The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response*

Eric Budish†, Peter Cramton‡, and John Shim§

February 3, 2015

Abstract

The high-frequency trading arms race is a symptom of flawed market design. Instead of the continuous limit order book (CLOB) that is currently predominant, we argue that financial exchanges should use frequent batch auctions: uniform price double auctions conducted, e.g., 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 CLOB market design 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 CLOB market design – 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 problems caused by the CLOB. 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.

*First version: July 2013. Project start date: Oct 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, and to 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. 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. University of Chicago Booth School of Business, eric.budish@chicagobooth.edu

‡University of Maryland, pcramton@gmail.com

§University of Chicago Booth School of Business, john.shim@chicagobooth.edu

1 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. 3 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 3000 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 glass. Since its emergence in around 2011, microwave technology has reduced round-trip transmission time first to around 10ms, then 9ms, then 8.5ms, and most recently to 8.1ms. 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 financial market design: *continuous-time trading.* That is, under the continuous limit order book (CLOB) market design that is currently predominant, it is possible to buy or sell stocks or other securities 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, of length, 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,

[1] Sources for this paragraph: "Wall Street's Speed War," Forbes, Sept 27th 2010; "The Ultimate Trading Weapon," ZeroHedge.com, Sept 21st 2010; "Wall Street's Need for Trading Speed: The Nanosecond Age," Wall Street Journal, June 2011; "Networks Built on Milliseconds," Wall Street Journal, May 2012; "Raging Bulls: How Wall Street Got Addicted to Light-Speed Trading," Wired, Aug 2012; "CME, Nasdaq Plan High-Speed Network Venture," Wall Street Journal March 2013; "Information Transmission between Financial Markets in Chicago and New York" by Laughlin et al (2014); McKay Brothers Microwave latency table, Jan 20th 2015, Aurora-Carteret route.

[2] Computers do not literally operate in continuous time; they operate in discrete time in increments of about 0.3 nanoseconds. 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.

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 of our paper uses millisecond-level direct-feed data from exchanges to document a series of stylized facts about the CLOB market design. Together, the facts suggest that the CLOB market design violates basic asset pricing principles at high-frequency time horizons; that is, the continuous-time market does not actually "work" in continuous time. Consider Figure 1.1. The figure depicts the price paths of the two largest securities 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 Panel A, we see that the two securities are nearly perfectly correlated over the course of the trading day, as we would expect given the near-arbitrage relationship between them. Similarly, the securities 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 whomever 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.

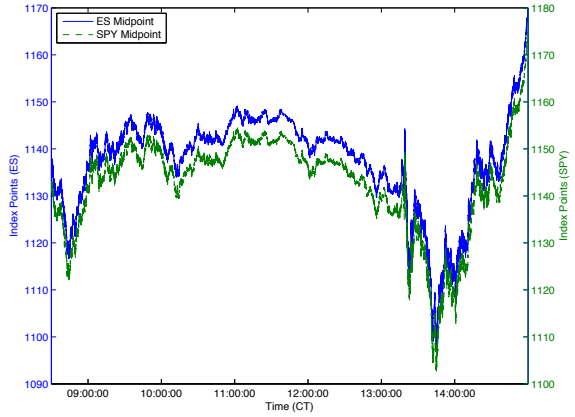
The usual economic intuition about obvious arbitrage opportunities is that, once discovered, competitive forces eliminate the inefficiency. But that is not what we find here. Over the time period of our data, 2005-2011, we find that the *duration* of ES-SPY arbitrage opportunities declines dramatically, from a median of 97ms in 2005 to a median of 7ms 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

[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 4ms 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 5 and Appendix B.1 for further details.

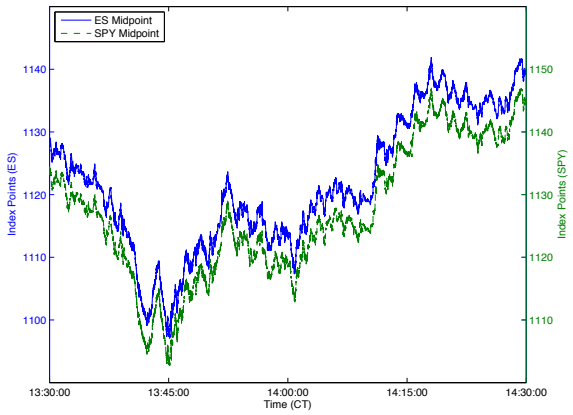
Figure 1.1: ES and SPY Time Series at Human-Scale and High-Frequency Time Horizons

Notes: 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 (Aug 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 securities due to differences in cost-of-carry, dividend exposure, and ETF tracking error; for details see Section 5.2.1. For details regarding the data, see Section 4.

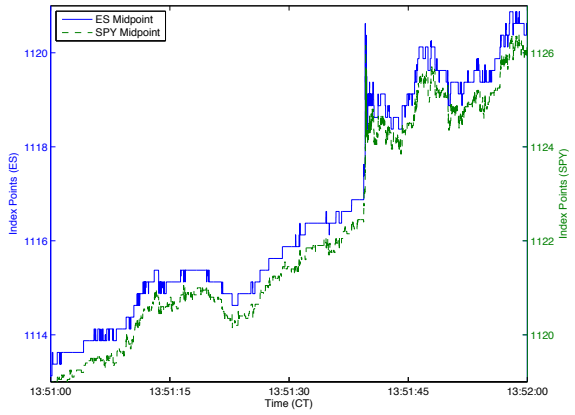
(a) Day



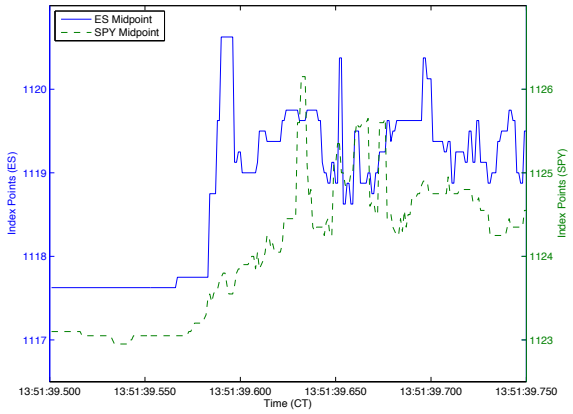
(b) Hour



(c) Minute



(d) 250 Milliseconds



decreasing over the time period 2005-2011; but, in all years, correlations are essentially zero at high-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 securities – there are hundreds if not thousands of other pairs of highly correlated securities, and, in fragmented equity markets, arbitrage trades that are even simpler, since the same stock trades on multiple venues. While we hesitate to put a precise estimate on the total size of the prize in the speed race, common sense extrapolation from our ES-SPY estimates suggests that the sums are substantial.

The second part of our paper presents a simple new theory model which is motivated by, and helps to explain and interpret, these empirical facts. The model serves as a critique of the CLOB market design, and it also articulates the economics of the HFT arms race. In the model, there is a security, x, that trades on a CLOB market, 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 CLOB, 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 CLOB 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 1.1 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 CLOB 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 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 CLOB market design. The obvious mechanical arbitrage opportunities we observed in the data are in a sense "built in" to the CLOB: *even symmetrically observed public information creates arbitrage rents.*

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 CLOB markets 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]

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 CLOB market design – a result which ties in closely with our empirical findings.

[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 6.3. Note too that while our model is extremely stylized, one thing we do not abstract from is the rules of the CLOB market design 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 CLOB. 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 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 (cf. Virtu, 2014, pg. 103; Angel, Harris and Spatt, 2013, pg. 23; and 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 6.5.4.

The third and final part of our paper shows that the frequent batch auction market design directly addresses the problems we have identified with the CLOB market design. Frequent batch auctions may sound like a very different market design from the CLOB, but there are really just two differences. First, time is treated as discrete instead of continuous.[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 to the CLOB. For instance, orders consist of 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 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 an auction – have two beneficial effects. First, they substantially reduce 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 CLOB market design, it is possible to earn a rent based on a piece of information that many traders observe at basically the same time, because CLOBs process orders in serial, and somebody 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 purely mechanical cost of liquidity provision in CLOB 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 associated with the CLOB market design, and in a manner that allocates the welfare savings to investors. In equilibrium, relative to the CLOB, frequent batch auctions eliminate sniping, enhance liquidity, and stop the HFT arms race.

We emphasize that the market design perspective we take in this paper sidesteps the "is HFT good or evil?" debate which seems to animate much of the current discussion about HFT among

[6]This paper does not characterize a specific optimal batch interval. See Section 7.4 and Appendix C.2 for a discussion of what the present paper's analysis does and does not teach us about the optimal batch interval.

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 which 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 – e.g., 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 paper is organized as follows. Section 2 discusses related literature. Section 3 briefly reviews the rules of the continuous limit order book. Section 4 describes our direct-feed data from NYSE and the CME. Section 5 presents the empirical results on correlation breakdown and mechanical arbitrage. Section 6 presents the model, and solves for and discusses the equilibrium of the CLOB. Section 7 analyzes frequent batch auctions, shows why they directly address the problems with the CLOB, and discusses their equilibrium properties. Section 8 discusses computational advantages of discrete-time trading over continuous-time trading. Section 9 concludes. Supporting materials are contained in a series of appendices. Appendix A uses our model to discuss alternative responses to the HFT arms race, including Tobin taxes, random delays, asymmetric delays, and minimum resting times. Appendix B provides backup materials for the empirical analysis. Appendix C provides backup materials for the theoretical analysis.

2 Related Literature

First, there is a well-known older academic literature on *in*frequent batch auctions, e.g., three times per day (opening, midday, and close). Important contributions to this literature include Cohen and Schwartz (1989); Madhavan (1992); 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 on 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

[7] In Economides and Schwartz (1995), the aggregation is achieved by conducting the auction at three significant points during the trading day (opening, 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.

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 CLOB to discrete-time batch-process auctions. The earlier literature then 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 that are 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 7.1 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 non-formal 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 CLOB 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 identifies the key flaws of the CLOB 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) departs more dramatically from the CLOB, 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 8. 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)

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

[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 (2013); ISN (2013); Schwartz and Wu (2013). In each case, the proposed market design either departs more dramatically from the CLOB than ours or there are important design details that are omitted.

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 above. 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); 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); Brogaard, Hendershott and Riordan (2014*a*,*b*); Hasbrouck and Saar (2013); Foucault, Kozhan and Tham (2014); Menkveld and Zoican (2014). We discuss the relationship between our results and aspects of this literature in Section 6.5.4. Biais, Foucault and Moinas (2013) study the equilibrium level of investment in speed technology in the context of a Grossman-Stiglitz style rational expectations 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. The Nasdaq "SOES Bandits" were an early incarnation of stale-quote snipers, in the context of an electronic market design that had an unusual feature – the prohibition of automated quote updates – that was exploited by the bandits. 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 paper.

Fourth, our paper 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 CLOB market design are new versions of classical ideas. Correlation breakdown is an extreme version of a phenomenon first documented by Epps (1979); see Section 5 for further discussion. Sniping, and its negative effect on liquidity, is closely related to Glosten and Milgrom (1985) adverse selection; see Section 6.3 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 a CLOB 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.

3 Brief Description of Continuous Limit Order Books

In this section we summarize the rules of the continuous limit order book (CLOB) 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 to sell, e.g., "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 also 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 1000 shares at $100.01 and 1000 shares at $100.02, a limit order to buy 1500 shares at $100.02 (or greater) would get filled by trading all 1000 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 1500 shares at $100.01 would get partially filled, by trading 1000 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 6.

In practice, there are many other order types that traders can use in addition to limit orders. These include market orders, stop-loss orders, fill-or-kill, 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.

4 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 dataset 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 Jan 1, 2005 - Dec 31, 2011. The NYSE dataset is called TAQ NYSE ArcaBook. While this data covers all US 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 Jan 1, 2005 - Dec 31, 2011, with the exception of a three-month gap from 5/30/2007-8/28/2007 resulting from data issues acknowledged to us by the NYSE data team. We also drop, from both datasets, 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 dataset's zip file is corrupted or truncated. We are left with 1560 trading days in total.

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.[11]

We emphasize that direct feed data are distinct from the so-called "regulatory feeds" provided by the exchanges to market regulators. In particular, the TAQ NYSE ArcaBook dataset is distinct from the more familiar TAQ NYSE Daily dataset (sometimes simply referred to as TAQ), which is an aggregation of orders and trades from all Consolidated Tape Association exchanges. The TAQ

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

[11] Our feed parsers will be made publicly available in the data appendix.

data is comprehensive in regards 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, regulatory feed data have time stamps that are based on the time at which the data are provided to the consolidated feed, and practitioners estimate that the TAQ's timestamps were 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.

5 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 1.1. The time-series evidence explores how the pattern depicted in Figure 1.1 has evolved over time.

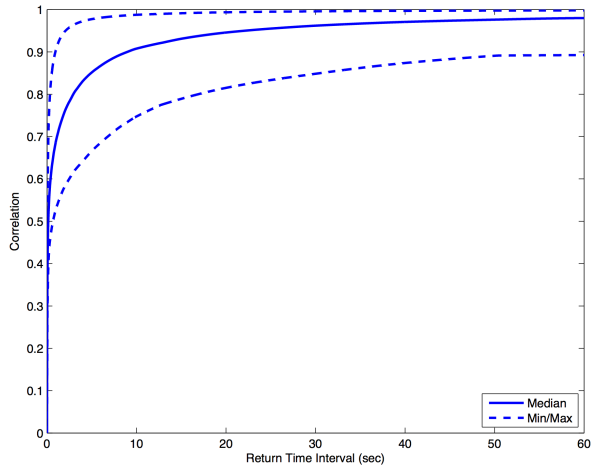
Before proceeding, we emphasize that the finding that correlations break down at high-enough frequency – which is an extreme version of a phenomenon discovered by Epps (1979)[12] – is obvious from introspection alone, at least ex-post. There is nothing in current market architecture – in which each security trades in continuous time on its own separate limit-order book, rather than in a single combinatorial auction market – that would allow different securities' prices to move at *exactly* the same time. It is only when we show that correlation breakdown is associated with frequent mechanical arbitrage opportunities that we can interpret it as a meaningful issue rather than a theoretical curiosity.

[12]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 ten minutes, and long meant a few days.

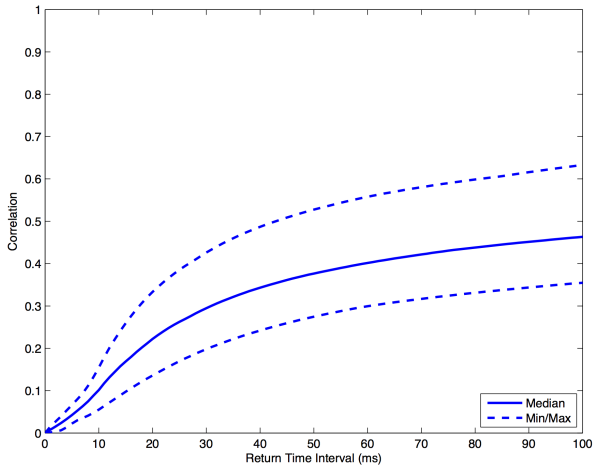
Figure 5.1: ES and SPY Correlation by Return Interval: 2011

Notes: 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 4.

(a) Correlations at Intervals up to 60 Seconds



(b) Correlations at Intervals up to 100 Milliseconds



5.1 Correlation Breakdown

5.1.1 Summary Statistics

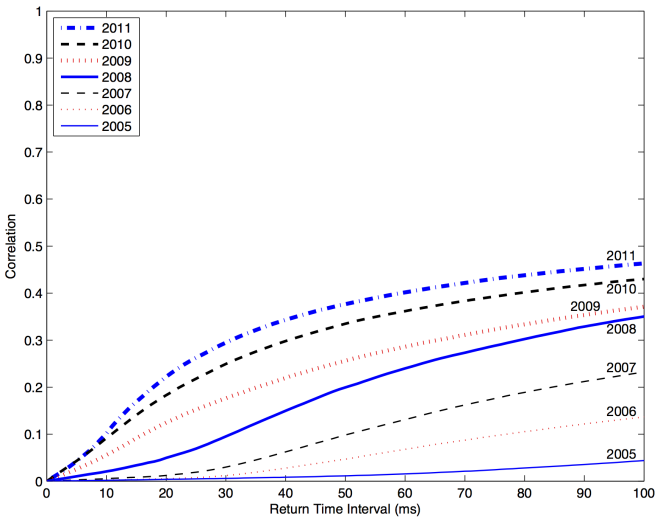
Figure 5.1 displays the median, min, and max daily return correlation between ES and SPY for time intervals ranging from 1 millisecond to 60 seconds, for our 2011 data, under our main specification for computing correlation. In this main specification, we compute the correlation of percentage changes 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,[13] 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 Appendix B.1.1. We also examine correlations for pairs of related equity securities in Appendix B.1.2, for which speed-of-light issues do not arise since they trade in the same physical location. In all cases,

[13]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 5.2: ES and SPY Correlation Breakdown Over Time: 2005-2011

Notes: 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 100ms. For years 2005-2008 the CME data is only at 10ms resolution, so we compute the median correlation for each multiple of 10ms and then fit a cubic spline. For more details on the data, refer to Section 4.



correlations break down at high frequency.

