Ms. Melissa D. Jurgens
Secretary of the Commission
Commodity Futures Trading Commission
Three Lafayette Centre
1155 21st Street NW
Washington DC 20581

Re: CFTC Concept Release – Risk Controls and System Safeguards for Automated Trading Environments
RIN 3038-AD52

Dear Ms. Jurgens,

We are pleased to provide our comment on the CFTC's concept release for Risk Controls and System Safeguards for Automated Trading Environments. We are market design researchers with academic and practical experience in a wide variety of market design contexts. The market design approach assumes that participants in a market act optimally in their rational self-interest with respect to market rules, but takes seriously the possibility that the market rules themselves may be sub-optimal.[1] We believe that this approach brings a useful perspective to the issues you raise in the concept release. In particular, it avoids the "is HFT good or evil?" debate, which we think is often counterproductive, and instead focuses attention on whether the current market design is optimal.

Our research suggests that the ongoing arms race amongst HFT firms – in which large sums of money are devoted to seemingly trivial speed improvements (sometimes measured in millionths of seconds) – is a *symptom* of a basic flaw in modern financial market design: continuous-time trading. We begin by showing that the continuous-time limit order book does not actually work in continuous-time: market correlations completely break down at high-frequency time horizons, which creates technical arbitrage opportunities available to whomever is fastest, inducing an arms race for speed. We then show that the arms race has two kinds of negative effects: it is socially wasteful, and it harms liquidity. The arms race and these negative effects are thus a consequence of flawed market design.

As an alternative, we recommend frequent batch auctions – *uniform price sealed-bid* double auctions conducted at frequent but discrete time intervals, such as once per second (or even once per 100ms). We show that frequent batching has two kinds of benefits relative to continuous trading. First, it stops the arms race (intuitively, if trading is just once per second, then microsecond speed improvements are much less valuable). Second, it transforms competition on speed into competition on price (intuitively, if there is new public information that many algorithmic trading firms observe at roughly the same time, firms will compete over who is willing to pay the most or sell for the least instead of competing over who can respond the fastest). Consequently, frequent batching leads to greater social welfare and improved liquidity.

[1] As Milton Friedman wrote over fifty years ago in Capitalism and Freedom: "The existence of a free market does not of course eliminate the need for government. On the contrary, government is essential both as a forum for determining the 'rules of the game' and as an umpire to interpret and enforce the rules decided on." More recently, the Nobel Prize in Economic Sciences was awarded in 2012 to Alvin E. Roth for his work on market design.

Our arguments are detailed in our paper, "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," which is attached. Here, we emphasize a few points that may be especially relevant to regulatory policy and implementation.

First, the change to frequent batching can be viewed as a backend, technocratic reform to market design that directly addresses an observed problem of the continuous-time limit order book market: the speed race. Exchanges would appear to operate essentially as they do today with the exception that orders would be accumulated over the batch interval before they are processed in batch at the clearing price, all in about the blink of an eye. Ordinary investors might not even notice the difference. Sophisticated algorithmic trading firms would continue to play a critical role in financial markets.

Second, batching helps fundamental investors by improving liquidity via deeper markets and narrower spreads. We show that the resources invested in the speed race ultimately come out of the pockets of investors, as opposed to HFT firms, via increased trading costs due to thinner markets and wider-than-necessary spreads. Unless one is willing to argue that waiting a second (or 100ms) is very costly for fundamental investors, batching makes fundamental investors better off.

Third, as we discuss in Section 8 of the paper, frequent batching has several potential market stability benefits. First, frequent batching is computationally simple for exchanges. Exchanges have a discrete block of time to compute and report the auction outcome, which prevents order backlog and incorrect time stamps. Second, batching gives trading algorithms a discrete time period to process current prices before deciding on future trades. There is no incentive to trade off code robustness for speed. Third, changing from continuous time to discrete time improves the regulatory paper trail. And fourth, the greater market depth from batching suggests that the market is apt to be less vulnerable to mini flash crashes.

A fundamental source of stability problems with continuous markets is that they implicitly assume that computers and communications are infinitely fast. Computers are fast but not infinitely so. Frequent batching respects the limits of computers.

We would be happy to provide any additional information or help that you may desire on this topic.

Sincerely yours,

Eric Budish, University of Chicago

Peter Cramton, University of Maryland

John Shim, University of Chicago

Attachment: Eric Budish, Peter Cramton, and John Shim, "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," University of Chicago, December 2013.

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

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

December 11, 2013

Abstract

We argue that the continuous limit order book is a flawed market design and propose that financial exchanges instead use frequent batch auctions: uniform-price sealed-bid double auctions conducted at frequent but discrete time intervals, e.g., every 1 second. Our argument has four parts. First, we use millisecond-level direct-feed data from exchanges to show that the continuous limit order book market design does not really "work" in continuous time: market correlations that function properly at human-scale time horizons completely break down at high-frequency time horizons. Second, we show that this correlation breakdown creates frequent technical arbitrage opportunities, available to whomever is fastest, which in turn creates an arms race to exploit such opportunities. Third, we develop a simple new theory model motivated by these empirical facts. The model shows that the arms race is not only socially wasteful – a prisoner's dilemma built directly into the market design – but that its cost is ultimately borne by fundamental investors via wider spreads and thinner markets. Last, we show that frequent batch auctions eliminate the arms race, both because they reduce the value of tiny speed advantages and because they transform competition on speed into competition on price. Consequently, frequent batch auctions lead to narrower spreads, deeper markets, and increased social welfare.

*First version: July 2013. 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, Berkeley, NBER Market Design, NYU, MIT, Harvard, Columbia, and Spot Trading, and to Susan Athey, Larry Ausubel, Eduardo Azevedo, 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, Emir Kamenica, Brian Kelly, Pete Kyle, Gregor Matvos, Paul Milgrom, Toby Moskowitz, Matt Notowidigdo, Mike Ostrovsky, David Parkes, Al Roth, Gideon Saar, Jesse Shapiro, Spyros Skouras, Lars Stole, Geoff Swerdlin, Richard Thaler, Brian Weller, Michael Wellman and Bob Wilson. We thank Daniel Davidson, 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. Microwave technology has further reduced round-trip transmission time, first to 10ms, then to 9ms, and most recently to 8.5ms. There are reports of analogous speed races occurring at the level of microseconds (millionths of a second) and even nanoseconds (billionths of a second).[1]

We argue that this high-frequency trading "arms race" is a manifestation of a basic flaw in financial market design: financial markets operate *continuously*. That is, it is possible to buy or sell stocks or other securities at literally any instant during the trading day. We argue that the continuous limit order book market design that is currently predominant in financial markets should be replaced by frequent batch auctions – uniform-price sealed-bid double auctions conducted at frequent but discrete time intervals, e.g., every 1 second. Our argument against continuous limit order books and in favor of frequent batch auctions has four parts.

The first part of our paper uses millisecond-level direct-feed data from exchanges to show that the continuous limit order book market design does not really "work" in continuous time: market correlations that function properly (i.e., obey standard asset pricing relationships) at human-scale time horizons completely break down at high-frequency time horizons. Consider Figure 1.1. The figure depicts the price paths of the two largest securities that track the S&P 500 index, the iShares SPDR S&P 500 exchange traded fund (ticker SPY) and the E-mini Future (ticker ES), on an ordinary trading day in 2011. In Panel A, we see that the two securities are nearly

[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.

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.[2] Similarly, we find that pairs of equity securities that are highly correlated at human time scales (e.g., the home-improvement companies Home Depot and Lowe's or the investment banks Goldman Sachs and Morgan Stanley) have essentially zero correlation at high frequency.

This correlation breakdown may seem like just a theoretical curiosity, and it is entirely obvious ex-post. There is nothing in current financial market architecture that would enable correlated securities' prices to move at *exactly* the same time, because each security trades on its own separate continuous limit order book; in auction design terminology, financial markets are a collection of separate single-product auctions, rather than a single combinatorial auction. Can correlation breakdown be safely ignored, much as the failure of Newtonian mechanics at the quantum level can safely be ignored in most of day-to-day life?

The second part of our argument is that this correlation breakdown has real consequences: it creates purely technical arbitrage opportunities, available to whomever is fastest, which in turn create an arms race to exploit these arbitrage opportunities. Consider again Figure 1.1, Panel D, at time 1:51:39.590 pm. At this moment, the price of ES has just jumped roughly 2.5 index points, but the price in the SPY market has not yet reacted. This creates a temporary profit opportunity – buy SPY and sell ES – available to whichever trader acts the fastest. We calculate that there are on average about a thousand such arbitrage opportunities per day in ES-SPY, worth on the order of $75 million per year. And, of course, ES-SPY is just the tip of the iceberg. 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 to the universe of trading opportunities very similar to ES-SPY – let alone to trading opportunities that exploit more subtle pricing relationships – suggests that the annual sums at stake are in the billions.

