Liquidity,” Tinbergen Institute Discussion Papers: 140-097/IV, 2014.

Milgrom, Paul, *Putting Auction Theory to Work* (Cambridge: Cambridge University Press, 2004).

———, “Critical Issues in the Practice of Market Design,” *Economic Inquiry*, 49 (2011), 311–320.

Moallemi, Ciamac, “The Value of Queue Position in a Limit Order Book,” Working Paper, 2014.

Najarian, Jon A., “The Ultimate Trading Weapon,” 2010, available at http://moneytalks.net/pdfs/37895070-The-Ultimate-Trading-Weapon.pdf

Nanex, “CQS Was Saturated and Delayed on May 6th, 2010,” July 25, 2011, available at http://www.nanex.net/Research/NewFlashCrash1/FlashCrash.CQS.Saturation.html.

———, “Dangerous Order Types,” November 15, 2012, available at http://www.nanex.net/aqck2/3681.html.

Niederauer, Duncan, “Market Structure: Ensuring Orderly, Efficient, Innovative and Competitive Markets for Issuers and Investors: Congressional Hearing Before the Subcommittee on Capital Markets and Government Sponsored Enterprises of the Committee on Financial Services US House of Representatives, 112th Congress,” Congressional Testimony, Panel I, 2012, available at http://financialservices.house.gov/uploadedfiles/112-137.pdf.

O’Hara, Maureen, “High Frequency Market Microstructure,” *Journal of Financial Economics*, 116 (2015), 257–270.

Patterson, Scott, and Jenny Strasburg, “How ‘Hide Not Slide’ Orders Work,” *Wall Street Journal*, September 18, 2012, available at http://online.wsj.com/article/SB10000872396390444812704577605840263150860.html.

Patterson, Scott, Jenny Strasburg, and Liam Pleven, “High-Speed Traders Exploit Loophole,” *Wall Street Journal*, May 1, 2013.

Rogers, Jonathan L., Douglas J. Skinner, and Sarah L. C. Zechman, “Run EDGAR Run: SEC Dissemination in a High-Frequency World,” Working Paper, 2014.

Rogow, Geoffrey, “Colocation: The Root of all High-Frequency Trading Evil?,” *Wall Street Journal*, September 20, 2012, available at http://blogs.wsj.com/market-beat/2012/09/20/collocation-the-root-of-all-high-frequency-trading-evil/.

Roth, Alvin E., “The Economist as Engineer: Game Theory, Experimentation and Computation as Tools for Design Economics,” *Econometrica*, 70 (2002), 1341–1378.

———, “What Have We Learned from Market Design?,” *Economic Journal*, 118 (2008), 285–310.

Roth, Alvin E., and Axel Ockenfels, “Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet,” *American Economic Review*, 92 (2002), 1093–1103.

Roth, Alvin E., and Xiaolin Xing, “Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions,” *American Economic Review*, 84 (1994), 992–1044.

———, “Turnaround Time and Bottlenecks in Market Clearing: Decentralized Matching in the Market for Clinical Psychologists,” *Journal of Political Economy*, 105 (1997), 284–329.

Sannikov, Yuliy, and Andrzej Skrzypacz, “Dynamic Trading: Price Inertia, Front-Running and Relationship Banking,” Working Paper, 2014.

Schwartz, Robert, ed., *The Electronic Call Auction: Market Mechanism and Trading* (Boston: Kluwer Academic, 2001).

Schwartz, Robert A., and Liuren Wu, “Equity Trading in the Fast Lane: The Staccato Alternative,” *Journal of Portfolio Management*, 39 (2013), 3–6.

SEC and CFTC, “Findings Regarding the Market Events of May 6, 2010,” *Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*, 10, (2010), 2012.

Securities and Exchange Commission, “Concept Release on Equity Market Structure,” 2010.

Sparrow, Chris, “The Failure of Continuous Markets,” *Journal of Trading*, 7 (2012), 44–47.

Steiner, Christopher, “Wall Street’s Speed War,” *Forbes Magazine*, September 27, 2010.

Stiglitz, Joseph E., “Tapping the Brakes: Are Less Active Markets Safer and Better for the Economy?,” Presented at the Federal Reserve Bank of Atlanta 2014 Financial Markets Conference, 2014.

Stoll, Hans R., “The Supply of Dealer Services in Securities Markets,” *Journal of Finance*, 33 (1978), 1133–1151.

Strasburg, Jenny, and Jacob Bunge, “Loss Swamps Trading Firm,” *Wall Street Journal*, August 2, 2012.

———, “Nasdaq Is Still on Hook as SEC Backs Payout for Facebook IPO,” *Wall Street Journal*, March 25, 2013, available at http://online.wsj.com/article/SB10001424127887323466204578382193806926064.html.

Summers, Laurence H., and Victoria P. Summers, “When Financial Markets Work Too Well: A Cautious Case for a Securities Transactions Tax,” *Journal of Financial Services Research*, 3 (1989), 261–286.

Tobin, James, “A Proposal for International Monetary Reform,” *Eastern Economic Journal*, 4 (1978), 153–159.

Troianovski, Anton, “Networks Built on Milliseconds,” *Wall Street Journal*, May 30, 2012, available at http://online.wsj.com/article/SB10001424052702304065704577426500918047624.html.

Vayanos, Dimitri, “Strategic Trading and Welfare in a Dynamic Market,” *Review of Economic Studies*, 66 (1999), 219–254.

Virtu, “Form S-1: Virtu Financial, Inc.,” 2014.

Wah, Elaine, and Michael Wellman, “Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model,” *Proceedings of the Fourteenth ACM Conference: Electronic Commerce*, 2013.

Yao, Chen, and Mao Ye, “Tick Size Constraints, High-Frequency Trading, and Liquidity,” Working Paper, 2014.

This page intentionally left blank