5.1.2 Correlation Breakdown Over Time

Figure 5.2 displays the ES-SPY correlation versus time interval curve that we depicted above as Figure 5.1(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 ms 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.

5.2 Mechanical Arbitrage

5.2.1 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 1.1d – buy cheap and sell expensive when one security 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 Appendix B.2.1. Here, we

mention the most important points.

First, there is a difference in levels between the two securities, 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, i.e., 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, e.g., due to a change in short-term interest rates.

Second, while Figure 1.1 depicts bid-ask midpoints, in computing the arbitrage opportunity we assume that the trader buys the cheaper security at its ask while selling the more expensive security 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.[14] Our arbitrageur only initiates a trade when the expected profit from doing so, accounting for bid-ask 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 4ms, the one-way speed-of-light travel time between New York and Chicago. Arbitrage opportunities that last fewer than 4ms 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.

5.2.2 Summary Statistics

Table 1 reports summary statistics on the ES-SPY arbitrage opportunity over our full dataset, 2005-2011.

An average day in our dataset has about 800 arbitrage opportunities, while an average arbitrage

[14] 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.

Table 1: ES-SPY Arbitrage Summary Statistics, 2005-2011

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 5.2.1 and Appendix B.2.1. The data, described in Section 4, cover January 2005 to December 2011. Variables are described in the text of Section 5.2.2.

		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	$927.07
Total Daily Profits - NYSE Data ($)	$79k	$5k	$9k	$18k	$33k	$57k	$204k	$554k
Total Daily Profits - All Exchanges ($)	$306k	$27k	$39k	$75k	$128k	$218k	$756k	$2,333k
% ES Initiated	88.56%							
% Good Arbs	99.99%							
% Buy vs. Sell	49.77%							

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 $79k per day, with profits on a 99th percentile day of $554k. 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.[15] This yields average profits of $306k per day, or roughly $75M per year. We discuss the total size of the arbitrage opportunity in more detail below in Section 5.3.

88.56% of the arbitrage opportunities in our dataset are initiated by a price change in ES, with the remaining 11.44% 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 Appendix

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

Table 2 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.

99.99% 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.

5.2.3 Mechanical Arbitrage Over Time: 2005-2011

In this sub-section we explore how the ES-SPY arbitrage opportunity has evolved over time.

Figure 5.3 explores the duration of ES-SPY arbitrage opportunities over the time of our data set, covering 2005-2011. As can be seen in Figure 5.3a, the median duration of arbitrage opportunities has declined dramatically over this time period, from a median of 97 ms in 2005 to a median of 7 ms in 2011. Figure 5.3b 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 10ms and most lasted at least 50ms, whereas by 2011 essentially none lasted 50ms and very few lasted even 10ms.

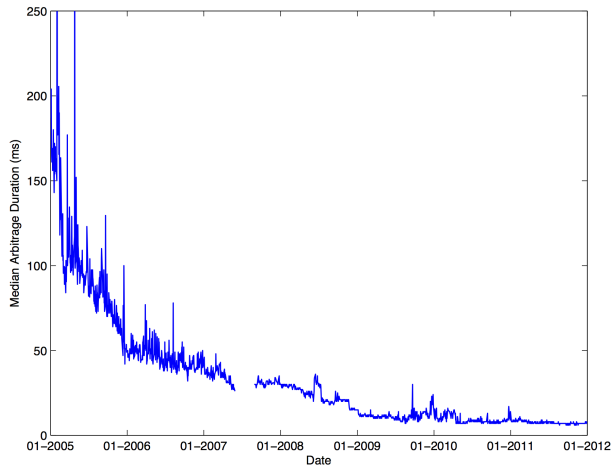
Figure 5.4 explores the per-arbitrage profitability of ES-SPY arbitrage opportunities over the time of our data set. In contrast to arbitrage durations, arbitrage profits have remained remarkably constant over time. Figure 5.4a shows that the median profits per contract traded have remained steady at around 0.08 index points, with the exception of the 2008 financial crisis when they were a bit larger. Figure 5.4b 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 5.5 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 5.5a 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 5/6/2010, 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. Figure 5.5b confirms this intuition formally. The figure plots the number of arbitrage opportunities on a given trading day against a measure 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 variable explains nearly all of

Figure 5.3: Duration of ES & SPY Arbitrage Opportunities Over Time: 2005-2011

Notes: Panel (a) shows the median duration of arbitrage opportunities between the E-mini S&P 500 future (ES) and the SPDR S&P 500 ETF (SPY) from January 2005 to December 2011. Each point represents the median duration of that day's arbitrage opportunities. The discontinuity in the time series (5/30/2007-8/28/2007) arises from omitted data resulting from data issues acknowledged by the NYSE. Panel (b) plots arbitrage duration against the proportion of arbitrage opportunities lasting at least that duration, for each year in our dataset. Panel (b) restricts attention to arbitrage opportunities with per-unit profits of at least 0.10 index points. We drop arbitrage opportunities that last fewer than 4ms, which is the one-way speed-of-light travel time between New York and Chicago. Prior to Nov 24, 2008, we drop arbitrage opportunities that last fewer than 9ms, which is the maximum combined effect of the speed-of-light travel time and the rounding of the CME data to centiseconds. See Section 5.2.1 for further details regarding the ES-SPY arbitrage. See Section 4 for details regarding the data.

(a) Median Arb Durations Over Time



(b) Distribution of Arb Durations Over Time

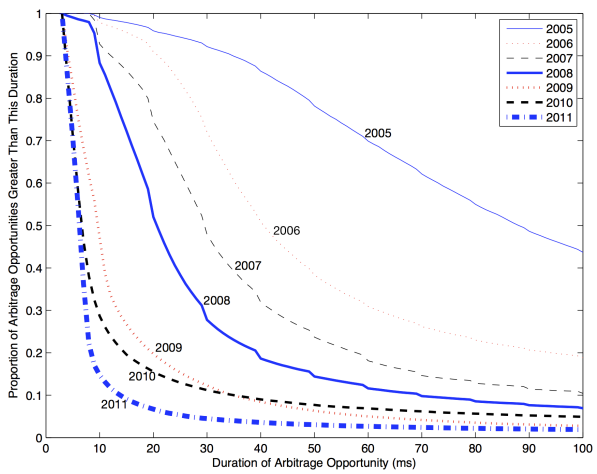
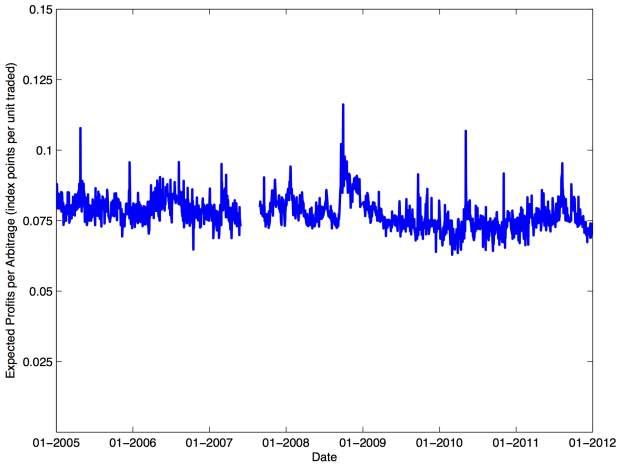


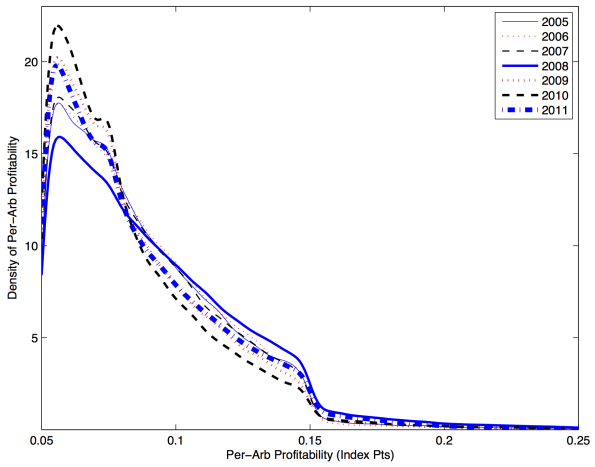
Figure 5.4: Profitability of ES & SPY Arbitrage Opportunities Over Time: 2005-2011

Notes: Panel (a) shows the median profitability of arbitrage opportunities (per unit traded) between the E-mini S&P 500 future (ES) and the SPDR S&P 500 ETF (SPY) from January 2005 to December 2011. Each point represents the median profitability per unit traded of that day's arbitrage opportunities. The discontinuity in the time series (5/30/2007-8/28/2007) arises from omitted data resulting from data issues acknowledged by the NYSE. Panel (b) plots the kernel density of per-arbitrage profits for each year in our dataset. See Section 5.2.1 for details regarding the ES-SPY arbitrage. See Section 4 for details regarding the data.

(a) Median Per-Arb Profits Over Time



(b) Distribution of Per-Arb Profits Over Time



the variation in the number of arbitrage opportunities per day: the R^2 of the regression of daily arbitrage frequency on daily distance traveled is 0.87.

Together, the results depicted in Figures 5.3, 5.4 and 5.5 suggest that the ES-SPY arbitrage opportunity should be thought of more as a mechanical "constant" of the CLOB market design than as a profit opportunity that is competed away over time. Competition has clearly reduced the amount of time that arbitrage opportunities last (Figure 5.3), but the size of arbitrage opportunities has remained remarkably constant (Figure 5.4), and the frequency of arbitrage opportunities seems to be driven mostly by market volatility (Figure 5.5). Figure 5.2 above, 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 5.3), but competition has not fixed the root issue that correlations break down at high enough frequency (the analogue of Figure 5.4).

These facts both inform and are explained by our model in Section 6.

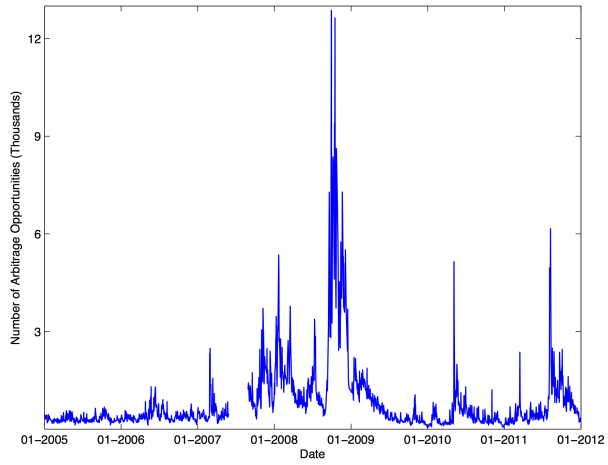
5.3 Discussion

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

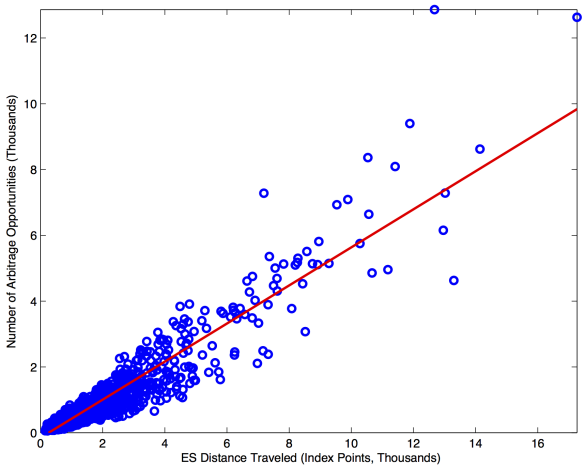
Figure 5.5: Frequency of ES & SPY Arbitrage Opportunities Over Time: 2005-2011

Notes: Panel (a) shows the time series of the total number of arbitrage opportunities between the E-mini S&P 500 future (ES) and the SPDR S&P 500 ETF (SPY), for each trading day in our data. The discontinuity in the time series (5/30/2007-8/28/2007) arises from omitted data resulting from data issues acknowledged by the NYSE. Panel (b) depicts a scatter plot of the total number of arbitrage opportunities in a trading day against that day's ES distance traveled. Distance traveled is 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. For more details on the trading strategy, see Section 5.2.1. See Section 5.2.1 for details regarding the ES-SPY arbitrage. See Section 4 for details regarding the data.

(a) Frequency of Arbitrage Opportunities



(b) Frequency vs. Distance Traveled



First, we suspect that our estimate of the annual value of the ES-SPY arbitrage opportunity– an average of around $75M per year, fluctuating as high as $151M in 2008 (the highest volatility year in our data) and as low as $35M in 2005 (the lowest volatility year in our data) – is an underestimate, for at least three reasons. One, 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. Two, 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 which economizes on transactions costs is to transact at market only in the security that lags – e.g., if ES jumps, transact at market in SPY but not in ES. Since 89% 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. Three, 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, due to the way that the CME's servers report 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 importantly, 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. One, there are hundreds of trades substantially similar to ES-SPY, consisting of securities that are highly correlated and with sufficient liquidity to yield meaningful profits from simple mechanical arbitrage strategies. Figure B.2 in the Appendix provides an illustrative partial list.[16] Two, 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 server farms in New Jersey – is measured in microseconds, which is finer than the current resolution of researcher-available exchange data. However, some

[16]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 $100M per security (calendar year 2011). It has not yet been possible to perform a similar screen on the universe of all securities, including, e.g., index futures, commodities, bonds, currencies, etc., due to data limitations. Instead, we include illustrative examples across all security types in Appendix Figure B.2.

indirect evidence for the importance and harmfulness of this type of arbitrage is that an entire new exchange, IEX, has recently been launched devoted to mitigating just this one aspect of the arms race (Patterson, 2013; IEX, 2014; Lewis, 2014). Three, securities that are meaningfully correlated, but with correlation far from one, can also be traded in a manner analogous to ES-SPY. For instance, even though the Goldman Sachs – Morgan Stanley correlation is far from one, 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 Appendix B.1.2, the 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. Four, 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, etc. 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). Last, in addition to the race to snipe stale quotes, there is also a race among liquidity providers to the top of the book (cf. Yao and Ye, 2014; Moallemi, 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 paper, 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.

6 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 securities 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 to make sense of these empirical facts. The model ultimately serves two related purposes: it is a critique of the CLOB market design, and it articulates the economics of the HFT arms race.

6.1 Preliminaries

Security x with perfect public signal y. There is a security x that trades on a CLOB, the rules of which are described in Section 3. 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 CLOB, abstracting from both asymmetric information and inventory costs.

We think of x and y as a metaphor for pairs or sets of securities that are highly correlated. For instance, x is SPY and y is ES. Alternatively, y can be interpreted more abstractly as simply 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 will 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 below in Section 6.2.4). 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 CLOB we derive below in Section 6.2, investors choose to transact immediately. In the equilibria of frequent batch auctions, studied in Section 7, 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 $N \geq 2$. Below, we will 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 amount weakly greater than the best outstanding ask (if buying) or an ask amount weakly lower than the best outstanding bid (if selling), whereas trading firms may use both marketable and non-marketable limit orders.[17]

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 CLOB. Trading firms observe innovations in the signal y with zero time delay, and there is zero latency in submitting orders to the exchange and receiving updates from the exchange. If multiple messages reach the CLOB 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 random latency, and then processed serially in the order received.[18]

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

In the exogenous entry case, we assume that investors observe y with any strictly positive latency; that is, they are slower than trading firms, but it is unimportant by how much. In the endogenous entry case we assume that investors observe y with the same latency as trading firms who do not invest in speed.

[17]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 non-standard. 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 more clear when we describe the role trading firms play in equilibrium below in Section 6.2.2.

[18]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 HFT's 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.