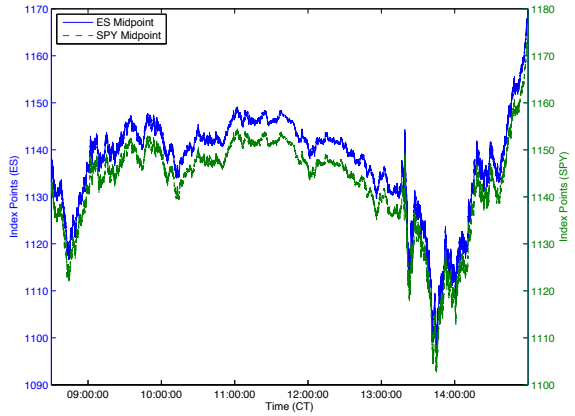
It is also instructive to examine how the ES-SPY arbitrage has evolved over time. Over the time period of our data, 2005-2011, we find that the *duration* of ES-SPY arbitrage opportunities

[2]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. The 4ms correlation is also essentially zero, for all three of these methods of handling the speed-of-light issue. See Section 4 for further details. We would also like to suggest that the fact that special relativity plays a role in these calculations is support for frequent batch auctions.

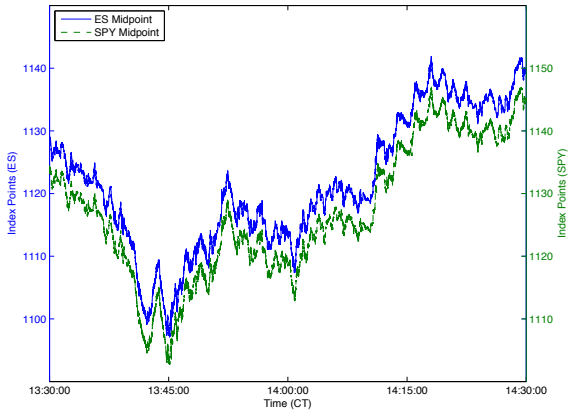
Figure 1.1: ES and SPY Time Series at Human-Scale and High-Frequency Time Horizons

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 an ordinary trading day (08/09/2011) at different time resolutions: the full day (a), an hour (b), a minute (c), and 250 milliseconds (d). Midpoints for each security are constructed by taking an equal-weighted average of the top-of-book bid and ask. 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 footnote 14. For details regarding the data, see Section 3.

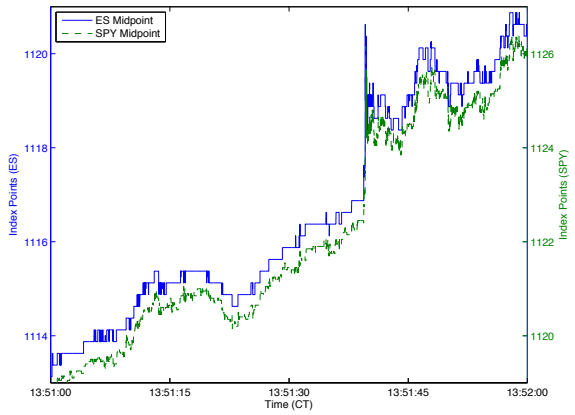
(a) Day



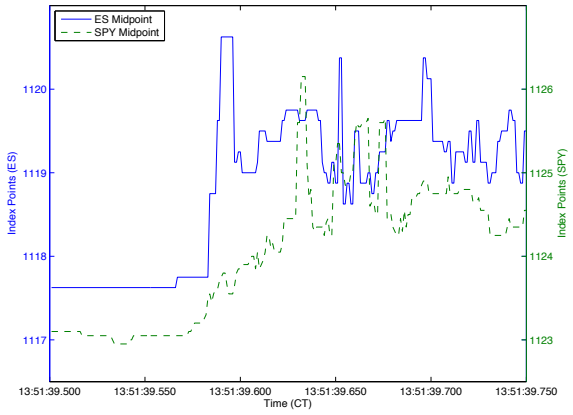
(b) Hour



(c) Minute



(d) 250 Milliseconds



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, which is intuitive given that it is changes in prices that create temporary relative mispricings. These findings suggest that while there is an arms race in speed, the arms race does not actually eliminate the arbitrage opportunities; rather, it just continually raises the bar for capturing them. A complementary finding, in the correlation breakdown analysis, is that the number of milliseconds necessary for economically meaningful correlations to emerge has been steadily decreasing over the time period 2005-2011; but, in all years, market correlations are essentially zero at high-enough frequency. Overall, our analysis suggests that the prize in the arms race should be thought of more as a mechanical "constant" of the continuous limit order book market design, rather than as an inefficiency that is competed away over time.

The third part of our paper develops a simple theory model informed by these empirical facts. The model serves two related purposes: it is a critique of the continuous limit order book market design, and it identifies the economic implications of the HFT arms race. In the model, there is a security, x, that trades on a continuous limit order book 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. This setup can be interpreted as a "best case" scenario for price discovery and liquidity provision in a continuous limit order book, abstracting from issues such as asymmetric information, inventory costs, etc.

Given the model setup, one might expect that Bertrand competition among market makers drives the bid-ask spread in the market for x to zero. But, consider what happens 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, market makers providing liquidity in the market for x (SPY) will send a message to the exchange to adjust their quotes – withdraw their old quotes and replace them with new, higher, quotes based on the new signal y (price of ES). At the exact same time, however, other market participants will try to "pick off" (or "snipe") the old quotes – send a message to the exchange attempting to buy x at the old ask price, before the liquidity providers can adjust. Hence, there is a race. And, since each one liquidity provider is in a race against many other market participants attempting to pick them off – and continuous limit order books process message requests in serial (i.e., one

at a time), so only the first message to reach the exchange matters – liquidity providers usually lose the race. This is the case even if liquidity providers can invest in speed technologies such as the Spread Networks cable – which they do in equilibrium of our model – since the other market participants looking to snipe stale quotes invest in speed as well. In a competitive market, liquidity providers must incorporate the cost of getting sniped into the bid-ask spread that they charge; this is a purely technical cost of liquidity provision caused by the continuous limit order book market design.[3]

This same phenomenon – liquidity providers getting picked off in the race to respond to purely public information – also causes continuous limit order book markets to be unnecessarily thin. That is, it is especially expensive for fundamental investors to trade large quantities of stock. The reason is that picking-off costs scale linearly with the quantity liquidity providers offer in the book – if quotes are stale, they will get picked off for the whole amount – whereas the benefits of providing a deep book scale less than linearly with the quantity offered, since only some fundamental investors wish to trade large amounts. Hence, not only is there a positive bid-ask spread even without asymmetric information about fundamentals, but markets are unnecessarily thin, too.

In addition to showing that the arms race induced by the continuous limit order book harms liquidity, our model also shows that the arms race is socially wasteful, and can be interpreted as a prisoner's dilemma. In fact, these two negative implications of the arms race – reduced liquidity and socially wasteful investment – can be viewed as opposite sides of the same coin. In equilibrium of our model, all of the money that market participants invest in the speed race comes out of the pockets of fundamental investors, via wider bid-ask spreads and thinner markets.[4] Moreover, these negative implications of the arms race are not competed away over time – they depend neither on the magnitude of potential speed improvements (be they milliseconds, microseconds, nanoseconds, etc.), nor on the cost of cutting edge speed technology (if speed costs grow lower over time there is simply more entry). These results tie in nicely with our empirical findings above which found that the prize in the arms race is essentially a constant.

[3]Our model can be interpreted as providing a new source of bid-ask spreads, incremental to the explanations of inventory costs (Roll, 1984), asymmetric information (Kyle, 1985; Glosten and Milgrom, 1985), and search costs (Duffie, Garleanu and Pedersen, 2005). Mechanically, our source of bid-ask spread is most similar to that in Glosten and Milgrom (1985), namely a liquidity provider sometimes gets exploited by another market participant who knows that the liquidity provider's quotes are mispriced. There are two key differences. First, in our model the liquidity provider has *exactly* the same information as the other market participants who are picking him off. There are no "informed traders" with asymmetric information. Second, the bid-ask spread in our model can be eliminated with a change to market design; under frequent batch auctions, Bertrand competition among market makers does in fact drive the bid-ask spread to zero. See further discussion in Section 6.3.1.

[4]A point of clarification: our claim is not that markets are less liquid today than before the rise of electronic trading and HFT; our claim is that markets are less liquid today than they would be under an alternative market design which eliminated sniping costs. See Section 6.3.1 for discussion.