6.2 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.[19]

6.2.1 Investors

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

6.2.2 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 provider is stochastic, or rotates, etc.[20] 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.[21] We will 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.

[19] 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.

[20] 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.

[21] 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 $s = \epsilon$.

Specifically, at time t she submits a message to the exchange to cancel her previous quotes, of $y_{t^-} - \frac{s}{2}$ and $y_{t^-} + \frac{s}{2}$, and also submits a message to the exchange 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 his trading partner is an investor as opposed to another trading firm, though he 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.

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 to the CLOB 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 to the CLOB 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 CLOB processes message requests 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.[22]

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, i.e., if her order is not the first of the N to be processed by the CLOB, then she immediately withdraws

[22]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 – e.g., 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 fn. 18). 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.

her order.[23]

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.

6.2.3 Equilibrium Bid-Ask Spread s

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

$$\lambda_{invest} \cdot \frac{s}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N} \tag{6.1}$$

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

$$\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N} \tag{6.2}$$

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 (6.1) and (6.2) yields the equilibrium indifference condition:

$$\lambda_{invest} \cdot \frac{s}{2} = \lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \tag{6.3}$$

Equation (6.3) 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.

[23] 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".

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 while 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 (6.3) 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 in Sections 6.2.1-6.2.3. The structure of this equilibrium is unique in the following sense. In any equilibrium:*

1. *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 (6.3). 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.*

2. *Investors trade immediately when their demand arises.*

3. *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 his 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 $\frac{N-1}{N}$.*

4. *Trading firms are indifferent between liquidity provision and stale-quote sniping, for both the bid and the ask.*

5. *As per (6.3), 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 Appendix C for further details about this equilibrium, such as behavior off the equilibrium path, which complete the proof of Proposition 1.

6.2.4 Market Depth

Consider the model of Section 6.1 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 $q \in \{1, \ldots, \bar{q}\}$, with $p_k > 0$ the probability that they need to transact k units, for $k = 1, \ldots, \bar{q}$. Above, 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 CLOB 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 6.2, 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.[24] In equilibrium, the bid-ask spread for the k^{th} unit of liquidity, s_k, is governed by indifference between liquidity provision (LHS) and stale-quote sniping (RHS) at the k^{th} level of the book:

$$\lambda_{invest} \cdot \sum_{i=k}^{\bar{q}} p_i \cdot \frac{s_k}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s_k}{2}) \cdot \mathbb{E}(J - \frac{s_k}{2} | J > \frac{s_k}{2}) \cdot \frac{N-1}{N} \tag{6.4}$$

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

The LHS of (6.4) represents the benefits less costs of liquidity provision in the k^{th} level of the book. Notice that the second term on the LHS of (6.4), which describes the costs of getting sniped, is exactly the same as the second term of (6.1). 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 (6.4), which represents the benefits of sniping the k^{th} level of the book, is exactly the same as (6.2).

By contrast, except for the case of $k = 1$, the first term on the LHS of (6.4), which describes the benefits of providing liquidity, is strictly smaller than the first term of (6.1). This is because only proportion $\sum_{i=k}^{\bar{q}} p_i$ of investors trade the k^{th} level of the order book.

Intuitively, the benefits of providing liquidity scale sub-linearly 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

[24] As above, we can assign trading firms to roles in an arbitrary fashion. Since there are now $2\bar{q}$ limit orders present in the book at any given instant, there is plenty of room for several trading firms to split up the role of liquidity provider. Each such trading firm will try to snipe any stale quotes that are not his own.

offered.[25] 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 in Section 6.2.4. The structure of this equilibrium is unique in the following sense. In any equilibrium:*

1. *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 $k = 1, \ldots, \bar{q}$. The bid-ask spread s_k^* for the k^{th} unit of liquidity is uniquely characterized by* (6.4)*. These $2\bar{q}$ 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.*

2. *Spreads are strictly increasing,*
$$s_1^* < s_2^* < \cdots < s_{\bar{q}}^*$$
   (a) *Hence, investors' per-unit cost of trading is strictly increasing in order size.*

3. *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 $\frac{N-1}{N}$.*

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

5. *As per* (6.4)*, the following two quantities are equivalent in any equilibrium and do not depend on N:*

   - *The total rents to trading firms: $\sum_{k=1}^{\bar{q}} \lambda_{jump} \cdot \Pr(J > \frac{s_k^*}{2}) \cdot \mathbb{E}(J - \frac{s_k^*}{2} | J > \frac{s_k^*}{2})$. 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, $\lambda_{invest} \sum_{k=1}^{\bar{q}} \cdot \sum_{i=k}^{\bar{q}} p_i \cdot \frac{s_k^*}{2}$.*

[25] 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).

6.3 Discussion: Sniping is "Built In" to the CLOB

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 CLOB 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 such as ES-SPY 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, i.e., 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.[26]

[26]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

The mechanical reason that our source of costly liquidity provision does not arise in these prior works is a subtle difference in how the CLOB is modeled. Our model uses the actual rules of the CLOB (cf. Section (3)) 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); Goettler, Parlour and Rajan (2005), use abstractions of the CLOB 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 CLOB market design causes symmetrically observed information to be processed by the market as if it were asymmetrically observed information. As we will see below, discrete-time batching eliminates this built-in adverse selection and restores the possibility of meaningfully symmetric information.

6.4 Equilibrium with Endogenous Entry: the HFT Arms Race

The equilibrium analysis in Section 6.2 shows that the CLOB market design creates rents for trading firms, and that the sniping associated with these rents harms liquidity provision. In that analysis, we assumed away latency and treated the number of trading firms as exogenous. In this section we incorporate latency into the model and endogenize entry, by allowing trading firms to invest in a costly technology which increases the speed with which they can capture the rents created by the CLOB. This modification induces an arms race for speed. The arms race dissipates the rents created by the CLOB while doing nothing to fix the underlying liquidity problem associated with sniping.

6.4.1 Speed Technology

We model investment in speed in a simple way. Both investors and 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 described in the introduction, or the microwaves that replaced it), the cost of cutting-edge

financial market regulation that certain kinds of information – company news releases, government data announcements, order book activity – should be disseminated symmetrically. As we will see below, frequent batch auctions restore the possibility of meaningfully symmetric information, during trading hours and not just when the market is closed.

hardware, the cost of colocation facilities, the cost of the relevant human capital, etc. We assume that the decision of whether or not to pay this speed cost is taken at the start of the game and is observable and irreversible.[27] Define $\delta = \delta_{slow} - \delta_{fast}$, the speed difference between fast and slow trading firms.

Above, 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 (6.7) in footnote 28, which ensures that in equilibrium 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 integer. This allows us to characterize the equilibrium level of N using a zero-profit condition. Alternatively we could require that N is integer, in which case equilibrium N is characterized by weakly positive profits for trading firms with N entrants and strictly negative with $N + 1$.

6.4.2 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 6.2.4 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 6.2.2, with the one modification that they now each react to jumps in y with latency δ_{fast}. Investors behave exactly as above, buying or selling immediately upon 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 (6.1) equal the costs of speed:

$$\lambda_{invest} \cdot \frac{s}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N} = c_{speed} \tag{6.5}$$

Second is the zero-profit condition for stale-quote snipers, which says that the rents from

[27] 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.

sniping as written in (6.2) equal the costs of speed:

$$\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N} = c_{speed} \tag{6.6}$$

Together, equations (6.5) and (6.6) characterize the equilibrium bid-ask spread s^* and the equilibrium quantity of entry N^*. Notice that subtracting (6.6) from (6.5) yields exactly (6.3); 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 (6.5) and $N-1$ times (6.6) to obtain

$$\lambda_{invest} \cdot \frac{s^*}{2} = N^* \cdot c_{speed} \tag{6.7}$$

The economic interpretation of (6.7) 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 (6.3) 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.[28] Hence, the rents created by the CLOB are dissipated by the speed race.[29]

Proposition 3. *There is an equilibrium of the CLOB market design with endogenous entry with play as described in the text of Section 6.4.2. The equilibrium number of fast trading firms N^* and the equilibrium bid-ask spread s^* are uniquely determined by the zero-profit conditions* (6.5) *and* (6.6)*. 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:*

- *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, $N^* \cdot c_{speed}$.*

[28] 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.

[29] 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.

6.5 Discussion of the Equilibrium

6.5.1 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 6.4 modified so that the number of potential fast 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 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.

6.5.2 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:*

1. *The size of the prize is increasing in the frequency of jumps, λ_{jump}.*
2. *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}.*
3. *The size of the prize is invariant to the cost of speed, c_{speed}.*
4. *The size of the prize is invariant to the speed of fast trading firms, δ_{fast}.*
5. *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 CLOB market design – 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 5.2). And, arbitrage durations decline each year – as consistent with our findings on the duration of ES-SPY opportunities (Figure 5.3). However, the arms race prize itself is unaffected by these advances in speed – which is consistent with Figures 5.4 and 5.5 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 as consistent with our findings in the arbitrage analysis.

6.5.3 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 (cf. Fama, 1970).[30]

Proposition 6 (Market Efficiency in Time Space but not Volume Space)**.** *In equilibrium, the midpoint of the bid-ask spread for x is equal to fast trading firms' information about x's current fundamental value, $y_{t-\delta_{fast}}$ for proportion one of the trading day. Nevertheless, a strictly positive*

[30]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" (Committee, 2013). Our empirical and theoretical results show that, in fact, prices are extremely easy to predict in the *extremely* short run.

proportion of trade, $\frac{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*}}{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*} + \lambda_{invest}}$*, is conducted at quotes that do not contain* $y_{t-\delta_{fast}}$ *between the bid and the ask.*

That said, while the weak-form EMH is violated in our model, there still is no free lunch. Since the arbitrage profits induce costly entry, in equilibrium, fast traders' economic profits are zero.

6.5.4 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.[31]

This distinction between roles is important to bear in mind in interpreting the historical evidence on the effect of HFT on liquidity. The rise of HFT over the last fifteen 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 (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, pg. 103) and the cost of executing large trades over time (Angel, Harris and Spatt, 2013, pg. 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 90s and early-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.

[31] 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 terms exploratory trading, is detrimental to investors, and it is clearly distinct from stale-quote sniping.

7 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 CLOB market design.

7.1 Frequent Batch Auctions: Definition

Informally, frequent batch auctions are just like the CLOB but with two departures: (i) time is treated as discrete, not continuous; and (ii) orders are processed in batch, using an auction, instead of serially in order of receipt. The remainder of this subsection defines frequent batch auctions formally.[32]

The trading day is divided into equal-length discrete time intervals, each of length $\tau > 0$. We will refer to the parameter τ as the *batch length* and to the intervals as *batch intervals*. We 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 CLOB, 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.[33] 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 CLOB, traders may freely modify or cancel their orders at any moment in time. Also, just as in the CLOB, 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 (i.e., 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

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

[33] 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).

to buy with bids strictly greater than p^* and offers to sell with asks strictly less than p^* transact their full quantity, whereas for bids and asks of exactly p^* it may be necessary to ration one side of the market to enable market clearing.[34] For this rationing, we adopt a time-priority rule analogous to current practice under the CLOB 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 CLOB, 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 orders are economically processed by the exchange and only then are they announced publicly; the only difference is that the economic processing occurs in discrete time, and hence the information dissemination occurs in discrete time as well.[35]

To further clarify the relationship to the CLOB, it is helpful to describe three scenarios. The

[34] A 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 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 CLOB. This is an open question for future research.

[35] 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, leading to needless gaming. 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.

first scenario is when there is no new activity during the batch interval. In this case, all outstanding orders simply carry over to the next batch interval, analogous to displayed liquidity in a CLOB. The second scenario is when there is a small amount of new activity during the batch interval, e.g., an investor arrives and buys 100 shares at market. This scenario, too, is closely analogous to the CLOB. The investor will buy the 100 shares at the best outstanding ask price, and which ask gets processed is based on our version of time priority. The third scenario is when there is a large amount of new activity in the batch interval; e.g., there is a news event that relates to the security and many trading algorithms are reacting at once. This scenario is where frequent batch auctions and the CLOB are importantly different: frequent batch auctions process all of the new activity together in batch, at the end of the interval, whereas the CLOB processes the new activity serially in order of arrival.

7.2 Why and How Frequent Batch Auctions Address the Problems with Continuous Trading

The frequent batch auction market design directly addresses the problems we identified in Section 6 with the CLOB market design, 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 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 7.1 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 6.2, with N trading firms exogenously in the market all equally fast. Consider a trading firm providing liquidity. In the

Figure 7.1: Illustration of How Batching Reduces the Value of Tiny Speed Advantages

Notes: 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. It is only events that occur between $\tau - \delta_{slow}$ and $\tau - \delta_{fast}$ that 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 7.2.

CLOB, 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.[36] If y jumps, say from $\underline{y}$ to $\bar{y} > \underline{y}$, then the liquidity provider cancels his old quotes based on $\underline{y}$ and submits new quotes 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.[37]

In the second example, suppose as in the first example 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 CLOB, 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 our model,

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

[37] 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, e.g., 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 $\underline{y}$.

all orders reach the exchange at exactly the same time, and then the exchange processes them in a random order). 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 an auction allows 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. In the CLOB, competition drives trading firms to be ever so slightly faster than the competition, so that they can be first to snipe the stale quote. And, even if many fast traders observe a piece of information at literally the same time, the serial processing of the CLOB ensures that there is still a rent. In the batch auction market, by contrast, information that is widely available induces competition on price instead of speed, eliminating the rent. Of course, with batch intervals of, say, 100 milliseconds, there is still plenty of scope for market participants to develop genuinely asymmetric information about security values, for which they would earn a rent. But, batching eliminates rents from information that many market participants observe at basically the same time.

We summarize this discussion as follows:

Proposition 7 (Batching Eliminates Sniping and the Arms Race)**.** *Consider the frequent batch auction market design in the model of Section 6.4.*

1. *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 $\frac{\delta}{\tau}$.*

2. *The proportion of the trading day during which jumps in y leave a fast liquidity provider vulnerable to being sniped is* 0*.*

3. *If there are $N \geq 2$ fast traders exogenously in the market, and there is a slow liquidity provider with a vulnerable stale quote – i.e., 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:

1. *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.

2. *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.*

3. *If there are $N \geq 2$ fast traders present in the market, and there is a slow liquidity provider with a vulnerable stale quote – i.e., 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.*

7.3 Equilibrium of Frequent Batch Auctions

Section 7.2 described why frequent batch auctions eliminate sniping and the HFT arms race, by reducing the value of tiny speed advantages and by 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 6, we first consider the case of exogenous entry and then consider endogenous entry.

7.3.1 Model

We study the equilibria of frequent batch auctions using the model of Section 6.1 that we used to study the CLOB, with one modification. In the model of Section 6.1, investors arrive according to a Poisson process with arrival rate λ_{invest}. In the context of the CLOB, 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 $\bar{Q} < \infty$ such that the absolute value of $D(\tau)$ is bounded by $\bar{Q} - 1$. We view this assumption as innocuous so long as $\bar{Q}$ is large relative to the standard deviation of the Poisson arrival process, $\sqrt{\tau \lambda_{invest}}$.

7.3.2 Exogenous Entry

We begin by considering the setting of Section 6.2 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 7.2. 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, $\bar{Q}$, at zero bid-ask spread,[38] 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 8 (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:*

1. *The bid-ask spread for the first-quoted unit is narrower: it is 0 instead of the spread characterized by* (6.3)*.*

2. *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 6.2.4, the bid-ask spread grows wider with the quantity traded.*

This equilibrium highlights the central differences between frequent batch auctions and the CLOB. 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.

7.3.3 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}.

[38] We maintain the convention from Section 6 that it is possible to offer a zero bid-ask spread. Formally, fast trading firms can be interpreted as offering to buy at least $\bar{Q}$ units at price $y_\tau - \epsilon$ and sell at least $\bar{Q}$ units at price $y_\tau + \epsilon$, in the limit as $\epsilon \to 0_+$.