The fourth and final part of our argument shows that frequent batch auctions are an attractive market design response to the HFT arms race. Batching eliminates the arms race for two reasons. First, and most centrally, batching substantially reduces the value of a tiny speed advantage. In our model, if the batching interval is τ, then a δ speed advantage is only $\frac{\delta}{\tau}$ as valuable as it is under continuous markets. So, for example, if the batching interval is 1 second, a 1 millisecond speed advantage is only $\frac{1}{1000}$ as valuable as it is in the continuous limit order book market design. Second, and more subtly, batching changes the nature of competition among fast traders, encouraging competition on price instead of speed. Intuitively, in the continuous limit order book market design, it is possible to earn a rent based on a piece of information that many fast traders observe at basically the same time – be it a mundane everyday event like a jump in the price of ES, or a more dramatic event such as a Fed announcement – because continuous limit order books process orders in serial, and *somebody is always first*.[5] In the batch market, by contrast, if multiple traders observe the same information at the same time, they are forced to compete on price instead of speed.

For both of these reasons, frequent batch auctions eliminate the purely technical cost of liquidity provision in continuous limit order book markets associated with stale quotes getting sniped. Batching also resolves the prisoner's dilemma associated with continuous limit order book markets, and in a manner that allocates the welfare savings to fundamental investors. In equilibrium of the frequent batch auction, relative to continuous limit order books, bid-ask spreads are narrower, markets are deeper, and social welfare is greater.

Our theoretical argument for frequent batch auctions as a response to the HFT arms race focuses on bid-ask spreads, market depth, and socially wasteful expenditure on speed. We also suggest several reasons why switching from the continuous limit order book to frequent batch auctions may have market stability benefits that are outside the model. First, frequent batch auctions give exchange computers a discrete period of time to process current orders before the next batch of orders needs to be dealt with. This simplifies the exchange's computational task, perhaps making markets less vulnerable to incidents like the August 2013 NASDAQ outage (Bunge, Strasburg and Patterson, 2013), and also prevents order backlog and incorrect time stamps, issues that were salient during the Facebook IPO and the Flash Crash (Strasburg and Bunge, 2013; Nanex, 2011). In a sense, the continuous limit order book design implicitly assumes that exchange computers are infinitely fast; computers are fast, but not infinitely so. Second, frequent batch auctions give trading algorithms a discrete period of time to process recent prices and outcomes before deciding on their next trades. While no market design can entirely prevent programming

[5] In fact, our model clarifies that fast traders can earn a rent even from information that they observe at *exactly* the same time as other fast traders. This can be viewed as the logical extreme of what Hirshleifer (1971) called "foreknowledge" rents, built directly into the continuous limit order book market design.

errors (e.g., the Knight Capital incident, see Strasburg and Bunge (2012)), batching makes the programming environment more natural, because algorithms can be written with certainty that they will know time t prices in time to make time $t+1$ trading decisions. Batching also reduces the incentive to trade off code robustness for speed; error checking takes time. Third, frequent batch auctions produce a better paper trail for regulators, exchanges, market participants and fundamental investors: all parties know exactly what occurred at time t, know exactly what occurred at time $t+1$, etc., which is not the case under the current equity market structure (cf. SEC and CFTC, 2010). Last, the market thickness results from the theory model can also be interpreted as a stability benefit of frequent batch auctions, since thin markets are more vulnerable to what have come to be known as "mini flash crashes". While these arguments are necessarily less formal than the main analysis, we include them due to the importance of market stability to current policy discussions (e.g., SEC and CFTC (2010); Niederauer (2012)).

We wish to reiterate that we are proposing batch auctions conducted at very fast intervals, such as once per second. The principle guiding this aspect of our proposal is that we seek a minimal departure from current market design subject to realizing the benefits of batching relative to continuous limit order books. There are two other recent papers, developed independently from ours and coming from different methodological perspectives, that also make cases for frequent batching: Farmer and Skouras (2012*b*) and Wah and Wellman (2013).[6] There is also an older literature arguing for batch auctions conducted at much lower frequency, such as just 3 times per day (Cohen and Schwartz (1989); Economides and Schwartz (1995)), however, one might worry that such a radical change would have unintended consequences; to give just one example, in the functioning of derivatives markets. Running batch auctions once per second, on the other hand, or even once per 100 milliseconds (respectively, 23,400 and 234,000 times per day per security) is more of a backend, technocratic proposal than a radical redesign. Sophisticated algorithmic trading firms would continue to play a critical role in financial markets. Ordinary investors might not even notice the difference.

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

[6]Farmer and Skouras (2012*b*) is a policy paper commissioned by the UK Government's Foresight report which makes a case for frequent batch auctions based on ideas from complexity theory, market ecology, and econophysics. Wah and Wellman (2013) uses a zero-intelligence agent-based simulation model to compare frequent batch auctions to continuous limit order books and study issues of market fragmentation.

among policy makers, the press, and market microstructure researchers.[7,8] The market design perspective assumes that market participants will optimize with respect to market rules as given, but takes seriously the possibility that we have the wrong market rules in place. Our question is not whether HFT firms perform a useful market function – our model takes as given that they do – but whether, through changing financial market design from continuous to discrete, this same function can be elicited more efficiently, by reducing the rent-seeking component of HFT.

The rest of the paper is organized as follows. Section 2 briefly reviews the rules of the continuous limit order book. Section 3 describes our direct-feed data from NYSE and the CME. Section 4 presents the correlation breakdown results. Section 5 presents the technical arbitrage results. Section 6 presents the model, and solves for and discusses the equilibrium of the continuous limit order book. Section 7 proposes frequent batch auctions, shows why they eliminate the arms race, and discusses their equilibrium properties. Section 8 discusses market stability. Section 9 concludes. Proofs are contained in the Appendix.

2 Brief Description of Continuous Limit Order Books

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

The basic building block of this market design is the limit order. A limit order specifies a price, a quantity, and whether the order is to buy or 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.

[7]Within the market design literature, some especially relevant papers include Roth and Xing (1994, 1997) on serial versus batch processing and the importance of the timing of transactions, Roth and Ockenfels (2002) on bid sniping, Klemperer (2004) for a variety of illustrative examples of failed real-world auction designs, and Bhave and Budish (2013) for a case study on the use of market design to reduce rent seeking. See Roth (2002, 2008) and Milgrom (2004, 2011) for surveys. See Jones (2013) for a recent survey of the burgeoning market microstructure literature on HFT. This literature mostly focuses on the impact of high-frequency trading on market quality, taking market design as exogenously fixed (e.g., Hendershott, Jones and Menkveld (2011); Brogaard, Hendershott and Riordan (2012); Hasbrouck and Saar (2013); Weller (2013)). A notable exception is Biais, Foucault and Moinas (2013), who study the equilibrium level of investment in speed technology, find that investment can be socially excessive, and informally discuss policy responses; see further discussion in Section 6.3.4. See also O'Hara (2003); Biais, Glosten and Spatt (2005); Vives (2010) for surveys of market microstructure more broadly.

[8]In policy discussions, frequent batch auctions have received some attention, but less so than other policy ideas such as minimum resting times, excessive order fees, and transaction taxes (cf. Jones (2013)). Our sense is that these latter ideas do not address the core problem, and seem to be motivated by the view that "HFT is evil and must be stopped." A notable exception is the policy paper by Farmer and Skouras (2012*b*) for the UK Government's Foresight report, mentioned in the previous footnote. Unfortunately, it was just one of 11 distinct policy papers commissioned for the report, and the executive summary of the report dismissed frequent batching as "unrealistic and draconian" without much explanation (The Government Office for Science (2012); pg. 14).

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.

3 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.[9] 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. From our perspective, the key advantage of direct-feed data is that the timestamps are as

[9]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.

accurate as possible.

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, as well as half days (e.g., day after Thanksgiving).

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 handler for each raw data format and update the state of the order book after every new message.[10]

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 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 and the Philadelphia Stock Exchange. However, regulatory feed data have time stamps that are based on the time at which the data are provided to market regulators, and practitioners estimate that the TAQ's timestamps are on the order of tens to hundreds of milliseconds delayed relative to the direct-feed data that comes directly from the exchanges (see Ding, Hanna and Hendershott (2013); our own informal comparisons confirm this as well). 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 millisecond-level time stamps are as accurate as possible. In particular, these are exactly the same data that HFT firms use to make trading decisions.

[10]Our feed handlers will be made publicly available in the data appendix.

4 Market Correlations Break Down at High-Enough Frequency

In this section we report two sets of results. First, we show that market correlations completely break down at high-enough frequency. That is, securities that are highly correlated at human time scales have essentially zero correlation at high-frequency time scales. Second, we show that the market has gotten faster over time in the sense that, in each year from 2005-2011, the number of milliseconds necessary for economically meaningful correlations to emerge has been steadily decreasing. This is evidence of the speed arms race. Invariably, however, correlations break down at high-enough frequency.

Before proceeding, we emphasize that the first finding – which is an extreme version of a phenomenon discovered by Epps (1979)[11] – is obvious from introspection alone, at least ex-post. There is nothing in current market architecture – in which each 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. We also emphasize that the first finding is difficult to interpret in isolation. It is only in Section 5, when we show that correlation breakdown is associated with frequent technical arbitrage opportunities, available to whomever is fastest, that we can interpret correlation breakdown as a meaningful issue as opposed to simply a theoretical curiosity.

4.1 Correlation Breakdown

Figure 1.1, in the introduction, depicts the price paths of ES and SPY on an ordinary day, over the course of the full trading day, an hour, a minute, and 250 milliseconds. As is expected, given that both securities track the S&P 500 index, their returns are highly correlated over the course of the trading day. The correlation remains high when we look at a one-hour or one-minute slice of the trading day. However, when we zoom in on a 250 millisecond interval, we see that the correlation between the two securities breaks down. For example, during the time interval 1:51:39.580 to 1:51:39.600 the price of ES changes a lot while the price of SPY is relatively flat, whereas during the time interval 1:51:39.630 to 1:51:39.640 the reverse is true.

In this section we show that the pattern depicted in Figure 1.1 holds throughout our data. We first present correlation breakdown results for ES and SPY, which is a canonical example of an asset pricing relationship governed by no arbitrage. We then present analogous results for

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

the correlation matrix amongst the largest equities in the US stock market, which suggests that the market covariance matrix, which underpins asset pricing models such as the CAPM, is not meaningful at high frequency.

4.1.1 ES and SPY

Figure 4.1 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,[12] but almost completely breaks down at high-frequency time intervals. The 10 millisecond correlation is just 0.1016, and the 1 millisecond correlation is just 0.0080.

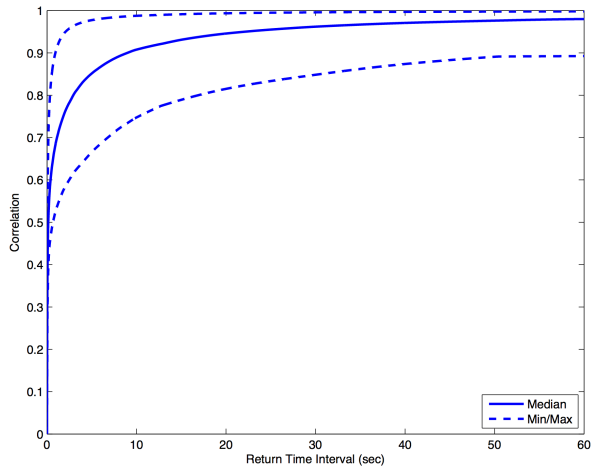
We consider several other measures of the ES-SPY correlation, varying along three dimensions. First, we consider both equal-weighted bid-ask midpoints and quantity-weighted bid-ask midpoints. Whereas equal-weighted midpoints place weight of $\frac{1}{2}$ on the bid and the ask, quantity-weighted midpoints 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} denotes the quantity offered at the bid at time t and Q_t^{ask} denotes the quantity offered at the ask. 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 New York time, treating Chicago events as occurring 4ms later in New York than they do in Chicago. That is, New York time 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 Chicago time, 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

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

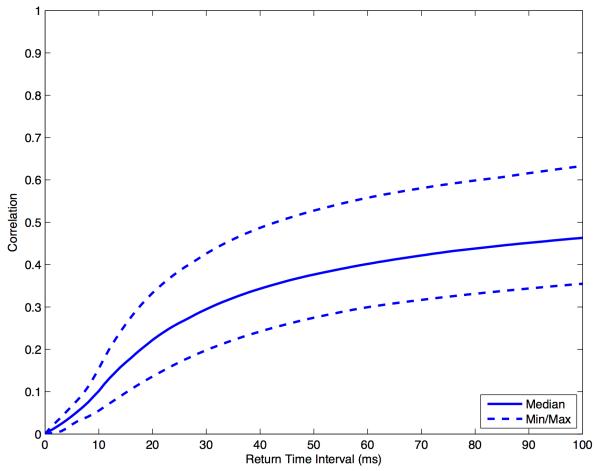
Figure 4.1: ES and SPY Correlation by Return Interval: 2011

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 midpoints are constructed using the equal-weighted average of the bid and ask in each security. The correlation is computed using simple arithmetic returns over a range of time intervals, measured in milliseconds. 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 regarding the computation of correlations, see the text of Section 4.1.1. For more details on the data, refer to Section 3.

(a) Correlations at Intervals up to 60 Seconds



(b) Correlations at Intervals up to 100 Milliseconds



Chicago and New York.

Table 1 displays the ES-SPY correlation for varying time intervals, averaged over all trading days in 2011, over each of our 12(= 2×2×3) methods of computing the correlation. As can be seen from the table the pattern depicted in Figure 4.1 is robust across these various specifications.[13]

4.1.2 Equities-Market Correlation Matrix

Table 2a displays the correlation at different time intervals between pairs of equity securities that are highly correlated, for instance, the oil companies Exxon-Mobil (XOM) and Chevron (CVX). Table 2b displays the correlation matrix amongst the 5 largest market capitalization US equities at varying time horizons. We follow the main specification used in Section 4.1.1 and use equal-weighted midpoints and simple returns. Note that the speed-of-light issue is not relevant for this exercise, since all of these securities trade on the NYSE.

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 amongst these securities, but not at high-frequency time scales such as 1ms or 100ms.

4.2 Correlation Breakdown Over Time

Figure 4.2 displays the ES-SPY correlation versus time interval curve that we depicted above as Figure 4.1 Panel (b), but separately for each year in the time period 2005-2011 that is covered in our data. As can be seen in the figure, the market has gotten faster over time in the sense that economically meaningful market correlations emerge more quickly in the later years of our data than in the early 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 Correlation Breakdown Creates Technical Arbitrage Opportunities

In this section we show that the correlation breakdown phenomenon we documented in Section 4 is associated with frequent technical arbitrage opportunities, available to whichever trader acts

[13] 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 1: 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 correlation 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^{ask} and Q_t^{bid} represent the quantity offered as 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), 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), and a trader equidistant from the two locations (Mid). For more details on these correlation computations, See Section 4.1.1. For more details on the data, refer to Section 3.

Panel 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

Panel 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

Table 2: 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 3.

(a) Pairs of Related Companies

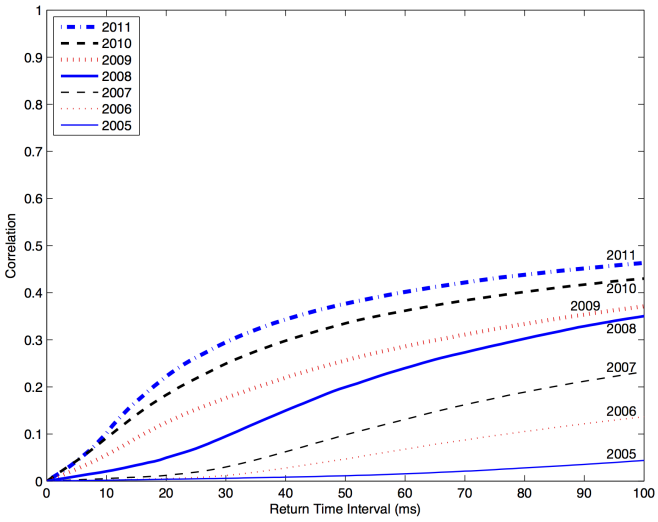
	1 ms	100 ms	1 sec	10 sec	1min	10 min	30 min
HD-LOW	0.008	0.101	0.192	0.434	0.612	0.689	0.704
GS-MS	0.005	0.094	0.188	0.405	0.561	0.663	0.693
CVX-XOM	0.023	0.284	0.460	0.654	0.745	0.772	0.802
AAPL-GOOG	0.001	0.061	0.140	0.303	0.437	0.547	0.650

(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	0.495	1.000			
GE	0.508	0.571	1.000		
JNJ	0.349	0.412	0.440	1.000	
IBM	0.554	0.512	0.535	0.464	1.000

Figure 4.2: 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. Correlations are computed using equal-weighted midpoints and simple arithmetic returns. 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 regarding the computation of correlations, see the text of Section 4.1.1. For more details on the data, refer to Section 3.



fastest. These are the kinds of profit opportunities that drive the arms race. We also explore how the nature of this arbitrage opportunity has evolved over the time period of our data, 2005-2011. The time series suggests that the prize in the speed race is more like a "constant" of continuous limit order book markets rather than an inefficiency that is competed away over time.