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 $\bar{Q}$ 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 $\bar{Q}$ 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 $\bar{Q}$ 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 $\bar{Q}$ units of depth in the limit order book, and there is, say, a positive jump, the entrant will wish to buy all $\bar{Q}$ 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 $\bar{Q}$ by the amount D, because the investors will outbid him for D of the $\bar{Q}$ units. On the other hand, if the imbalance D is negative, the fast trader can transact not just the $\bar{Q}$ 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 $\bar{Q}$ 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 $\bar{Q}$ 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 $\frac{p_{jump}}{\tau}\mathbb{E}(J') \cdot \bar{Q} \approx \frac{\delta}{\tau} \cdot \lambda_{jump}\mathbb{E}(J) \cdot \bar{Q}$. Note that a difference versus the analogous expression in (6.2) 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,

$$\frac{\delta}{\tau} \cdot \lambda_{jump} \cdot \mathbb{E}(J) \cdot \bar{Q} < c_{speed} \tag{7.1}$$

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 7.1), and these jumps occur at rate λ_{jump}. The LHS of (7.1) 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 $\bar{Q}$, equation (7.1) 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 $\bar{Q}$ consistent with (7.1) 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 9 (Equilibrium of Frequent Batch Auctions with Endogenous Entry)**.** *Fix any batch interval τ satisfying $\frac{p_{jump}}{\tau}\mathbb{E}(J') \cdot \bar{Q} < c_{speed}$, the exact version of* (7.1)*. 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:*

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

2. *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 6.2.4, the bid-ask spread grows wider with the quantity traded.*

3. *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 CLOB.*
   - *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 above in Proposition 4 can be understood as a socially wasteful prisoner's dilemma.[39] 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 paper, 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.[40]

[39]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.

[40]The working paper version of this paper considers the case where τ fails (7.1). 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

In Appendix C.2, we use a combination of our ES-SPY analysis and information from HFT public documents to calibrate equation (7.1). The goal of this exercise is not to determine the optimal batch interval, but rather to get an extremely rough 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 intrepretation of δ, when we plug in estimates for the other parameters in (7.1), we obtain a lower bound for τ on the order of 10 to 100 milliseconds. Under the second interpretation of δ we obtain a lower bound for τ on the order of 100ms to 1 second. Again, we caveat that the exercise is rough and at best gives a sense of magnitudes.

Appendix C.2 also 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 that in Section 7.3.3 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 to 10 milliseconds.

7.4 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 7.3.2, 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

vulnerable to sniping. Instead, liquidity is provided by a fast trading firm who pays c_{speed}, as in the equilibrium of the CLOB with endogenous entry. The key difference versus the CLOB is that the *fast* trading firm is no longer vulnerable to sniping, as per Proposition 7. As a result, the equilibrium bid-ask spread is narrower and depth is greater than in the CLOB with endogenous entry, though the spread is wider and the market is less deep than in the case of τ satisfying (7.1). Equilibrium expenditure on speed also lies between the CLOB with endogenous entry and frequent batch auctions with τ satisfying (7.1). In the limiting case of $\tau \to 0_+$ 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 (7.1) – and zero costs, because investor delay costs vanish as the delay goes to zero.

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 Section 7.3.3, 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 7.2 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 7.2 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 7.3.3, 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 (7.1). In a modification of the model in which information arrives in discrete time, as discussed in Appendix C.2.2, the batch interval should be long in *absolute* terms compared to the speed of slow traders, δ_{slow}. The calibration exercise in Appendix C.2.1, while extremely rough, suggests that a batch interval on the order of 10ms or 100ms 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. That is, while we have shown that there is a discontinuous benefit from moving from continuous time to discrete time, intuition suggests that the costs of moving from continuous time to discrete time are second order in the time interval.[41]

[41]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. The cost of this risk would be proportional to the square of the batch interval.

8 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.[42] While a formal analysis of the effect of frequent batch auctions on market stability is beyond the scope of the present paper, in this section we provide some informal discussion of several computational 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 the exchanges. Uniform-price auctions are fast to compute,[43] and exchange computers can be allocated a discrete block of time during which to perform this computation.[44] By contrast, in the CLOB market design, exchange computers are not allocated a block of time during which to perform order processing, but instead process orders and other messages in serial order of their 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 and the state of the limit order book. 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 (Strasburg and Bunge, 2013; Nanex, 2011; Jones, 2013).

A second computational benefit of frequent batching is that it gives algorithmic traders a discrete period of time to process recent prices and outcomes before deciding on their next trades.

[42] Duncan Niederauer, former CEO of NYSE Euronext, testified to Congress in June 2012 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 (Commission, 2013*a*), and policy papers by Haldane (2011) and Farmer and Skouras (2012).

[43] 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 10ms in this simple computational environment.

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

That is, algorithms can observe all of the relevant information from the time t batch auction, process it, and then decide on their actions in the time $t+1$ batch auction. By contrast, in the continuous-time market, trading algorithms are incentivized to race to react whenever any piece of relevant information is received. This means, first, that trading algorithms are incentivized to trade off "smarts" for speed, i.e., to make trading decisions based on only partial information and with only simple economic logic, since accumulating information and using more complicated economic logic each take time. And, second, that trading algorithms are incentivized to trade off code robustness and risk checking for speed, because error and risk checking each take time.[45] While frequent batching 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.

Third, frequent batch auctions improve the paper trail for regulators and other market observers. The regulatory authorities can observe exactly what happened at time t, at time $t+1$, etc. In a continuous-time market the paper trail can be much less clear, because the relationship between the time an order is submitted and the time it is processed by the relevant exchange is stochastic, due to backlog, because the sequence of time stamps across exchanges may not reflect the actual sequence of events, due to varying processing delays across exchanges, and because of the need to adjust timestamps to account for relativity. 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. Regulatory concerns regarding clock synchronization across exchanges (Hope, 2014) and the discrepancy in latency between direct feeds and the SIP feed (Ding, Hanna and Hendershott, 2014; see also Lewis, 2014) become moot in a discrete-time market.

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.

9 Conclusion

This paper argues that the continuous limit order book is a flawed market design and proposes that financial exchanges instead use discrete-time frequent batch auctions – uniform-price double auctions conducted, e.g., every tenth of a second. To recap, our basic argument is as follows.

[45] 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."

First, we show empirically that the CLOB market design does not really "work" in continuous time: market 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 CLOB 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 CLOB 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 in this paper is extremely stylized. This level of abstraction is appropriate both for making stark the key design flaw of the CLOB 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, multi-leg 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 8 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 non-issues 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.

References

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

Angel, James J., Lawrence E. Harris, and Chester S. Spatt. 2013. "Equity Trading in the 21st Century: An Update."

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

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

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

Biais, Bruno, Theirry Foucault, and Sophie Moinas. 2013. "Equilibrium Fast Trading." *IDEI Working Papers*.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2014*a*. "High Frequency Trading and Price Discovery." *Review of Financial Studies*, 27(8): 2267–2306.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2014*b*. "High Frequency Trading and the 2008 Short Sale Ban." *Working Paper*.

Budish, Eric, Peter Cramton, and John Shim. 2014. "Implementation Details for Frequent Batch Auctions: Slowing Down Markets to the Blink of an Eye." *American Economic Review: Papers and Proceedings*, 104(5): 418–424.

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

Cinnober. 2010. "Using Adaptive Micro Auctions to Provide Efficient Price Discovery When Access in Terms of Latency is Differentiated Among Market Participants."

Clark-Joseph, Adam. 2013. "Exploratory Trading." *Working Paper*.

Cohen, Kalman J., and Robert A. Schwartz. 1989. "The Challenge of Information Technology for the Securities Markets: Liquidity, Volatility and Global Trading." , ed. Henry Lucas and Robert Schwartz, Chapter An Electronic Call Market: Its Design and Desirability, 15–58. Dow Jones-Irwin.

Commission, Commodity Futures Trading. 2013*a*. "Concept Release on Risk Controls and System Safeguards for Automated Trading Environments."

Commission, European. 2013*b*. "FTT – Non-technical Answers to Some Questions on Core Features and Potential Effects."

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

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

Copeland, Thomas E., and Dan Galai. 1983. "Information Effects on the Bid-Ask Spread." *Journal of Finance*, 38(5): 1457–1469.

Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms.* . Third ed., MIT Press.

Demsetz, Harold. 1968. "The Cost of Transacting." *The Quarterly Journal of Economics*, 82(1): 33–53.

Ding, Shengwei, John Hanna, and Terrence Hendershott. 2014. "How Slow is the NBBO? A Comparison with Direct Exchange Feeds." *Financial Review*, 49(2): 313–332.

Duffie, Darrell, Nicolae Garleanu, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica*, 73(6): 1815–1847.

Economides, Nicholas, and Robert A. Schwartz. 1995. "Electronic Call Market Trading: Let Competition Increase Efficiency." *The Journal of Portfolio Management*, 21(3): 10–18.

Einstein, Albert. 1905. "Zur Elektrodynamik bewegter Körper (On the Electrodynamics of Moving Bodies)." *Annalen der Physik*, 322(10): 891–921.

Epps, Thomas. 1979. "Comovements in Stock Prices in the Very Short Run." *Journal of the American Statistical Association*, 74(366): 291–298.

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance*, 25(2): 383–417.

Farmer, J. Doyne, and Spyros Skouras. 2012. "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.

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

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

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham. 2014. "Toxic Arbitrage." *CEPR Discussion Papers: 9925*.

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

Glosten, Lawrence R. 1994. "Is the Electronic Open Limit Order Book Inevitable?" *Journal of Finance*, 49(4): 1127–1161.

Glosten, Lawrence R., and Paul Milgrom. 1985. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics*, 14(1): 71–100.

Goettler, Ronald L., Christine A. Parlour, and Uday Rajan. 2005. "Equilibrium in a Dynamic Limit Order Market." *The Journal of Finance*, 60(5): 2149–2192.

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

Harris, Jeffrey, and Paul Schultz. 1998. "The Trading Profits of SOES Bandits." *Journal of Financial Economics*, 39–62.

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

Harris, Larry. 2012. "Stop the High-Frequency Trader Arms Race." *Financial Times*, December 27.

Hasbrouck, Joel, and Gideon Saar. 2013. "Low-Latency Trading." *Journal of Financial Markets*, 16(4): 646–679.

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

Hirshleifer, Jack. 1971. "The Private and Social Value of Information and the Reward to Inventive Activity." *The American Economic Review*, 61(4): 561–574.

Hope, Bradley. 2014. "Clock Synchronization with Traders is Challenge for Regulators." *The Wall Street Journal.*

IEX. 2014. "Form ATS."

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

ISN. 2013. "Toward A Fairer and More Efficient Market." *ISN Research Report.*

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

KCG. 2013. "Form S-4 Registration Statement."

Klemperer, Paul. 2003. "Why Every Economist Should Learn Some Auction Theory." *Advances in Economics and Econometrics: Theory and Applications*, 1: 25–55.

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

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica*, 53(6): 1315–1335.

Laughlin, Gregory, Anthony Agiurre, and Joseph Grundfest. 2014. "Information Transmission between Financial Markets in Chicago and New York." *The Financial Review*, 49: 283–312.

Lewis, Michael. 2014. *Flash Boys: A Wall Street Revolt.* W.W. Norton & Co.

MacKenzie, Donald. 2014. "Be Grateful for Drizzle." *London Review of Books*, 36(17): 27–30.

Madhavan, Anath. 1992. "Trading Mechanisms in Securities Markets." *The Journal of Finance*, 47(2): 607–641.

McKay Brothers Microwave Latencies Table. 2015. `http://www.mckay-brothers.com/product-page/#latencies` Accessed Feb 3, 2015.

McPartland, John. 2013. "Recommendations for Equitable Allocation of Trades in High Frequency Trading Environments." *Federal Reserve Bank of Chicago Policy Discussion Paper.*

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

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

Milgrom, Paul. 2011. "Critical Issues in the Practice of Market Design." *Economic Inquiry*, 49(2): 311–320.

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

Najarian, Jon A. 2010. "The Ultimate Trading Weapon." *(blog)*, `http://moneytalks.net/pdfs/37895070-The-Ultimate-Trading-Weapon.pdf`.

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

Nanex. 2012. "Dangerous Order Types." November 15. `http://www.nanex.net/aqck2/3681.html`.

Niederauer, Duncan. 2012. "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. `http://financialservices.house.gov/uploadedfiles/112-137.pdf`.

O'Hara, Maureen. 2015. "High Frequency Market Microstructure." *Journal of Financial Economics.*

Patterson, Scott. 2013. "Upstart Pitches Trading Sanctum." *Wall Street Journal*, July 29. http://online.wsj.com/article/SB10001424127887324170004578634040178310664.html.

Patterson, Scott, and Jenny Strasburg. 2012. "How 'Hide Not Slide' Orders Work." *Wall Street Journal*, September 18. http://online.wsj.com/article/SB10000872396390444812704577605840263150860.html.

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

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

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

Roth, Alvin E. 2002. "The Economist as Engineer: Game Theory, Experimentation and Computation as Tools for Design Economics." *Econometrica*, 70(4): 1341–1378.

Roth, Alvin E. 2008. "What Have We Learned From Market Design?" *Economic Journal*, 118(527): 285–310.

Roth, Alvin E., and Axel Ockenfels. 2002. "Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet." *American Economic Review*, 92(4): 1093–1103.

Roth, Alvin E., and Xiaolin Xing. 1994. "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions." *American Economic Review*, 84(4): 992–1044.

Roth, Alvin E., and Xiaolin Xing. 1997. "Turnaround Time and Bottlenecks in Market Clearing: Decentralized Matching in the Market for Clinical Psychologists." *Journal of Political Economy*, 105(2): 284–329.

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

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

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

SEC, and CFTC. 2010. "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: 2012.

Securities and Exchange Commission. 2010. "Concept Release on Equity Market Structure."

Sparrow, Chris. 2012. "The Failure of Continuous Markets." *The Journal of Trading*, 7(2): 44–47.

Steiner, Christopher. 2010. "Wall Street's Speed War." *Forbes Magazine*, September.

Stiglitz, Joseph E. 2014. "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*.

Stoll, Hans R. 1978. "The Supply of Dealer Services in Securities Markets." *The Journal of Finance*, 33(4): 1133–1151.

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

Strasburg, Jenny, and Jacob Bunge. 2013. "Nasdaq Is Still on Hook as SEC Backs Payout for Facebook IPO." *Wall Street Journal*, March 25. `http://online.wsj.com/article/SB10001424127887323466204578382193806926064.html`.

Summers, Laurence H., and Victoria P. Summers. 1989. "When Financial Markets Work Too Well: A Cautious Case for a Securities Transactions Tax." *Journal of Financial Services Research*, 3(2-3): 261–286.

Tobin, James. 1978. "A Proposal for International Monetary Reform." *Eastern Economic Journal*, 4(3-4): 153–159.

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

Vayanos, Dimitri. 1999. "Strategic Trading and Welfare in a Dynamic Market." *Review of Economic Studies*, 66(2): 219–254.

Virtu. 2014. "Form S-1: Virtu Financial, Inc."

Wah, Elaine, and Michael Wellman. 2013. "Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model." *14th ACM Conference on Electronic Commerce*, June.

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

A 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 two recent private sector market design innovations, TMX's asymmetric delay to executable orders and IEX's combination of a symmetric delay and price-sliding logic for stale quotes.

A.1 Tobin Taxes

Tobin taxes (also known as financial transactions taxes) were originally proposed as "sand in the gears" to curb perceived excessive speculation and excessive volatility in foreign exchange markets (cf. 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,[46] 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 6.4. For expositional simplicity assume that the tax is paid by the liquidity-taking side of the trade (the analysis is economically identical if the two sides split the tax or the liquidity providing side pays the tax). This tax has two effects. The direct effect is that it simply increases the cost of trading by θ. The indirect effect, however, 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. 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.

Let $s(\theta)$ denote the bid-ask spread as a function of the level of the Tobin tax, and denote investors' total cost of trading by $\kappa(\theta)=\theta+\frac{s(\theta)}{2}$. We have $\frac{s(0)}{2} = \kappa(0) = \frac{s^*}{2}$, where s^* is the bid-ask spread from our original model as characterized by (6.5)-(6.6). We have the following results:[47]

Proposition 10 (Effect of Tobin Tax)**.** *In the model of the CLOB with endogenous entry as studied in Section 6.4, adding a Tobin tax of $\theta > 0$ per unit traded has the following effects:*

[46] 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 Commission (2013*b*).

[47] Proofs are in Appendix C.

1. *Investment in speed is lower. The reduction in investment in speed is increasing in the level of the tax: letting N_{fast} denote the equilibrium number of fast traders, we have $N'_{fast}(\theta) \leq 0$.*

2. *The bid-ask spread, i.e., the sniping-cost component of transactions costs, is lower. This reduction in sniping costs is increasing in the level of the tax: $s'(\theta) \leq 0$.*

3. *Investors' all-in trading costs are higher. This increase in investor trading costs is increasing in the level of the tax: $\kappa'(\theta) > 0$.*

Parts 1 and 2 of the proposition tell us that an appropriately set Tobin tax is an alternative way to mitigate sniping and the HFT arms race. However, Part 3 gives an important caveat, which is that Tobin taxes achieve these benefits at the expense of making investors worse off.[48] A second caveat is that the Tobin tax is a relatively blunt instrument: in order 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% would require a Tobin tax on the order of 10 basis points.[49] Reducing arms race profits by 99% would require a Tobin tax on the order of 50 basis points.