5.1 Computing the ES-SPY Arbitrage

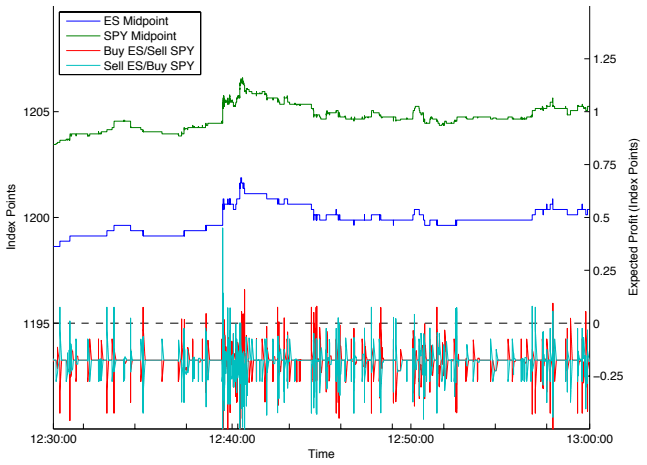
Figure 5.1 illustrates the exercise we conduct. The top portion depicts the midpoint prices of ES and SPY over the course of a fairly typical 30-minute period of trading (Panel a) and a volatile period of trading during the financial crisis (Panel b). Notice that, while there is a difference in levels between the two securities,[14] 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

[14]There are three differences between ES and SPY that drive the difference in levels. First, 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. Second, SPY is larger than ES by a term that represents S&P 500 dividends, since SPY holders receive dividends (which accumulate and then are distributed at the end of each quarter) and ES holders do not. Third, the basket of stocks in the SPY creation-redemption basket typically differs slightly from the basket of stocks in the S&P 500 index; this is known as ETF tracking error.

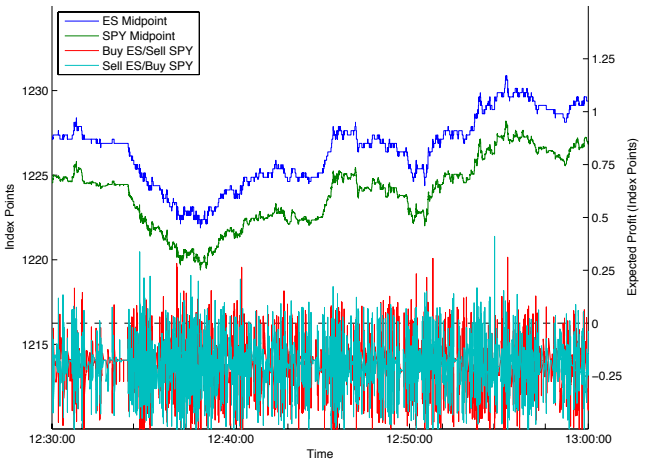
Figure 5.1: Technical Arbitrage Illustrated

Notes: This figure illustrates the technical 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 3. For details regarding the computation of instantaneous arbitrage profits, see Section 5.1.

(a) Ordinary Day



(b) Financial Crisis



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. 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. These trading opportunities represent the prize at stake in the high-frequency trading arms race, for this particular trade in this particular market.

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

$$S_t = P^{mid}_{ES,t} - 10P^{mid}_{SPY,t}, \tag{5.1}$$

where $P^{mid}_{j,t}$ 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. Define the moving-average spread between ES and SPY at millisecond t as

$$\bar{S}_t = \frac{1}{\tau^*} \sum_{i=t-\tau^*}^{t-1} S_i, \tag{5.2}$$

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 of ES and SPY at intervals of length τ^* implies that prices over this time horizon produce relatively stable spreads.[15] We define a trading rule based on the presumption that, at high-frequency time horizons, deviations of S_t from $\bar{S}_t$ are driven mostly by the correlation breakdown phenomenon we documented in Section 4. 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 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 implicit spread market according to $S_t^{bid} = P_{ES,t}^{bid} - 10P_{SPY,t}^{ask}$ and $S_t^{ask} = P_{ES,t}^{ask} - 10P_{SPY,t}^{bid}$. Note that $S_t^{bid} < S_t < S_t^{ask}$ 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.[16] Expected profits, on a

[15] 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 (cf. footnote 14) – 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.

[16] Our understanding is that this is the best simple estimate of transactions costs. A richer estimate of transactions costs 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, which 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, but might seek to trade in ES at its new price as a liquidity provider, potentially earning rather than paying half the bid-ask spread. Also complicating matters are that high-frequency trading firms' trading fees are often substantially offset by exchange rebates for liquidity provision.

per-unit spread basis, are thus:

$$\pi \;=\; \begin{cases} \bar{S}_{t_{start}} - S^{ask}_{t_{start}} & \text{if } S^{ask}_{t_{start}} < \bar{S}_{t_{start}} \\ S^{bid}_{t_{start}} - \bar{S}_{t_{start}} & \text{if } S^{bid}_{t_{start}} > \bar{S}_{t_{start}}. \end{cases} \tag{5.3}$$

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. We define the ending time of the arbitrage, t_{end}, as the first 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.[17]

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 (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.[18]

[17] Throughout the interval $[t_{start}, t_{end}]$ we compute both the actual empirical order book and a hypothetical order book which accounts for our arbitrageur's trade activity. The reason this matters is that 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 need to account for this to avoid double counting. Here is an example to illustrate. Suppose that at time t_{start} an arbitrage opportunity starts which involves buying all 10000 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 2000 shares of SPY available at price p – either a trader with 8000 shares offered at p just removed his ask, or another trader just purchased 8000 shares at the ask. Our arbitrageur buys all 10000 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 2000 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 12000 shares of SPY available at price p, then our arbitrageur would have purchased 10000 shares at time t_{start}, and then an additional 2000 (=12000-10000) shares at time t'.

[18] 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).

Table 3: 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 5.1. The data, described in Section 3, cover January 2005 to December 2011. # of Arbs/Day indicates the number of arbitrage opportunities for each trading day. Qty denotes the size of each arbitrage opportunity, measured in the number of ES lots traded. Per-Arb Profits are computed in index points as described in the text and in dollars by multiplying index points times quantity in ES lots times 50, because each ES contract has notional value of 50 times the S&P 500 index. Total Daily Profits - NYSE Data indicates the total profits from all arbitrage opportunities over the course of a trading day, based on the depth we observe in our NYSE data. Total Daily Profits - All Exchanges indicates the total profits from all arbitrage opportunities over the course of a trading day, under the assumption that including the depth from other equities exchanges multiplies the quantity available to trade by a factor of (1 / NYSE market share in SPY), as discussed in the text. % ES initiated indicates the percentage of arbitrage opportunities that are initiated by a change in the price of ES, with the remainder initiated by a change in the price of SPY. % Good Arbs indicates the percentage of arbitrage opportunities where the market uncrosses within a τ^* time interval, as described in the text, with the remainder being bad arbs. % Buy vs. Sell indicates the percentage of arbitrage opportunities in which the arbitrage involves buying spread, defined as buying ES and selling SPY, with the remainder being opportunities in which the arb involves selling spread.

		Percentile						
	Mean	1	5	25	50	75	95	99
# of Arbs/Day	801	118	173	285	439	876	2498	5353
Qty (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%							

5.2 Results on ES-SPY Arbitrage

5.2.1 Summary Statistics

Table 3 reports summary statistics on the ES-SPY arbitrage opportunity over our full dataset, 2005-2011. Throughout this section, we drop arbitrage opportunities with per-unit profitability π of strictly less than 0.05 index points, or one-half of one penny in the market for SPY.

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

Total daily profits in our data are on average $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.[19] This yields average profits of $306k per day, or roughly $75mm 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 Section 4.1.1 that correlations are higher when we treat the New York market as lagging Chicago than when we treat the Chicago market as lagging New York. Note, though, that the equities underlying the S&P 500 index trade in New York, so innovations in the index that are driven by news for particular stocks may be incorporated into SPY before ES. This may partly explain why 11% of the arbitrage opportunities are initiated by SPY rather than ES.

99.99% of the arbitrage opportunities we identify are "good arbs", meaning that large deviations of the instantaneous ES-SPY spread S_t from its moving-average level $\bar{S}_t$ 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.

5.2.2 Evolution Over Time: 2005-2011

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

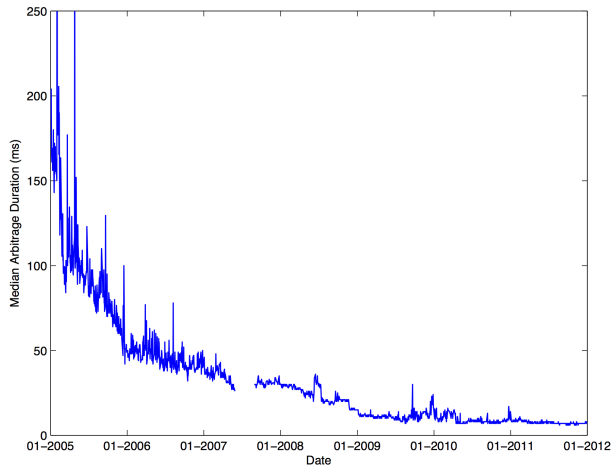
Figure 5.2 explores the duration of ES-SPY arbitrage opportunities over the time of our data set, covering 2005-2011. As can be seen in Figure 5.2a, the median duration of arbitrage opportunities has declined dramatically over this time period, from a median of 97 ms in 2005 to a median of 7 ms in 2011. Figure 5.2b 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.

[19] 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

Figure 5.2: 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. Panel (b) plots arbitrage duration against the proportion of arbitrage opportunities lasting at least this 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. The discontinuity in the time series (5/30/2007-8/28/2007) arises from omitted data resulting from data issues acknowledged by the NYSE. 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 centisecond level. See Section 5.1 for further details regarding the ES-SPY arbitrage. See Section 3 for details regarding the data.

(a) Median Arb Durations Over Time



(b) Distribution of Arb Durations Over Time

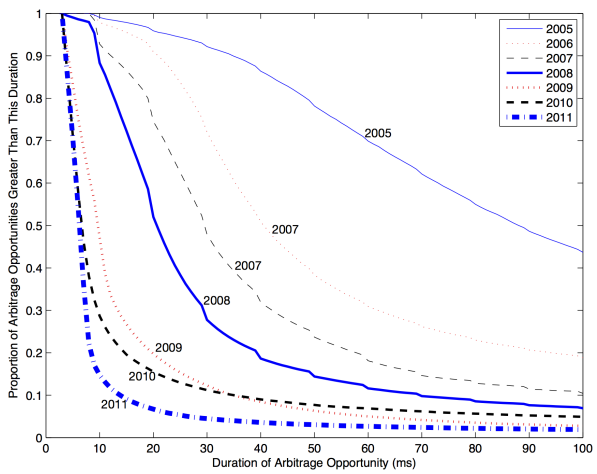
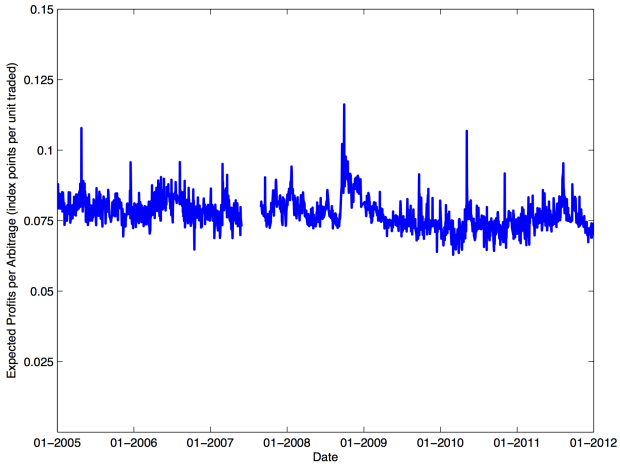


Figure 5.3: Profitability of ES & SPY Arbitrage Opportunities Over Time: 2005-2011

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. Panel (b) plots the kernel density of per-arbitrage profits for each year in our dataset. The discontinuity in the time series (5/30/2007-8/28/2007) arises from omitted data resulting from data issues acknowledged by the NYSE. We also omit 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. See Section 5.1 for details regarding the ES-SPY arbitrage. See Section 3 for details regarding the data.

(a) Median Per-Arb Profits Over Time



(b) Distribution of Per-Arb Profits Over Time

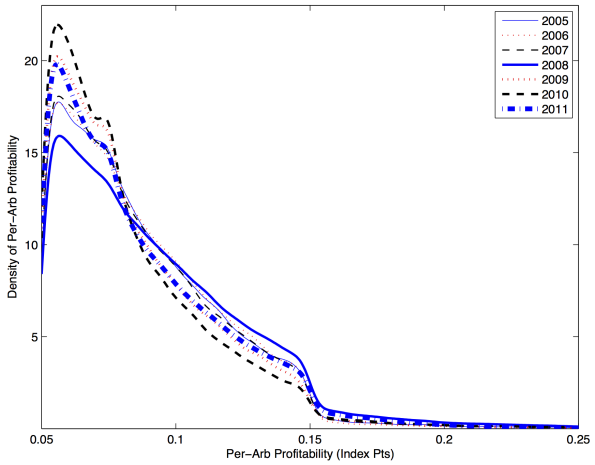


Figure 5.3 explores the per-arbitrage profitability of ES-SPY arbitrage opportunities over the time of our data set. In contrast to arbitrage durations, arbitrage profits have remained remarkably constant over time. Figure 5.3a shows that the median profits per contract traded have remained steady at around 0.08 index points, with the exception of the 2008 financial crisis when they were a bit larger. Figure 5.3b 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.

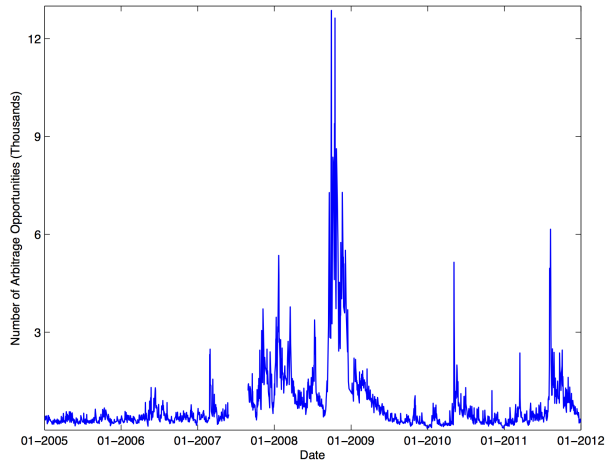
Figure 5.4 explores the frequency of ES-SPY arbitrage opportunities over the time of our data set. Unlike per-arb profitability, the frequency of arbitrage opportunities varies considerably over time. Figure 5.4a shows that the median arbitrage frequency seems to track the overall volatility of the market, 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 in light of Figure 5.1 above: 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

between the other three largest exchanges, NASDAQ, BATS and DirectEdge.

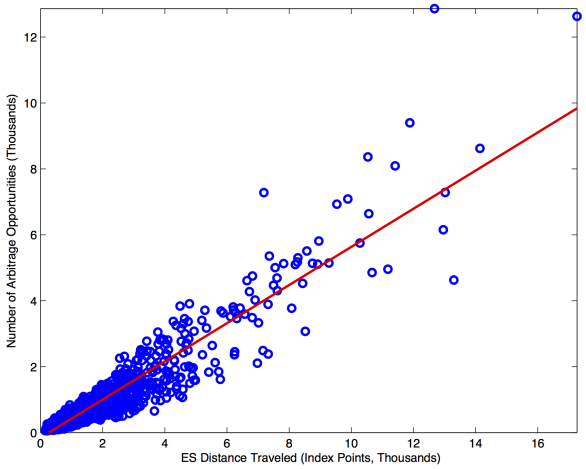
Figure 5.4: Frequency of ES & SPY Arbitrage Opportunities Over Time: 2005-2011

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. 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 5.1. The discontinuity in the time series (5/30/2007-8/28/2007) arises from omitted data resulting from data issues acknowledged by the NYSE. We also omit 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. See Section 5.1 for details regarding the ES-SPY arbitrage. See Section 3 for details regarding the data.

(a) Frequency of Arbitrage Opportunities



(b) Frequency vs. Distance Traveled



5.4, Panel (b) documents this observation rigorously. 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 simple statistic explains nearly all of 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 5.2, 5.3 and 5.4 suggest that the ES-SPY arbitrage opportunity should be thought of more as a mechanical "constant" of the continuous limit order book 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 5.2), but the size of arbitrage opportunities has remained remarkably constant (Figure 5.3), and the frequency of arbitrage opportunities seems to be driven mostly by market volatility (Figure 5.4). These facts both inform and are explained by our model in Section 6.

5.3 Discussion

We have shown that the continuous limit order book market design leads to frequent technical arbitrage opportunities, available to whomever is fastest, which in turn induces an arms race in speed. Moreover, the arms race does not actually compete away the prize, but rather just raises the bar for capturing it. In this section, we briefly discuss the magnitude of the prize. We make two sets of remarks.

First, we suspect that our estimate of the annual value of the ES-SPY arbitrage opportunity– an average of around $75mm per year, fluctuating as high as $151mm in 2008 (the highest volatility year in our data) and as low as $35mm 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 optimized and/or complicated 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 order book messages that are transmitted publicly to CME data feed subscribers, but we do not have access to the trade notifications that are transmitted privately only to the parties involved in a particular trade. It has recently been reported (Patterson, Strasburg and Pleven, 2013) that order-book updates lag 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 four 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 5.5 provides an illustrative partial list.[20] Two, because equity markets are fragmented –

[20]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 $100mm per security (calendar year 2011).