A.2 "Bans" on HFT: Message-to-Trade 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.[50] 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 500ms. 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.

[48] 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.

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

[50] The SEC's Concept Release on Equity Market Structure listed these as 2 of 5 common characteristics of HFT trading strategies, along with the use of speed technology, the use of co-location, and ending the trading day close to flat (Securities and Exchange Commission, 2010).

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 CLOB. 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 CLOB may lead to what they term "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, e.g., multiple jumps within the same 500ms, 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 one. Minimum resting times are in a sense the opposite of the asymmetric deterministic delays which we show are an attractive idea below in Section A.4, because they slow down attempts by liquidity providers to cancel rather than slowing down snipers.

A.3 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 firms responding to the same stimulus, which in turn reduces the incentive to invest in tiny speed improvements.

While intuitively appealing, there are two important concerns with random message delays.

First, random message delays do not eliminate 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, in the hopes that one of them will be processed with short random delay. Consider our model of Section 6.2, 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 1 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.

A full equilibrium analysis of the CLOB with random message delays is beyond the scope of this paper. The source of intractability is the possibility that multiple events – jumps in y, investor arrivals – occur within the random delay period. If we assume away the possibility of multiple events within the delay window we can get the following simple characterization of equilibrium.

Proposition 11 (Effect of Random Message Delays)**.** *Consider the model of the CLOB with endogenous entry as studied in Section 6.4. Incorporate a random message delay that is a uniform draw from* $[0, \epsilon]$*, for some* $\epsilon > 0$*. Assume that, after any jump in* y *or investor arrival, both Poisson processes pause for* 2ϵ *time. There is an equilibrium with the following features:*

1. *The number of fast trading firms,* N^**, and equilibrium expenditure on speed,* $N^* \cdot c_{speed}$*, is identical to that in Section 6.4.*

2. *The fast trading firms divide into one liquidity provider and* $N^* - 1$ *stale-quote snipers, just as in Section 6.4.*

3. *The bid-ask spread* s^* *is identical to that in Section 6.4.*

4. *After each sufficiently large jump in* y*, each of the fast trading firms sends infinitely many messages; formally, each fast trading firm sends* $\bar{M}$ *messages and we consider the limit as* $\bar{M} \to \infty$*. The liquidity provider is sniped with probability* $\frac{N^*-1}{N^*}$*, just as in Section 6.4.*

This proposition is stark: it suggests that the random message delay has no effect on liquidity or the arms race, its only effect is to increase message traffic by fast trading firms looking to circumvent the random delay. A natural idea in response to this proposition might be to place a cap on the number of redundant messages any one firm can send, for instance, a cap of 1 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.

A.4 Asymmetric Delay to Immediately Executable Orders

Our analysis in the previous section showed that random message delays do not address sniping and encourage redundant message traffic. Consider, however, the following alternative, which captures the key idea of recent market design innovations by TMX and IEX: apply an asymmetric

delay of $\Delta > 0$ only to immediately executable orders.[51,52] 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 6.4, 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 6, the asymmetric delay eliminates sniping and hence stops the arms race, just like frequent batch auctions. However, there are two disadvantages of the deterministic 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 CLOB 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.

First, the deterministic delay does not address the race to the top of the book; see Yao and Ye (2014); Moallemi (2014) for analyses of this component of the speed race. Formally, consider a modification to the model of Section 6 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 very common case in practice, as documented by Yao and Ye (2014) and others. In this case, in the CLOB, trading firms strictly prefer the role of liquidity provider to the role of stale-quote sniper (cf. fn 20). In equilibrium, after jumps in y, there can be 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 not only the race to snipe to stale quotes, but also the race to the top of the book. The advantage a fast trading firm has over a slow trading firm with respect to obtaining priority in the order book is proportional to $\frac{\delta}{\tau}$,[53] just as is the advantage a fast trading firm has over a slow trading firm with respect to

[51]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 non-executable 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 A.3 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.

[52]The key details of the IEX Alternative Trading System are as follows. 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 other US equity 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. See IEX Group (2014) for more details on the IEX market design.

[53]The fast trading firm obtains time priority in the book over the slow trading firm only if their order reaches

sniping stale quotes. By contrast, the deterministic delay has zero effect on the race to the top of the book.

Second, the deterministic delay does not transform competition on speed into competition on price in the event that there are unmonitored quotes in the book that become stale based on public information. Formally, consider a modification of our model in Section 6 in which a strictly positive proportion of investors attempt to satisfy their demand to trade with unmonitored non-marketable limit orders; e.g., an investor who arrives at time t, instead of transacting immediately at the ask $y_t + \frac{s^*}{2}$, submits a bid of $y_t - \frac{s^*}{2}$ which he then leaves unmonitored until it is filled or the trading day ends. This behavioral type captures the idea that some investors attempt to trade without paying the bid-ask spread even though their monitoring technology is sub-par. Suppose an unmonitored quote becomes stale based on an innovation in the public signal y, observed by all trading firms but not by the investor. In the CLOB with asymmetric delay, the unmonitored 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, by contrast, the unmonitored quote induces competition on price, and will get filled at a price determined by the batch auction based on the new public information rather than at the stale price. Hence, unmonitored quotes are a second source of incentive to race for speed in the CLOB with asymmetric delay that is eliminated by frequent batch auctions.

For both of these reasons, the asymmetric delay is only a partial response to the HFT arms race. While it ingeniously addresses sniping, it does not change the continuous-time serial-process nature of the market design, and for this reason does not fully eliminate the incentive to invest in speed. Further research on these new market designs would be a valuable direction for future research.

B Backup Materials for the Empirical Analysis

B.1 Correlation Breakdown Appendix

B.1.1 Alternative Measures of the ES-SPY Correlation

In this section, we describe alternative measures of the ES-SPY correlation as a robustness check for the results presented in section 5.1.1. These correlation measures vary along three dimensions. First, we consider both equal-weighted bid-ask midpoints and quantity-weighted bid-ask midpoints. Whereas equal-weighted midpoints place weight of $\frac{1}{2}$ on the bid and the ask, quantity-

the order book in an earlier batch interval. Hence, just as depicted in Figure 7.1, 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.

weighted midpoints place weight $\omega_t^{bid} = \frac{Q_t^{ask}}{Q_t^{ask}+Q_t^{bid}}$ on the bid and weight $\omega_t^{ask} = 1-\omega_t^{bid}$ on the ask, where Q_t^{bid} and Q_t^{ask} denote the quantity offered at the bid and ask at time t. Second, we consider correlation measures based on both simple returns and on average returns. Specifically, given a time interval τ and a time t, the simple return is the percentage change in price from time $t-\tau$ to time t, and the average return is the percentage change between the average price in the interval $(t-2\tau, t-\tau]$ and the average price in the interval $(t-\tau, t]$. Last, we consider three different ways to handle the concern that the speed-of-light travel time between New York and Chicago is roughly 4 milliseconds, which, per the theory of special relativity, represents a lower bound on the amount of time it takes information to travel between the two locations. One approach is to compute correlations based on the perspective of a trader in New York, which treats Chicago events as occurring 4ms later in New York than they do in Chicago. That is, the New York perspective treats Chicago events with time stamp t as contemporaneous with New York events with time stamp $t+4ms$. A second approach is to compute correlations based on the perspective of a trader in Chicago, in which case New York events with time stamp t are treated as contemporaneous with Chicago events with time stamp $t+4ms$. A last approach is to adjust neither dataset; this can be interpreted either as ignoring speed-of-light concerns or as taking the vantage point of a trader equidistant between Chicago and New York.

Table 2 displays the median ES-SPY correlation for varying time intervals over all trading days in 2011 for each of our 12 ($= 2\times 2\times 3$) methods of computing the correlation. As can be seen from the table, the pattern depicted in Figure 5.1, which uses our main specification of equal-weighted midpoints, simple returns, and no speed of light adjustment, is robust across all of the alternative specifications – at high enough frequency the ES-SPY correlation completely breaks down.[54]

B.1.2 Equities Correlation Breakdown

In this section, we compute correlations between several equity securities. This addresses two potential concerns. First, equity correlations address potential speed-of-light concerns. Since all of the equity securities analyzed below trade in the same physical location, the speed-of-light issue is not relevant for this exercise (at least not given the precision of our data). Second, the equity correlation results suggest that correlation breakdown applies more broadly to securities in general and not only to ES and SPY.

Table 3a displays the correlation at different time intervals between pairs of equity securities

[54] We also examined the correlogram of ES and SPY, for year 2011. The correlogram suggests that the correlation-maximizing offset of the two datasets treats Chicago events as occurring roughly 8-9 milliseconds earlier than New York events. At the correlation-maximizing offset, using simple returns and equal-weighted midpoints, the 1ms correlation is 0.0447, the 10ms correlation is 0.2232, and the 100ms correlation is 0.4863. Without any offset, the figures are 0.0080, 0.1016, and 0.4633.

Table 2: Correlation Breakdown in ES & SPY

Notes: This table shows the correlation between the return of the E-mini S&P 500 future (ES) and SPDR S&P 500 ETF (SPY) bid-ask midpoints as a function of the return time interval, reported as a median over all trading days in 2011. We compute correlations several different ways. First, we use either equal-weighted or quantity-weighted midpoints in computing returns. Quantity-weighted midpoints weight the bid and ask by $\omega_t^{bid} = Q_t^{ask} / \left(Q_t^{ask} + Q_t^{bid}\right)$ and $\omega_t^{ask} = 1 - \omega_t^{bid}$, respectively, where Q_t^{bid} and Q_t^{ask} represent the quantity offered at the bid and ask. Second, we use either simple or averaged returns. Simple returns use the conventional return formula and averaged returns use the return of the average midpoint of two non-overlapping intervals. Third, we compute correlations from the perspective of a trader in New York (Chicago events occurring at time t in Chicago are treated as contemporaneous with New York events occurring at time $t + 4ms$ in New York; labeled NY), a trader in Chicago (New York events occurring at time t in New York are treated as contemporaneous with Chicago events occurring at time $t + 4ms$ in Chicago; labeled Chi), and a trader equidistant from the two locations (labeled Mid). For more details on these correlation computations, See Section B.1.1. For more details on the data, refer to Section 4.

(a) Equal-Weighted Midpoint Correlations

Returns:	Simple			Average		
Location:	NY	Mid	Chi	NY	Mid	Chi
1 ms	0.0209	0.0080	0.0023	0.0209	0.0080	0.0023
10 ms	0.1819	0.1016	0.0441	0.2444	0.1642	0.0877
100 ms	0.4779	0.4633	0.4462	0.5427	0.5380	0.5319
1 sec	0.6913	0.6893	0.6868	0.7515	0.7512	0.7508
10 sec	0.9079	0.9076	0.9073	0.9553	0.9553	0.9553
1 min	0.9799	0.9798	0.9798	0.9953	0.9953	0.9953
10 min	0.9975	0.9975	0.9975	0.9997	0.9997	0.9997

(b) Quantity-Weighted Midpoint Correlations

Returns:	Simple			Average		
Location:	NY	Mid	Chi	NY	Mid	Chi
1 ms	0.0432	0.0211	0.0100	0.0432	0.0211	0.0100
10 ms	0.3888	0.2389	0.1314	0.5000	0.3627	0.2301
100 ms	0.7323	0.7166	0.6987	0.7822	0.7782	0.7717
1 sec	0.8680	0.8666	0.8647	0.8966	0.8968	0.8969
10 sec	0.9602	0.9601	0.9599	0.9768	0.9768	0.9769
1 min	0.9906	0.9906	0.9906	0.9965	0.9965	0.9965
10 min	0.9987	0.9987	0.9987	0.9998	0.9998	0.9998

that are relatively highly correlated, for instance, the oil companies Exxon-Mobil (XOM) and Chevron (CVX). Table 3b displays the correlation matrix for the 5 largest market capitalization US equities at a short and long time horizon. We follow the main specification used in Section 5.1, using equal-weighted midpoints and simple returns. As can be seen from the tables, the equities market correlation structure breaks down at high frequency. At human time scales such as one minute there is economically meaningful correlation between these securities, but not at high-frequency time scales such as 1ms or 100ms.

B.2 Mechanical Arbitrage Appendix

B.2.1 Computing the ES-SPY Arbitrage

In this section, we describe the mechanical relationship between ES and SPY that we use to estimate arbitrage frequency, duration and profitability.

Figure B.1 illustrates the exercise we conduct. The top portion depicts the midpoint prices of ES and SPY over the course of a fairly typical 30-minute period of trading (Panel a) and a volatile 30-minute period of trading during the financial crisis (Panel b). Notice that, while there is a difference in levels between the two securities, which is described in the main text (cf. Section 5.2.1), the two securities' price paths are highly correlated at this time resolution. The bottom portion depicts our estimate of the instantaneous profits (described below) associated with simultaneously buying one security (at its ask) and selling the other (at its bid). Most of the time these instantaneous profits are negative, reflecting the fact that buying one security while selling the other entails paying half the bid-ask spread in each market, constituting 0.175 index points in total. However, every so often the instantaneous profits associated with these trades turn positive. These are the moments where one security's price has just jumped a meaningful amount but the other security's price has not yet changed – which we know is common from the correlation breakdown analysis in Section 5.1. At such moments, buying the cheaper security and selling the more expensive security (with cheap and expensive defined relative to the difference in levels between the two securities) is sufficiently profitable to overcome bid-ask spread costs. Our exercise is to compute the frequency, duration, and profitability of such trading opportunities.

To begin, define the instantaneous spread between ES and SPY at millisecond t as

$$S_t^{mid} = P_{ES,t}^{mid} - 10 \cdot P_{SPY,t}^{mid}, \tag{B.1}$$

where $P_{j,t}^{mid}$ denotes the midpoint between the bid and ask at millisecond t for security $j \in \{ES, SPY\}$, and the 10 reflects the fact that SPY tracks $\frac{1}{10}$ the S&P 500 index. Next, define the

Table 3: Correlation Breakdown in Equities

Notes: This table shows the correlation between the returns of various equity pairs as a function of the return time interval, reported as a median over all trading days in 2011. Correlations are computed using equal-weighted midpoints and simple arithmetic returns. Speed-of-light considerations are not relevant for this exercise since all of these securities trade at the same geographic location. For more details on the data, refer to Section 4.