Figure 5.5: 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)

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, DirectEdge, 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 indirect evidence for the importance and harmfulness of this type of arbitrage is that an entire new exchange, IEX, is being launched devoted to mitigating just this one aspect of the arms race (Patterson, 2013). 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 GS-MS 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 showed in Section 4.1.2, the equities market correlation matrix breaks down at high frequency, suggesting that such trading opportunities – whether they involve pairs of stocks or statistical relationships among sets of stocks – may be important. Four, in addition to the race to snipe stale quotes, there is also a race among liquidity providers to the top of the book (cf. Farmer and Skouras (2012*a*)). 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 in the billions.

6 Model: Economic Implications of the Arms Race

We have established three empirical facts about continuous limit order book markets. First, market correlations completely break down at high-enough frequency, even for securities that are nearly perfectly correlated at longer frequencies, such as SPY and ES. Second, this correlation breakdown is associated with frequent technical arbitrage opportunities, available to whomever 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 the first two facts and seeks to make sense of the third. The model ultimately serves two related purposes: it is a critique of the continuous limit order book market design, and it identifies the economic implications of the HFT arms race.

Unfortunately, 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 Figure 5.5.

6.1 Preliminaries

Security x with perfect public signal y There is a security x that trades on a continuous limit order book market, the rules of which are described in Section 2. There is a publicly observable signal y of the value of security x. We make the following purposefully strong assumption: the fundamental value of x is *perfectly* correlated to the public signal y, and, moreover, x can always be costlessly liquidated at this fundamental value. This is a "best case" scenario for price discovery and liquidity provision in a continuous limit order book.

We think of x and y as a metaphor for pairs or sets of securities that are highly correlated. In our leading example, x is SPY and y is ES. Numerous other examples are discussed in Section 5.3. An alternative interpretation of y is as publicly observable news about the fundamental value of x. For example, y could correspond to public news coming from Fed announcements, earnings announcements, consumer confidence reports, etc.

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 (i.e., discrete) 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. To fix ideas, a simple example of a jump distribution is where the support is $\{-1, +1\}$ and positive and negative jumps are equally likely; in this case, all jumps have jump size equal to 1. Referring back to the S&P 500 arbitrage example, a jump in y can be interpreted as a discrete change in the price level of the S&P 500 futures contract in Chicago. Such jumps naturally have discrete support because futures contracts trade in units of 0.25 index points.

Players: Fundamental Investors and Market Makers There are two types of players, fundamental investors and market makers. The players we call fundamental investors could equivalently be called "liquidity traders" as in Glosten and Milgrom (1985) or "noise traders" as in Kyle (1985). The players we call market makers could equivalently be called "trading firms", "market participants" or "HFTs". Both types of players are risk neutral and there is no discounting.

Fundamental investors stochastically arrive to the market with an inelastic need to either buy or sell a unit of x. The arrival process is Poisson with rate λ_{fund}, and, conditional on arrival, it is equal probability that the fundamental trader needs to buy as opposed to sell. Payoffs for fundamental traders are defined as follows. If a trader 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 $(y_{t'} - p) - f_{delaycost}(t' - t)$, where $y_{t'}$ is the fundamental value of x at the time she trades, and the function $f_{delaycost} : \mathbb{R}^+ \to \mathbb{R}^+$ gives the cost to the fundamental trader of waiting $t' - t$ units of time to execute her trade. If

the trader arrives to market at time t needing to sell one unit, and then sells a unit at time $t' \geq t$ for price p, her payoff is $(p - y_{t'}) - f_{delaycost}(t' - t)$. We assume that the cost of delay function satisfies $f_{delaycost}(0) = 0$, and is strictly increasing and continuous. In words, all else equal, fundamental investors prefer to transact sooner than later. In the equilibrium we derive below in Section 6.2, fundamental investors choose to transact immediately. In the equilibria of frequent batch auctions, studied in Section 7, fundamental investors will choose to transact at the next available batch auction. Once a fundamental investor transacts, they exit the game.

Whereas fundamental investors arrive to market stochastically and depart once they have satisfied their demand, market makers are always present in the market. The number of market makers, N, will be governed by an equilibrium zero-profit condition. Market makers 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 market maker 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 $p - y_t$. Their objective is to maximize profits per unit time, or equivalently, total profits over the course of the trading day.

We assume that fundamental investors only use market orders (or, equivalently, immediately executable limit orders), whereas market makers use both market orders and limit orders. That is, we exogenously assume that fundamental investors act only as "takers" of liquidity, whereas market makers act as both "makers" and takers of liquidity.[21]

Signal Latency and Speed Technology The public signal y of security x's value is observable by fundamental investors and market makers with a small time delay ("signal latency"). This time delay can be interpreted as the time it takes information to travel, be processed, etc. We assume that all players can observe the signal y costlessly at delay $\delta_{slow} > 0$, meaning that the value of signal y at time t is observed at time $t + \delta_{slow}$. In addition, all players can invest in technology that allows them to observe the signal faster. We model this in a simple way: players can pay nothing and observe the signal y with delay δ_{slow}, or they can pay a cost c_{speed}, interpreted as a rental cost per unit time, and observe the signal y with delay of $\delta_{fast} < \delta_{slow}$. The cost c_{speed} is a metaphor for the cost of access to high-speed fiber optic cables (such as the Spread Networks cable described in the introduction), the cost of cutting-edge computers, the cost of the relevant human capital, etc. We assume that investment in speed is publicly observable.

[21] The assumption that fundamental investors (or noise traders) only use market orders is standard in the market microstructure literature. Our treatment of market makers as using both limit and market orders is slightly non-standard. This is because our market makers will play a role that combines aspects of what the traditional market microstructure literature calls a market maker (who use limit orders) and what the traditional literature calls an informed trader (who use market orders). This will become more clear when we describe the role market makers play in equilibrium below in Section 6.2.2.

Define $\delta = \delta_{slow} - \delta_{fast}$ as the speed difference between fast and slow players. For ease of exposition we normalize $\delta_{fast} = 0$, so $\delta = \delta_{slow}$.

We assume that all players in the market for x can submit orders and other types of messages instantaneously. That is, if any player decides to submit a message at time t, it reaches the market at exactly time t. If multiple messages reach the market at the same time, they are processed in a random order, one-at-a-time in serial. This random tie-breaking can be interpreted as messages being transmitted with small random latency, and then processed serially in the order received.[22]

6.2 Equilibrium

We construct a Nash equilibrium as follows.

6.2.1 Fundamental Investors

Fundamental 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 stationary in equilibrium, so fundamental investors have no incentive to delay trade. Fundamental investors do not choose to pay the cost c_{speed} to be faster than the competition.[23]

6.2.2 Market Makers

Market maker entry is governed by a zero-profit condition. In equilibrium, N market makers enter and pay the cost c_{speed} to be faster than the competition, and zero market makers enter but do not pay the cost. For simplicity, we allow N to take on any real value greater than or equal to 1, rather than require that N be an integer; alternatively we could require that N is integer and require that market-maker profits are weakly positive with N entrants and strictly negative with $N + 1$ entrants.

Of the N market makers, 1 plays a role we call "liquidity provider" and $N - 1$ play a role we call "stale-quote sniper".[24] Market makers will be indifferent between these two roles in equi-

[22]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 the same order to the exchange at the same time, it really is random which will be processed first. See Rogow (2012) for more details on colocation.

[23]There is nothing in our setup that prevents a fundamental trader from paying the cost c_{speed} and behaving as a market maker as described below, but there is also no particular reason for them to do so. That is, a fundamental trader who pays the cost c_{speed} and acts like a market maker can be conceptualized as two distinct entities; there is no complementarity between the two activities in our equilibrium.

[24]The term "sniper" originated in the context of eBay auctions; see Roth and Ockenfels (2002). Snipers in eBay auctions attempt to bid as *late* as possible before the auction closes. Snipers here will attempt to bid as *soon* as possible after an exploitable jump in y_t, as we will see below.

librium. 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. In practice, this sorting is stochastic, and many HFT firms perform both roles over time.[25]

Liquidity Provider The liquidity provider behaves as follows. When the trading day opens at 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 will be the opening bid and ask, respectively, and $s \geq 0$ is the bid-ask spread. We will derive the equilibrium value of s below. For simplicity, we allow s to be real-valued rather than discrete, just as we did for N. The bid-ask spread will be stationary throughout the trading day.

If the signal y jumps at time t, from y_{t^-} to y_t (we use the notation $y_{t^-} = \lim_{t' \to t^-} y_{t'}$), per the Poisson arrival process described above, the liquidity provider immediately adjusts her quotes. Specifically, at time t she submits a message to the exchange to remove 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 a fundamental trader arrives in 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 a fundamental trader 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 fundamental trader books profits of $\frac{s}{2}$. Note that the liquidity provider does not directly observe that his trading partner is a fundamental trader as opposed to another market maker, 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 fundamental trader, the liquidity provider does not take any action. Thus, at all times t, there is a single unit offered at both the bid and the ask.