(a) Pairs of Related Companies

	1 ms	10 ms	100 ms	1 sec	10 sec	1min	10 min	30 min
HD-LOW	0.008	0.039	0.102	0.194	0.433	0.605	0.703	0.676
GS-MS	0.005	0.031	0.095	0.183	0.402	0.551	0.664	0.704
CVX-XOM	0.023	0.128	0.281	0.456	0.652	0.745	0.760	0.789
AAPL-GOOG	0.001	0.013	0.060	0.134	0.302	0.434	0.533	0.631

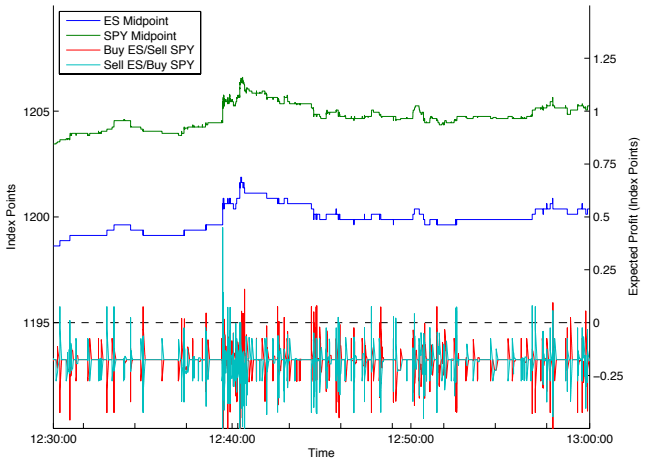
(b) Largest Components of the S&P 500 Index

	AAPL	XOM	GE	JNJ	IBM
			1 ms		
AAPL	1.000				
XOM	0.005	1.000			
GE	0.002	0.005	1.000		
JNJ	0.003	0.010	0.004	1.000	
IBM	0.002	0.005	0.002	0.004	1.000
			30 Min		
AAPL	1.000				
XOM	0.466	1.000			
GE	0.462	0.550	1.000		
JNJ	0.323	0.439	0.405	1.000	
IBM	0.498	0.513	0.513	0.446	1.000

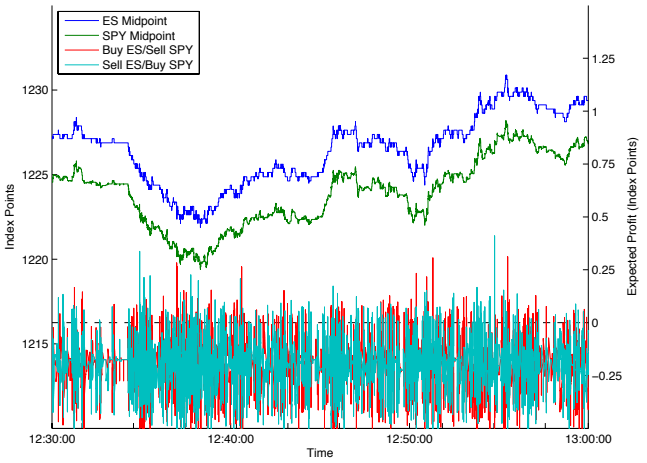
Figure B.1: Mechanical Arbitrage Illustrated

Notes: This figure illustrates the mechanical arbitrage between ES and SPY on an ordinary trading day (5/3/2010) in Panel (a) and a day during the financial crisis (9/22/2008) in Panel (b). In each panel, the top pair of lines depict the equal-weighted midpoint prices of ES and SPY, with SPY prices multiplied by 10 to reflect the fact that SPY tracks $\frac{1}{10}$ the S&P 500 index. The bottom pair of lines depict our estimate of the instantaneous profits associated with buying one security at its ask and selling the other security at its bid. These profits are measured in S&P 500 index points per unit transacted. For details regarding the data, see Section 4. For details regarding the computation of instantaneous arbitrage profits, see Section B.2.1.

(a) Ordinary Day



(b) Financial Crisis



moving-average spread between ES and SPY at millisecond t as

$$\bar{S}_t = \frac{1}{\tau^*} \sum_{i=t-\tau^*}^{t-1} S_i^{mid}, \tag{B.2}$$

where τ^* denotes the amount of time it takes, in milliseconds, for the ES-SPY averaged-return correlation to reach 0.99, in the trailing month up to the date of time t. The high correlation between ES and SPY at intervals of length τ^* implies that prices over this time horizon produce relatively stable spreads.[55] We define a trading rule based on the presumption that, at high-frequency time horizons, deviations of S_t^{mid} from $\bar{S}_t$ are driven mostly by the correlation breakdown phenomenon we documented in Section 5.1. For instance, if ES and SPY increase in price by the same amount, but ES's price increase occurs a few milliseconds before SPY's price increase, then the instantaneous spread will first increase (when the price of ES increases) and then decrease back to its initial level (when the price of SPY increases), while $\bar{S}_t$ will remain essentially unchanged.

We consider a deviation of S_t^{mid} from $\bar{S}_t$ as large enough to trigger an arbitrage opportunity if it results in the instantaneous spread market "crossing" the moving-average spread. Specifically, define the bid and ask in the instantaneous spread market according to $S_t^{bid} = P_{ES,t}^{bid} - 10 \cdot P_{SPY,t}^{ask}$ and $S_t^{ask} = P_{ES,t}^{ask} - 10 \cdot P_{SPY,t}^{bid}$. Note that $S_t^{bid} < S_t^{mid} < S_t^{ask}$ at all times t by the fact that the individual markets cannot be crossed, and that typically we will also have $S_t^{bid} < \bar{S}_t < S_t^{ask}$. If at some time t there is a large enough jump in the price of ES or SPY such that the instantaneous spread market crosses the moving-average spread, i.e., $\bar{S}_t < S_t^{bid}$ or $S_t^{ask} < \bar{S}_t$, then we say that an arbitrage opportunity has started at time t, which we now denote as t_{start}. We treat the relevant transactions cost of executing the arbitrage opportunity as the bid-ask spread costs associated with buying one security at its ask while selling the other at its bid. As discussed in the text in Section 5.2.1, this is a conservative and simple way to account for transactions costs. Expected profits, on a per-unit spread basis, are thus:

$$\pi \quad = \quad \begin{cases} \bar{S}_{t_{start}} - S_{t_{start}}^{ask} & \text{if } S_{t_{start}}^{ask} < \bar{S}_{t_{start}} \\ S_{t_{start}}^{bid} - \bar{S}_{t_{start}} & \text{if } S_{t_{start}}^{bid} > \bar{S}_{t_{start}}. \end{cases} \tag{B.3}$$

If our presumption is correct that the instantaneous market crossing the moving-average is due to correlation breakdown, then in the data the instantaneous market will uncross reasonably quickly, i.e., $S_t^{bid} < \bar{S}_t < S_t^{ask}$. We define the ending time of the arbitrage, t_{end}, as the first

[55] Economically, spreads are stable at such time horizons because the three differences between ES and SPY which drive the difference in levels – cost of carry until contract expiration, quarterly S&P 500 dividends, and ETF tracking error – are approximately stationary at time horizons on the order of seconds or a minute. Over longer time horizons, however, such as days or weeks, there is noticeable drift in the ES-SPY spread, mostly due to the way the cost of carry difference between the two securities changes as the ES contract approaches expiration.

millisecond after t_{start} in which the market uncrosses, the duration of the arbitrage as $t_{end} - t_{start}$, and label the opportunity a "good arb." If the expected profitability of the arbitrage varies over the time interval $[t_{start}, t_{end}]$, i.e., the instantaneous spread takes on multiple values before it uncrosses the moving average, then we record the full time-path of expected profits and quantities and compute the quantity-weighted average profits. This requires maintaining both the actual empirical order book and a hypothetical order book which accounts for our arbitrageur's trade activity. It is common that the trades in ES and SPY that our arbitrageur makes overlap with trades in ES and SPY that someone in the data makes, and we account for this in order to avoid double counting.[56]

In the event that the instantaneous market does not uncross the moving-average of the spread after a modest amount of time (we use τ^*) – e.g., what looked to us like a temporary arbitrage opportunity was actually a permanent change in expected dividends or short-term interest rates – then we declare the opportunity a "bad arb."

If an arbitrage opportunity lasts fewer than 4ms, the one-way speed-of-light travel time between New York and Chicago, it is not exploitable under any possible technological advances in speed. 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.[57]

B.2.2 Illustrative List of Highly Correlated Securities

Figure B.2 below provides an illustrative list of highly correlated securities, similar to ES-SPY. These pairs of securities are highly correlated and have sufficient liquidity to yield meaningful profits from simple mechanical arbitrage strategies.

[56] Here is an example to illustrate. Suppose that at time t_{start} an arbitrage opportunity starts which involves buying all 10,000 shares of SPY available in the NYSE order book at the ask price of p. Suppose that the next message in the NYSE data feed, at time $t' < t_{end}$, reports that there are 2,000 shares of SPY available at price p – either a trader with 8,000 shares offered at p just removed his ask, or another trader just purchased 8,000 shares at the ask. Our arbitrageur buys all 10,000 shares available at time t_{start}, but does not buy any additional shares at time t'. Even though the NYSE data feed reports that there are 2,000 shares of SPY at p at t', our hypothetical order book regards there as being 0 shares of SPY left at p at t'. If, on the other hand, the next message in the NYSE data feed at time t' had reported that there are 12,000 shares of SPY available at price p, then our arbitrageur would have purchased 10,000 shares at time t_{start}, and then an additional 2,000 shares at time t'.

[57] Prior to Nov 24, 2008, when the CME data was only at the centisecond level but the NYSE data was at the millisecond level, we filter out arbitrage opportunities that last fewer than 9ms, to account for the maximum combined effect of the rounding of the CME data to centisecond level (up to 5ms) and the speed-of-light travel time (4ms).

Figure B.2: Illustrative List of Highly Correlated Securities

E-mini S&P 500 Futures (ES) vs. SPDR S&P 500 ETF (SPY)
E-mini S&P 500 Futures (ES) vs. iShares S&P 500 ETF (IVV)
E-mini S&P 500 Futures (ES) vs. Vanguard S&P 500 ETF (VOO)
E-mini S&P 500 Futures (ES) vs. ProShares Ultra (2x) S&P 500 ETF (SSO)
E-mini S&P 500 Futures (ES) vs. ProShares UltraPro (3x) S&P 500 ETF (UPRO)
E-mini S&P 500 Futures (ES) vs. ProShares Short S&P 500 ETF (SH)
E-mini S&P 500 Futures (ES) vs. ProShares Ultra (2x) Short S&P 500 ETF (SDS)
E-mini S&P 500 Futures (ES) vs. ProShares UltraPro (3x) Short S&P 500 ETF (SPXU)
E-mini S&P 500 Futures (ES) vs. 9 Select Sector SPDR ETFs
E-mini S&P 500 Futures (ES) vs. E-mini Dow Futures (YM)
E-mini S&P 500 Futures (ES) vs. E-mini Nasdaq 100 Futures (NQ)
E-mini S&P 500 Futures (ES) vs. E-mini S&P MidCap 400 Futures (EMD)
E-mini S&P 500 Futures (ES) vs. Russell 2000 Index Mini Futures (TF)
E-mini Dow Futures (YM) vs. SPDR Dow Jones Industrial Average ETF (DIA)
E-mini Dow Futures (YM) vs. ProShares Ultra (2x) Dow 30 ETF (DDM)
E-mini Dow Futures (YM) vs. ProShares UltraPro (3x) Dow 30 ETF (UDOW)
E-mini Dow Futures (YM) vs. ProShares Short Dow 30 ETF (DOG)
E-mini Dow Futures (YM) vs. ProShares Ultra (2x) Short Dow 30 ETF (DXD)
E-mini Dow Futures (YM) vs. ProShares UltraPro (3x) Short Dow 30 ETF (SDOW)
E-mini Nasdaq 100 Futures (NQ) vs. ProShares QQQ Trust ETF (QQQ)
E-mini Nasdaq 100 Futures (NQ) vs. Technology Select Sector SPDR (XLK)
Russell 2000 Index Mini Futures (TF) vs. iShares Russell 2000 ETF (IWM)
Euro Stoxx 50 Futures (FESX) vs. Xetra DAX Futures (FDAX)
Euro Stoxx 50 Futures (FESX) vs. CAC 40 Futures (FCE)
Euro Stoxx 50 Futures (FESX) vs. iShares MSCI EAFE Index Fund (EFA)
Nikkei 225 Futures (NIY) vs. MSCI Japan Index Fund (EWJ)
Financial Sector SPDR (XLF) vs. Direxion Daily Financial Bull 3x (FAS)
Euro Futures (6E) vs. Spot EURUSD
Euro Futures (6E) vs. E-mini Euro Futures (E7)
Euro Futures (6E) vs. E-micro EUR/USD Futures (M6E)
E-mini Euro Futures (E7) vs. Spot EURUSD
E-mini Euro Futures (E7) vs. E-micro EUR/USD Futures (M6E)
E-micro EUR/USD Futures (M6E) vs. Spot EURUSD
Japanese Yen Futures (6J) vs. Spot USDJPY
Japanese Yen Futures (6J) vs. E-mini Japanese Yen Futures (J7)
E-mini Japanese Yen Futures (J7) vs. Spot USDJPY
British Pound Futures (6B) vs. Spot GBPUSD
Australian Dollar Futures (6B) vs. Spot AUDUSD
Swiss Franc Futures (6S) vs. Spot USDCHF
Canadian Dollar Futures (6C) vs. Spot USDCAD
New Zealand Dollar Futures (6N) vs. Spot NZDUSD
Mexican Peso Futures (6M) vs. Spot USDMXN
Gold Futures (GC) vs. miNY Gold Futures (QO)
Gold Futures (GC) vs. Spot Gold (XAUUSD)
Gold Futures (GC) vs. E-micro Gold Futures (MGC)
Gold Futures (GC) vs. SPDR Gold Trust (GLD)
Gold Futures (GC) vs. iShares Gold Trust (IAU)
miNY Gold Futures (QO) vs. E-micro Gold Futures (MGC)
miNY Gold Futures (QO) vs. Spot Gold (XAUUSD)
miNY Gold Futures (QO) vs. SPDR Gold Trust (GLD)
miNY Gold Futures (QO) vs. iShares Gold Trust (IAU)
E-micro Gold Futures (MGC) vs. SPDR Gold Trust (GLD)
E-micro Gold Futures (MGC) vs. iShares Gold Trust (IAU)
E-micro Gold Futures (MGC) vs. Spot Gold (XAUUSD)
Market Vectors Gold Miners (GDX) vs. Direxion Daily Gold Miners Bull 3x (NUGT)
Silver Futures (SI) vs. miNY Silver Futures (QI)
Silver Futures (SI) vs. iShares Silver Trust (SLV)
Silver Futures (SI) vs. Spot Silver (XAGUSD)
miNY Silver Futures (QI) vs. iShares Silver Trust (SLV)
miNY Silver Futures (QI) vs. Spot Silver (XAGUSD)
Platinum Futures (PL) vs. Spot Platinum (XPTUSD)
Palladium Futures (PA) vs. Spot Palladium (XPDUSD)
Eurodollar Futures Front Month (ED) vs. (12 back month contracts)
10 Yr Treasury Note Futures (ZN) vs. 5 Yr Treasury Note Futures (ZF)
10 Yr Treasury Note Futures (ZN) vs. 30 Yr Treasury Bond Futures (ZB)
10 Yr Treasury Note Futures (ZN) vs. 7-10 Yr Treasury Note
2 Yr Treasury Note Futures (ZT) vs. 1-2 Yr Treasury Note
2 Yr Treasury Note Futures (ZT) vs. iShares Barclays 1-3 Yr Treasury Fund (SHY)
5 Yr Treasury Note Futures (ZF) vs. 4-5 Yr Treasury Note
30 Yr Treasury Bond Futures (ZB) vs. iShares Barclays 20 Yr Treasury Fund (TLT)
30 Yr Treasury Bond Futures (ZB) vs. ProShares UltraShort 20 Yr Treasury Fund (TBT)
30 Yr Treasury Bond Futures (ZB) vs. ProShares Short 20 Year Treasury Fund (TBF)
30 Yr Treasury Bond Futures (ZB) vs. 15+ Yr Treasury Bond
Crude Oil Futures Front Month (CL) vs. (6 back month contracts)
Crude Oil Futures (CL) vs. ICE Brent Crude (B)
Crude Oil Futures (CL) vs. E-mini Crude Oil Futures (QM)
Crude Oil Futures (CL) vs. United States Oil Fund (USO)
Crude Oil Futures (CL) vs. ProShares Ultra DJ-UBS Crude Oil (UCO)
Crude Oil Futures (CL) vs. iPath S&P Crude Oil Index (OIL)
ICE Brent Crude Front Month (B) vs. (6 back month contracts)
ICE Brent Crude Front Month (B) vs. E-mini Crude Oil Futures (QM)
ICE Brent Crude (B) vs. United States Oil Fund (USO)
ICE Brent Crude (B) vs. ProShares Ultra DJ-UBS Crude Oil (UCO)
ICE Brent Crude (B) vs. iPath S&P Crude Oil Index (OIL)
E-mini Crude Oil Futures (QM) vs. United States Oil Fund (USO)
E-mini Crude Oil Futures (QM) vs. ProShares Ultra DJ-UBS Crude Oil (UCO)
E-mini Crude Oil Futures (QM) vs. iPath S&P Crude Oil Index (OIL)
Natural Gas (Henry Hub) Futures (NG) vs. United States Nat Gas Fund (UNG)

C Backup Materials for the Theoretical Analysis

C.1 Omitted Proofs

C.1.1 Proof of Proposition 1

To complete the argument that the behavior described in Sections 6.2.1-6.2.3 constitutes a static Nash equilibrium, and that this equilibrium is essentially unique as described in the proposition statement, we make the following observations.

First, investors are optimizing given trading firm behavior. Investors have no benefit to delaying trade, since the bid-ask spread s^* is stationary, y is a martingale, they are unable to successfully snipe since they have greater latency than trading firms, they are risk neutral, and their costs of delay are strictly increasing.

Second, let us confirm that the liquidity-provider's behavior is optimal given the behavior of investors and the stale-quote snipers. If at any moment in time the liquidity provider offers a bid less than $y_t - \frac{s^*}{2}$ or an ask greater than $y_t + \frac{s^*}{2}$, then one of the other trading firms will want to undercut her. For instance, if the liquidity provider sets an ask of $y_t + \frac{s''}{2}$ with $s'' > s^*$, then one of the other trading firms will immediately respond with an ask of $y_t + \frac{s'}{2}$ with $s'' > s' > s^*$. Our analysis in Section 6.2.3 which shows that providing liquidity at a bid-ask spread of s^* is exactly as attractive as sniping when the bid-ask spread is s^* implies that providing liquidity at $s' > s^*$ is strictly preferred to sniping. Hence, a deviation which widens the bid-ask spread (either on one or both sides) is not possible in equilibrium.[58] If the liquidity provider offers a narrower bid-ask spread, $s' < s^*$, then her profits are strictly lower than they are with a spread of s^*, so this is not an attractive deviation either. Third, if the liquidity provider offers a first unit of liquidity at s^* and then additional units of liquidity on either side of the book at a spread weakly greater than s^*, her benefits of providing liquidity stay the same (as it is, she satisfies all investor demand) but her costs of getting sniped will strictly increase, since she would get sniped for the full quantity. (The exception is if she offers additional liquidity at a spread so wide that it is never sniped; this is allowed for in the proposition statement). Last, if the liquidity provider offers zero units on either side of the book, then it is attractive for other trading firms to provide liquidity, and the reasoning above implies that they will do so at a bid of $y_t - \frac{s^*}{2}$ and/or an ask of $y_t + \frac{s^*}{2}$; this is just a permutation of roles, which is allowed for in the proposition statement.

[58] One might have expected that the liquidity provider will attempt to exploit an investor who happens to arrive to market in the interval between a change in the value of y and the time when this change is observable to investors. For instance, if y just jumped down in value, the liquidity provider might hope to sell to an investor at the old value of y (plus $\frac{s}{2}$). This discussion shows that this is not possible in equilibrium, because then other trading firms would no longer be indifferent between sniping and liquidity provision. They would prefer to offer more attractive quotes to investors.

Third, let us confirm that each stale-quote sniper's behavior is optimal given the behavior of the investors, the liquidity-provider, and the other stale-quote snipers. First, we note that in the event of a jump in y that is larger than $\frac{s^*}{2}$, it is clearly optimal for each stale-quote sniper to try to trade at the stale price; trying to do so has benefits and no costs. Next, we confirm that stale-quote snipers do not do anything else in equilibrium. Offering quotes narrower than the liquidity provider's quotes is not an attractive deviation, since such a deviation would yield negative profits per the analysis above. Offering quotes that are wider is not an attractive deviation, since such quotes have costs (of getting sniped) but no benefits. Last, offering quotes that are the same as the liquidity provider's is not an attractive deviation. More specifically, if the sniper's quotes reach the order book first (i.e., he wins the random tie-breaking against the liquidity provider's quotes) then he is simply playing the role of the liquidity provider (the original liquidity provider, off path, will remove his quotes and become a sniper), and our analysis in Section 6.2.3 shows that the two roles have equivalent payoffs. If the sniper's quotes reach the order book second, then such quotes derive less benefit than the quotes that are first – quotes that are second in time priority only get to transact if there are multiple investor arrivals before the next jump in y – but have the same sniping costs as the quotes that are first in time priority. So, this is not a profitable deviation either.

Last, to complete the proof of the proposition statement we confirm the uniqueness claims. Claim (1) is implied by the discussion of trading firm behavior above; note that, if $\bar{J}$ is the maximum jump size, a bid less than $y_t - \bar{J}$ and an ask greater than $y_t + \bar{J}$ can be placed in the book with zero benefit, because such orders trade with probability zero (because at almost all times there is a bid of $y_t - \frac{s^*}{2}$ and an ask of $y_t + \frac{s^*}{2}$ in the book) and zero cost, because such quotes are too wide to be vulnerable to sniping. Claim (2) is confirmed by the discussion of investor behavior above. Claim (3) is confirmed by the discussion of trading firm behavior above. Claim (4) follows from (6.1)-(6.3), which describe any equilibrium per the discussion above, and Claim (5) follows from (6.3). Note that what is not unique in equilibrium is the assignment of trading firms to roles. In particular, all that is pinned down is that at almost any moment in time t there is one trading firm providing liquidity at bid $y_t - \frac{s^*}{2}$ and one (possibly the same) trading firm providing liquidity at ask $y_t + \frac{s^*}{2}$, and that, in the event of a jump larger than $\frac{s^*}{2}$, the trading firm whose quote is stale attempts to cancel and the other $N - 1$ trading firms attempt to snipe.

C.1.2 Proof of Proposition 2

Equation (6.4) represents indifference between liquidity provision and stale quote sniping at the kth level of the book, for $k = 1, \ldots, \bar{q}$. It can be rearranged to obtain the multi-unit analogue of (6.3), which characterizes the equilibrium bid-ask spread at each level of the book:

$$\lambda_{invest} \cdot \sum_{i=k}^{\bar{q}} p_i \cdot \frac{s_k}{2} = \lambda_{jump} \cdot \Pr(J > \frac{s_k}{2}) \cdot \mathbb{E}(J - \frac{s_k}{2} | J > \frac{s_k}{2}) \tag{C.1}$$

For each k, the solution to (C.1) is unique because the LHS is strictly increasing in s_k (and is equal to zero at $s_k = 0$) whereas the RHS is strictly positive for $s_k = 0$ and then is strictly decreasing in s_k until it reaches its minimum of zero at s_k equal to the upper bound of the jump size distribution. The fact that $s_1^* < s_2^* < \cdots < s_{\bar{q}}^*$ follows from (C.1), because the probability that an investor wants to trade k units, $\sum_{i=k}^{\bar{q}} p_i$, is strictly decreasing in k. With these observations, the proposition follows from arguments analogous to those for Proposition 1.

Claim (1) follows from the same argument as for Claim (1) in Proposition 1, applied separately to each level of the book. At level k, widening the spread to $s'' > s_k^*$ induces another trading firm to undercut to $s'' > s' > s_k^*$; narrowing the spread is not profitable; and similarly waiting in queue at exactly s_k^* is not profitable.

Claim (2) follows from the observation above that spreads are strictly increasing with quantity. Note that for the multi-unit demand case we assumed that investor behavior is mechanical whereas for Proposition 1 investor behavior was microfounded.

Claims (3)-(5) follow from the same arguments as for Claim (3)-(5) in Proposition 1.

Note, again, that the assignment of trading firms to roles is not unique. All that is pinned down is that at almost any time t there are $2\bar{q}$ quotes in the book, as characterized by (C.1), which can belong to any number of trading firms. And, after a jump, for each quote that is stale, the 1 firm whose quote it is tries to cancel and the $N - 1$ other firms try to snipe. Note too that after a large jump some firms may be engaged in both cancelation of their own stale quotes and attempting to snipe others' stale quotes.

C.1.3 Proof of Proposition 3

First, note that investors are behaving optimally given trading firm behavior. The argument that investors should trade immediately is identical to that in the proof of Proposition 1: the bid-ask spread s^* is stationary, y is a martingale, they are unable to successfully snipe, they are risk neutral, and their costs of delay are strictly increasing.

Second, given any number of fast trading firms $N' \geq 2$, under the hypothesis (to be confirmed below) that slow trading firms do not play any role in equilibrium, the proof of Proposition 1 carries over exactly. In particular, the bid-ask spread is uniquely characterized by (6.3), and the N' fast trading firms endogenously sort into 1 liquidity provider and $N' - 1$ stale-quote snipers.

Third, we show that there is no opportunity for entry by a slow trading firm for any number $N' \geq 2$ of fast trading firms. Clearly, slow trading firms will never succeed at sniping stale quotes

when there are fast trading firms present. So we need to rule out the possibility of a slow trading firm providing liquidity. Suppose a slow trading firm provides liquidity at a spread of $s' < s^*$, i.e., a slow trading firm attempts to undercut the spread of the fast trading firm providing liquidity. The benefits to the slow trading firm from investor arrivals, per unit time, are $\lambda_{invest} \cdot \frac{s'}{2}$. The costs from getting sniped, per unit time, are $\lambda_{jump} \cdot \Pr(J > \frac{s'}{2}) \cdot \mathbb{E}(J - \frac{s'}{2} | J > \frac{s'}{2})$. Notice that whereas the fast trading firm is sniped with probability $\frac{N'-1}{N'}$, the slow trading firm is sniped with probability one.[59] Since $s^* > s'$, we have $\lambda_{invest} \cdot \frac{s'}{2} < \lambda_{invest} \cdot \frac{s^*}{2}$ and $\lambda_{jump} \cdot \Pr(J > \frac{s'}{2}) \cdot \mathbb{E}(J - \frac{s'}{2} | J > \frac{s'}{2}) > \lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$. But, (6.3) shows that $\lambda_{invest} \cdot \frac{s^*}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2}) = 0$ in equilibrium, so the slow trading firm's benefits less costs of liquidity provision, $\lambda_{invest} \cdot \frac{s'}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s'}{2}) \cdot \mathbb{E}(J - \frac{s'}{2} | J > \frac{s'}{2})$, are strictly negative. This shows that slow trading firms cannot profitably enter as liquidity providers with a bid-ask spread $s' < s^*$. It is obvious that they cannot profitably enter as liquidity providers with a bid-ask spread of $s' \geq s^*$ either. Hence, slow trading firms cannot profitably enter as liquidity providers.

Last, we show that the equilibrium quantity of entry by fast trading firms is N^*. Clearly, entry of $N' > N^*$ fast trading firms is not an equilibrium, since all such firms would lose money once speed costs are accounted for. Similarly, entry of $N' < N^*$ fast trading firms is not an equilibrium, since a marginal entrant could enter profitably. Hence, in any equilibrium, the number of fast trading firms is N^*.[60]

Hence, the behavior described in the text of Section 6.4.2 constitutes an equilibrium (sub-game perfect Nash at the entry stage, static Nash throughout the trading day, per fn. 19), as claimed, and the equilibrium number of fast trading firms N^* and the equilibrium bid-ask spread s^* are uniquely characerized by the zero-profit conditions (6.5)-(6.6), as claimed. The final statement of the proposition follows directly from (6.3) and (6.7).

C.1.4 Proof of Proposition 4

Formally, there are N^* trading firms, each of whom must choose the action *fast* or *slow*. If all N^* trading firms choose slow, they each earn profits of c_{speed}, as per Section 6.2. If all N^* trading

[59] This expression for the cost of getting sniped assumes that the slow trading firm avoids being sniped multiple times on the same jump in y. Formally, the exercise is to consider a slow trading firm who provides liquidity at spread s' until she trades a single time, then exits.

[60] Note that we assumed in Section 6.4 that N can take on any real value, i.e., we allowed for fractional entry. Mathematically, if N^* is non-integer, then a single fractional entrant pays cost pc_{speed}, with $p = N^* - \lfloor N^* \rfloor$ denoting the fraction with which he enters, and, when there is a stale quote, the fractional entrant's request to snipe is submitted with probability p, i.e., he transacts with total probability $\frac{p}{N^*}$. The game form allows the large fringe of slow trading firms to enter at any fractional rate between 0 and 1. Our argument shows that in any equilibrium, the total quantity of entry is exactly N^*. Alternatively, we could restrict attention to integer entry, in which case the quantity of entry would be $\lfloor N^* \rfloor$ in any equilibrium; all of the above analysis would carry through essentially unchanged. If an $\lfloor N^* \rfloor + 1^{st}$ trading firm entered it would lose money.

firms choose fast, they each earn profits of zero, as per Section 6.4 and equations (6.3) and (6.7). To show that *fast* is a dominant strategy, we make the following observations. If the number of trading firms who choose fast satisfies $1 < N < N^*$, then in equilibrium of the subgame the N fast trading firms play exactly as in Section 6.2, because indifference among the fast trading firms between liquidity provision and stale-quote sniping is still characterized by equation (6.3). The only difference is that each fast trading firm earns larger profits than when all N^* enter, since they split the revenues from investors of $\lambda_{invest}\frac{s^*}{2}$ among N instead of splitting it among N^*. If the number of trading firms who choose fast is 1, then one strategy available to the fast trading firm is to charge the same bid-ask spread s^* as when there are multiple trading firms, but to do so without any risk of being sniped; also, by the same argument as in the third step of the proof of Proposition 3, a slow trading firm cannot profitably undercut a bid-ask spread of s^*. The profits from this strategy are larger than the profits from the case where all trading firms are slow. Hence, for any number of fast trading firms $0 \leq N < N^*$, any slow trading firm prefers to be fast than slow. Hence, fast is a dominant strategy, and we have a prisoner's dilemma.

C.1.5 Proof of Proposition 5

Claim 1 follows from (6.3). Increasing λ_{jump} while holding λ_{invest} and the distribution of J constant, the only variable that can change in response is s. Since $\Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2})$ on the RHS is decreasing in s, and $\frac{s}{2}$ on the LHS is of course strictly increasing in s, an increase in λ_{jump} must increase s (else, the LHS would be weakly lower while the RHS would be strictly higher). If s is higher, then so too is the size of the prize which, per (6.3), is equal to $\lambda_{invest} \cdot \frac{s}{2}$.

Claim 2 follows similarly from (6.3). A mean-preserving spread of F_{jump} increases $\Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2})$ for any fixed level of s. This in response must increase the equilibrium s (else, the LHS would be weakly lower while the RHS is strictly higher), and if s is higher then so too is the size of the prize.

Claims 3-5 follow from the observation that none of the elements of (6.3), which per (6.5)-(6.6) characterizes s^* in the equilibrium with endogenous entry as well, are affected by c_{speed}, δ_{fast}, or δ. Note that (6.7) then implies that $N^* \cdot c_{speed}$ is a constant, meaning that if c_{speed} is higher N^* is lower, and vice versa; but, the total size of the prize is unaffected by c_{speed}.

C.1.6 Proof of Proposition 6

The observation that the midpoint of the bid-ask spread is equal to fast trading firms' information about fundamental value, $y_{t-\delta_{fast}}$, for proportion one of the trading day follows from the equilibrium behavior of the liquidity provider as described in Section 6.2.2.

The proportion of trade conducted at quotes that do not contain the fundamental value is computed by observing that the rate at which trade occurs between the liquidity provider and a sniper is $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*}$, whereas the rate at which trade occurs between the liquidity provider and an investor is λ_{invest}. In equilibrium, the former trades occur at quotes that are stale, i.e., where the quotes do not contain the fundamental value y_t which has just jumped, whereas the latter trades occur at quotes that are not stale (but for the probability zero event that an investor arrival and a jump occur at the exact same time). Hence, trade at stale quotes as a proportion of all trade is $\frac{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*}}{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*} + \lambda_{invest}}$.

C.1.7 Proof of Proposition 7

The three claims for frequent batch auctions are established in the text of Section 7.2. The three claims for CLOB markets follow from the description of equilibrium in Section 6.

C.1.8 Proof of Proposition 8

First, observe that it is a static Nash equilibrium for each of the N trading firms to offer depth of $\frac{\bar{Q}}{N-1}$ at zero bid-ask spread (i.e., to set bid and ask of y_τ for $\frac{\bar{Q}}{N-1}$ units), and for investors to trade at market in the batch auction immediately following their arrival. Investors clearly cannot do better since they fulfill their demand at zero cost as soon as is possible. Trading firms earn zero profits in this equilibrium. However, given the behavior of the other trading firms, each individual trading firm can do no better than to offer $\frac{\bar{Q}}{N-1}$ at zero bid-ask spread; if it sets a strictly positive spread for any unit the probability that that unit trades is zero, given the assumption that $\bar{Q}$ is an upper bound on the imbalance of investor demand in a batch interval. Similarly, it is not possible for a trading firm to snipe, since all trading firms are equally fast (cf. Proposition 7). Hence, there is no deviation that earns strictly positive profits.

Second, we show that this equilibrium is essentially unique. Suppose that there is an equilibrium in which trading firms earn strictly positive profits in the auction ending at τ. This means that there exists an investor imbalance quantity q (without loss consider $q > 0$) such that (i) the probability that the imbalance is q is strictly positive, and (ii) if the imbalance is q the market-clearing price is strictly greater than y_τ. Since imbalance is bounded, there exists a largest such imbalance satisfying (i) and (ii); call it $\hat{q}$, and call the price that results under this imbalance $\hat{p} > y_\tau$. Let $\hat{a}$ denote the quantity of asks in the book at price $\hat{p}$. Since $N \geq 2$ there exists at least one firm whose own quantity of asks at price $\hat{p}$ is strictly less than $\hat{a}$. This firm has a profitable deviation: keep its supply function unchanged except offer $\hat{a}$ units at a price of $\hat{p} - \epsilon$ for $\epsilon > 0$ and sufficiently small. Under this deviation the firm earns strictly higher profits if the imbalance is $\hat{q}$

(or any other imbalance that results in a market-clearing price of $\hat{p}$) which occurs with positive probability, and the same profits otherwise. A contradiction.[61]

C.1.9 Proof of Proposition 9

First, we established in the text that it is not profitable to enter as a fast trading firm given the hypothesized equilibrium behavior of slow trading firms. Picking off stale quotes is not sufficiently profitable, as shown by (7.1) and the surrounding discussion. Additionally, it is not profitable to enter as a fast trading firm in an effort to provide liquidity, because slow trading firms are already providing the maximum necessary liquidity at zero bid-ask spread. One last thing to point out is that the discussion in the text already covers the possibility of providing liquidity in the event that there is a jump between times $\tau - \delta_{slow}$ and $\tau - \delta_{fast}$; the fast trading firm's activity in such an event both exploits the stale quotes of the slow trading firms and provides liquidity to the net demand of investors, yielding $\bar{Q}$ of total volume in expectation. As discussed, this is not sufficiently profitable to induce the fast trader to enter.

Second, we describe equilibrium behavior by individual slow trading firms that generates the aggregate behavior described in the proposition. Of the large competitive fringe of slow trading firms, $\bar{Q}$ slow trading firms each offer a bid and an ask in each batch auction for a single unit at price $y_{\tau-\delta_{slow}}$, where $y_{\tau-\delta_{slow}}$ represents the best available information for a slow trading firm in the auction ending at time τ. Given that the other trading firms behave this way, each individual slow trading firm has no incentive to deviate; in particular, since the upper bound of investor imbalance is $\bar{Q}-1$, any individual slow trading firm that raises their price trades with probability zero.

Last, we show that this equilibrium is essentially unique. We do this by ruling out four cases. First, suppose that there is an equilibrium in which there are only slow trading firms but in which some of the slow trading firms earn strictly positive profits. The argument in the second paragraph of the proof of Proposition 8 applies to show that this generates a contradiction; there will be a profitable opportunity to undercut. Second, suppose that there is an equilibrium in which two or more fast trading firms enter. The proof of Proposition 8 applies to show that if multiple fast trading firms enter they will compete spreads to zero, so this is not a profitable deviation. Third, suppose that there is an equilibrium in which a single fast trading firm enters, alongside the fringe of slow trading firms, and the fast trading firm provides all liquidity to investors. Consider a hypothetical equilibrium liquidity supply schedule $s_1 \leq s_2 \leq \cdots \leq s_{\bar{Q}}$, where s_k denotes the

[61] Note that this argument is essentially the same argument used to establish the uniqueness of marginal-cost pricing in symmetric Bertrand price competition with bounded demand. As Klemperer (2003) and others have pointed out, there can also exist approximate equilibria of these games with mixed-strategies that yield prices in excess of marginal cost.

difference between the fast trading firm's kth best ask and kth best bid. Let $Rev(k)$ denote the hypothetical equilibrium revenue for the kth unit of liquidity provision, per unit time.[62] Let $Cost(k)$ denote the hypothetical cost to a slow trading firm, per unit time, of getting sniped by the fast trading firm if she provided the kth unit of liquidity at spread s_k instead of the fast trading firm.[63] There must exist a k for which $Rev(k) > \frac{c_{speed}}{\bar{Q}}$; otherwise, the total price of liquidity for investors could not exceed c_{speed} per unit time, which would mean that the fast trading firm cannot recover her costs. However, we know from (the exact version of) (7.1) that the sniping cost per unit of liquidity is strictly less than $\frac{c_{speed}}{\bar{Q}}$ per unit time even if the spread is zero, hence for all k we have $\frac{c_{speed}}{\bar{Q}} > Cost(k)$. Hence, there exists a k for which $Rev(k) > \frac{c_{speed}}{\bar{Q}} > Cost(k)$, i.e., there exists a unit of liquidity for which the liquidity-provision revenues strictly exceed the sniping costs. Hence, a slow trading firm can profitably undercut the fast trading firm, a contradiction. Fourth, suppose that there is an equilibrium in which a single fast trading firm enters, alongside the fringe of slow trading firms, and liquidity is provided by both the fast trading firm and slow trading firms. Since sniping is zero sum, for it to be the case that neither the fast trading firm wishes to undercut a slow trading firm providing liquidity nor that a slow trading firm wishes to undercut either the fast trading firm or another slow trading firm, we must have $Rev(k) = Cost(k)$ for all $k = 1, \ldots, \bar{Q}$. But, per the argument for the previous case, for the fast trading firm to recover her costs her profits per unit time must exceed $\frac{c_{speed}}{\bar{Q}}$ for at least one level of the book. Hence, for some k we have $Cost(k) = Rev(k) > \frac{c_{speed}}{\bar{Q}}$, which generates a contradiction since, as described in the previous case, (7.1) implies that $\frac{c_{speed}}{\bar{Q}} > Cost(k)$ for all k.

C.1.10 Proof of Proposition 10

Equilibrium with the Tobin tax is still governed by zero-profit conditions for liquidity provision and stale-quote sniping, as in Proposition 3, which also encode indifference between the two roles. The zero-profit conditions are now

$$\lambda_{invest} \cdot \frac{s(\theta)}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s(\theta)}{2} + \theta) \cdot \mathbb{E}(J - \frac{s(\theta)}{2} | J > \frac{s(\theta)}{2} + \theta) \cdot \frac{N(\theta) - 1}{N(\theta)} = c_{speed} \quad \text{(C.2)}$$

for the liquidity provider and

$$\lambda_{jump} \cdot \Pr(J > \frac{s(\theta)}{2} + \theta) \cdot \mathbb{E}(J - \frac{s(\theta)}{2} - \theta | J > \frac{s(\theta)}{2} + \theta) \cdot \frac{1}{N(\theta)} = c_{speed} \quad \text{(C.3)}$$

[62] Formally, the definition is $Rev(k) = \frac{1}{\tau} \cdot \Pr(D(\tau) \geq k) \cdot \sum_{j=k}^{\bar{Q}} \frac{\Pr(D(\tau)=j)}{\Pr(D(\tau)\geq k)} \cdot \frac{s_j}{2}$.

[63] The formal definition is $Cost(k) = \frac{p_{jump}}{\tau} \Pr(J' > \frac{s_k}{2}) \cdot \mathbb{E}(J' - \frac{s_k}{2} | J' > \frac{s_k}{2}) \cdot$. If we use the same approximations as in (7.1), the definition becomes $Cost(k) = \frac{\delta}{\tau} \lambda_{jump} \Pr(J > \frac{s_k}{2}) \cdot \mathbb{E}(J - \frac{s_k}{2} | J > \frac{s_k}{2}) \cdot$

for the stale-quote snipers. Notice that, in the event of a jump larger than $\frac{s(\theta)}{2} + \theta$ which results in a successful stale-quote snipe, the Tobin tax θ is paid by the sniper, so her net profits are $J - \frac{s(\theta)}{2} - \theta$, whereas the liquidity provider's losses are $J - \frac{s(\theta)}{2}$. That is, sniping is no longer zero sum among trading firms but is actually negative sum in the amount θ.

To derive the desired comparative statics, first examine the zero-profit condition for the snipers, (C.3). First, observe that all else equal sniping profits are strictly decreasing in the all-in trading cost $\frac{s(\theta)}{2} + \theta$ and strictly decreasing in N. Therefore, to maintain sniping profits of c_{speed} per unit time we must have $\frac{\partial s(\theta)}{\partial \theta} < 0$ and $\frac{\partial N(\theta)}{\partial \theta} < 0$. Next, examining the zero-profit condition for the liquidity provider, (C.2), notice that if $\frac{\partial \frac{s(\theta)}{2}}{\partial \theta} \leq -1$, i.e., if the all-in trading cost $\frac{s(\theta)}{2} + \theta$ is actually weakly decreasing in θ, then we would have a contradiction, because the liquidity provider's revenue $\lambda_{invest} \cdot \frac{s(\theta)}{2}$ is strictly decreasing in θ whereas the liquidity provider's losses from getting sniped would be weakly increasing. Hence, we have $-1 < \frac{\partial \frac{s(\theta)}{2}}{\partial \theta} < 0$. Together with $\frac{\partial N(\theta)}{\partial \theta} < 0$ this establishes the three claims in the proposition statement: an increase in θ decreases investment in speed $N(\theta)$, decreases the bid-ask spread $s(\theta)$, and increases investors' all-in trading costs $\theta + \frac{s(\theta)}{2}$.

C.1.11 Proof of Proposition 11

If is straightforward to see that there is an equilibrium with the same structure as in Section 6.4, with N^* trading firms of whom 1 provides liquidity and $N^* - 1$ snipe, with the bid-ask spread s^*, and equilibrium characterized as in Section 6.4 by the zero-profit conditions (6.5)-(6.6). The only difference is that after each jump in y that is larger than $\frac{s^*}{2}$, all N^* trading firms send their desired message (either a snipe or a cancel) as often as possible, namely $\bar{M} \to \infty$ times. It is random which of the $N^* \cdot \bar{M}$ messages reaches the exchange first, so just as in Section 6.4 each fast trading firm has a $\frac{1}{N^*}$ probability of winning the race. The 2ϵ pause assumption ensures that the liquidity provider can maintain depth of exactly one at all times when investors might arrive to market and ensures that stale-quote snipers can cancel unsuccessful snipes without risk that they themselves get sniped in the event of a subsequent jump in the opposite direction. The $\bar{M} \to \infty$ limit ensures that the probability that a slow trading firm can win the race to snipe goes to zero because of the δ speed disadvantage. Hence there is no role for slow trading firms in this equilibrium, just as in Section 6.4.

C.2 Supporting Materials for Section 7.3.3: How Long is Long Enough to Stop the Speed Race?

C.2.1 Calibration of Equation (7.1)

While we lack the data necessary to calibrate (7.1) in a fully satisfactory way, we can use a combination of our ES-SPY analysis, information from HFT public filings and information from discussions with market participants to obtain a rough sense of magnitudes.

There are two potential interpretations of δ. The first interpretation is that it represents the year-on-year speed improvements of state-of-the-art HFTs. Figures from the website of microwave provider McKay Brothers suggest that, in New York - Chicago trades like ES-SPY, the difference in one-way latency between state-of-the-art in 2014 versus 2013 was comfortably less than 100 microseconds. For equities only trades, since all trading venues are in server farms in New Jersey, this figure would be comfortably less than 10 microseconds. A second interpretation of δ is that it represents the speed difference between HFTs and sophisticated algorithmic trading firms that are not at the cutting edge of speed. A simple way to proxy for this speed difference is to use the difference in speed between microwaves and fiber-optic cables: about 2.5 milliseconds one-way for New York - Chicago trades, and on the order of 50 microseconds for equities-only trades.

λ_{jump} and $\mathbb{E}(J)$ represent the frequency and size of sniping opportunities, or more precisely jumps that would be sniping opportunities if they occur during the correct δ interval of the batch interval. In our ES/SPY data, there are roughly 200,000 sniping opportunities per year (800 per day times 250 days), averaging roughly 0.01 per share. $\bar{Q}$ represents the depth of the order book in the auction. In our SPY data, top-of-book depth averages about 40,000 shares.[64] Our theory predicts that frequent batch auctions should narrow spreads – which both increases the likelihood that a jump creates a sniping opportunity and increases the profits of any given sniping opportunity – and increase depth. We thus double the product of these figures from ES/SPY, i.e., we use $\lambda_{jump}\mathbb{E}(J)\bar{Q} = 2 * 200,000 * 0.01 * 40,000 = 160,000$ annualized.

c_{speed} represents the cost of speed. Data on speed expenditures by high-frequency trading firms are mostly proprietary. An exception is that the high-frequency trading firm GETCO's financial data was released publicly when GETCO merged with Knight Capital Group, because the merger filing detailed the financials of each of the merging firms separately (KCG, 2013). In 2012, the last year for which standalone GETCO data are available, GETCO spent $84M on colocation and data line expenses, $31M on capital expenditures, and $161M on employee compensation. For each of these expenses it is not possible to know how much of the expense was related to speed

[64] This figure represents 2011 NYSE SPY depth at the top of the book, multiplied by the inverse of NYSE's market share to get an estimate for market-wide depth (cf. Section 5.2.2).

per se, so to be conservative suppose that $100M of the total relates to speed. This $100M figure then represents the annual cost of speed for a single firm, for all of its trading activity, not just ES/SPY. If we assume that ES/SPY represents 1% of the speed race – under this assumption, our $75M estimate for the total prize in ES/SPY would imply a total prize overall of $7.5bn per year – then the annual cost of speed that should be attributed to ES/SPY is 1% of the $100mm, for c_{speed} of $1 million per year.

Using these estimates, we can rearrange (7.1) as $\tau > \frac{\delta \lambda_{jump} \mathbb{E}(J) \bar{Q}}{c_{speed}}$ to bound τ. Under the first interpretation of δ, we obtain a lower bound of 16ms, and under the second interpretation of δ the lower bound is 400ms.

As should be clear from the discussion of each of the inputs above, these figures should be interpreted as giving no more than an extremely rough sense of magnitudes. One can easily tinker with each of the inputs above to get a bound for τ that is an order of magnitude larger (e.g., if depth $\bar{Q}$ is considerably higher, ES/SPY is considerably less than 1% of the speed race, or c_{speed} is lower), or an order of magnitude smaller.

Below we analyze a modification of our 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 that in Section 7.3.3 but with a simpler and less stringent sufficient condition under which frequent batch auctions stop the speed race: $\tau > \delta_{slow}$. For New York - Chicago trades, δ_{slow} is about 4ms under the first interpretation of slow traders and about 7ms under the second interpretation. For equities only trades, δ_{slow} would be less than 1ms under the first interpretation of slow traders and at most a few milliseconds under the second interpretation.

C.2.2 Modification to the Model: Endogenous Entry with Discrete-Time Information Arrival

To the extent that the information y about the value of the security x is information about other security prices – e.g., if x is SPY, y is information about price changes in ES – then the widespread adoption of frequent batch auctions would change the arrival process for information from a continuous-time process to a discrete-time process. In this appendix we briefly discuss a modification of the model in which changes in y only occur in discrete time, as one batch interval ends and the next begins. Formally, assume that y_t is any discrete-time martingale process with updates at time $t = 0, \tau, 2\tau, 3\tau, \ldots$, for $\tau > 0$. Under this modification, we obtain an equilibrium analogous to that in Section 7.3.3 but with a simpler condition characterizing the batch interval

necessary to stop the speed race. The condition for τ is:

$$\tau\text{: there is no integer } k \text{ such that } \delta_{fast} < k\tau < \delta_{slow} \tag{C.4}$$

Under condition (C.4), any time there is an update to y_t, both slow and fast traders observe the update during the same batch interval. A simple necessary condition for (C.4) is $\tau > \delta$ and a simple sufficient condition for (C.4) is

$$\tau > \delta_{slow} \tag{C.5}$$

The following proposition describes the equilibrium.

Proposition 12 (Equilibrium of Frequent Batch Auctions with Endogenous Entry and Discrete–Time Information Arrival)**.** *Consider a modification of the model of Section 7.3.3 in which information y evolves in discrete time, with updates occurring as one batch interval ends and the next begins. Let τ satisfy* (C.4)*, a sufficient condition for which is* (C.5)*. Then any equilibrium is analogous to the equilibrium in Proposition 9:*

- *Slow trading firms collectively provide at least $\bar{Q}$ of depth at zero bid-ask spread.*
- *There is zero investment in speed.*
- *Investors have to wait a positive amount of time to trade.*

Proof. Observe that given the assumption about information arrival there is zero benefit to speed if τ satisfies (C.4); both slow trading firms and fast trading firms have exactly the same information at the conclusion of each batch interval. Hence, there is no reason to pay the cost c_{speed} and there are only slow trading firms in equilibrium. Given this observation, the result that there is at least $\bar{Q}$ of depth at zero bid-ask spread, in any equilibrium, follows directly from the arguments in the proof for the exogenous entry case, Proposition 8.