Stale-Quote Snipers The $N-1$ stale-quote snipers behave as follows. Suppose that at time t the signal y jumps from y_{t^-} to y_t. If $y_t > y_{t^-} + \frac{s}{2}$, the snipers immediately submit a limit order to buy a single unit at price $y_{t^-} + \frac{s}{2}$, the ask price of the liquidity provider who, at the same time, submits a message to the exchange to remove this ask. Each sniper's bid is successful with

[25] In practice tick sizes are discrete (penny increments), whereas we allow for bids and asks to be any real value. If we used discrete ticks, then the role of liquidity provider would be strictly preferred to the role of stale-quote sniper at the equilibrium bid-ask spread. In this case, the N market makers 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.

probability $\frac{1}{N}$: there are $N-1$ snipers attempting to buy at this ask price, 1 liquidity provider attempting to remove this ask price, and the order in which the exchange processes these N messages is random.[26] If the sniper's bid is successful she books profits of $y_t - y_{t^-} - \frac{s}{2}$. If the sniper's bid is unsuccessful, she immediately withdraws her bid.[27]

Symmetrically, if $y_t < y_{t^-} - \frac{s}{2}$ the snipers immediately submit a limit order to sell a single unit at price $y_{t^-} - \frac{s}{2}$, the bid price of the market-maker who, at the same time, submits a message to the exchange to remove this bid. If the sniper's ask is successful, which occurs in equilibrium with probability $\frac{1}{N}$, then she books profits of $y_{t^-} - y_t - \frac{s}{2}$. Else, she immediately withdraws her ask.

If $y_{t^-} - \frac{s}{2} < y < y_{t^-} + \frac{s}{2}$, then the sniper does nothing. Last, if in some time interval there is no jump in the signal y, the sniper does nothing.

6.2.3 Equilibrium Bid-Ask Spread s

In equilibrium, the bid-ask spread s balances off two forces.

If, as occurs at arrival rate λ_{fund}, a fundamental trader arrives to market, the liquidity provider will earn profits of $\frac{s}{2}$, or half the bid-ask spread. The benefits of providing liquidity are thus $\lambda_{fund} \cdot \frac{s}{2}$ per unit time.

If, as occurs at arrival rate λ_{jump}, the signal y jumps, the liquidity provider will attempt to instantaneously adjust her stale quotes. However, if the jump is larger in size than $\frac{s}{2}$, the snipers simultaneously attempt to pick off her stale quotes. The liquidity provider loses this race 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 costs of providing liquidity, per unit time, are $\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N}$.

The zero-profit condition is satisfied for the liquidity provider when benefits less costs equal the rental cost of the speed technology:

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

[26] In our model, all fast market makers 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 market maker's message transmission – e.g., a uniform-random draw from $[0, \epsilon]$ – and then whichever market maker had the smallest draw from $[0, \epsilon]$ would win the race. This would yield exactly the same probability of winning the race of $\frac{1}{N}$. See also footnote 22.

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

6.2.4 Equilibrium Entry Quantity N

The equilibrium number of stale-quote snipers, $N-1$, can be determined as follows.

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 per-person expected profits, per unit time, are $\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N}$. Notice that, summed over all $N-1$ snipers, this equals the liquidity provider's cost of providing liquidity; this captures that trade amongst market makers is zero sum.

The zero-profit condition for stale quote snipers is satisfied when the benefits of sniping equal the rental cost of the speed technology:

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

6.2.5 Solving for s and N

Equations (6.1) and (6.2) together constitute two equations in two unknowns, N and s. Adding (6.1) and $N-1$ times (6.2) yields

$$\lambda_{fund} \cdot \frac{s}{2} = N c_{speed} \tag{6.3}$$

Equation (6.3) has a natural economic interpretation. The right-hand side is the total expenditure by market makers on speed. The left-hand side is the total revenue earned by the liquidity provider from providing liquidity to fundamental investors. Since stale-quote sniping is a zero-sum activity amongst market makers, this in turn is equal to the total *profits* earned by market makers as a whole from providing liquidity to fundamental investors. The equation thus tells us that all of the expenditure by market makers on speed technology ultimately is borne by fundamental investors, via the bid-ask spread.

If we multiply (6.2) by N and substitute in (6.3) we obtain a single equation with a single unknown, s:

$$\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) = \lambda_{fund} \cdot \frac{s}{2} \tag{6.4}$$

The left-hand side of (6.4) is strictly positive when s is zero, and then is strictly decreasing in s until its value is zero when $\frac{s}{2}$ is equal to the upper bound of the jump size distribution (i.e., when $\frac{s}{2} = \max J$). The right-hand side of (6.4) has value zero at $s = 0$ and then is strictly increasing in s. Hence, (6.4) has a unique solution. Plugging this unique solution for s into (6.3) then gives a unique solution for N.[28]

[28] While s and N are uniquely characterized, the sorting of market makers into roles is not. In particular, there are equilibria in which (i) it is deterministic which market maker serves as liquidity provider and which serve as stale-quote snipers; (ii) market makers stochastically sort into the two roles, e.g., by racing to perform the role of liquidity provider, with losers of the race performing the role of stale-quote sniper; (iii) market makers rotate who

We summarize with the following proposition.

Proposition 1 (Equilibrium)**.** *There is a Nash equilibrium of the continuous limit order book market design with fundamental investor play as described in Section 6.2.1 and market maker play as described in Section 6.2.2. The amount of market maker entry N^* and the equilibrium bid-ask spread s^* are uniquely determined by the market maker zero profit conditions* (6.1) *and* (6.2)*. The sorting of market makers into the roles of 1 liquidity provider and $N^* - 1$ stale-quote snipers is not unique.*

Per (6.3)-(6.4), the following three quantities are equivalent in equilibrium:

1. *The total prize at stake in the arms race, $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$. That is, the sum of the value of all arbitrage opportunities that the snipers are racing to capture (and that the liquidity provider is racing to avoid being captured).*

2. *The total equilibrium expenditure by market makers on speed technology, $N^* c_{speed}$.*

3. *The total revenue the liquidity provider earns from fundamental investors via the bid-ask spread, $\lambda_{fund} \cdot \frac{s^*}{2}$.*

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

6.3 Discussion of the Equilibrium

6.3.1 Why is there a Positive Bid-Ask Spread?

Given the setup of our model, one might have guessed that Bertrand competition among market makers drives the bid-ask spread to zero. There is not an asymmetrically informed "insider" as in the models of Kyle (1985) or Glosten and Milgrom (1985); instead, all market makers 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 Roll (1984) 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. Yet, the equilibrium bid-ask spread s^* is strictly positive.

Our model highlights that the continuous limit order book market design creates an additional, purely technical cost of liquidity provision – the cost of getting sniped, i.e., of getting picked off in the race to react to public news. Since the continuous limit order book processes message

performs the role liquidity provider; and (iv) versions of the deterministic, stochastic, and rotation equilibria in which the liquidity provider role is split into two sub-roles, one of which provides liquidity at the bid and the other of which provides liquidity at the ask.

requests one-at-a-time in serial, a liquidity provider's quotes are vulnerable to being picked off if they become stale, *even if the liquidity provider learns at exactly the same time as other market participants that his quotes are now stale.* All the liquidity provider can do is send a message to the exchange to remove his stale quotes, knowing full well that at the same time other market makers are sending messages to the exchange attempting to exploit his stale quotes. It is random which of this barrage of messages will get processed first, and with probability $\frac{N^*-1}{N^*}$, it will not be the liquidity provider's message that is first, and he will get sniped.

Mechanically, our source of bid-ask spread is most similar to that in Glosten and Milgrom (1985), namely, a liquidity provider sometimes gets exploited by another player who knows that the liquidity provider's quote is mispriced.[29] The key difference is that in Glosten and Milgrom (1985) there is asymmetric information between the liquidity provider and this other player, whom Glosten and Milgrom (1985) call an "informed trader", whereas in our model the liquidity provider and these other players, the stale-quote snipers, are symmetrically informed. Both the liquidity provider and the stale-quote snipers observe the innovation in y at *exactly* the same time, but, because the continuous limit order book processes message requests one-at-a-time in serial, the liquidity provider's request to withdraw his quote may get processed after some stale-quote sniper's request to accept his quote. A potentially useful way to state the relationship is that our model shows that Glosten-Milgrom adverse selection is "built in" to the continuous limit order book market design, even in the absence of asymmetric information. We describe this source of bid-ask spread as technical as opposed to fundamental since it is caused by the market design and can be eliminated by modifying the market design.

The difference between our source of bid-ask spread and that in Glosten and Milgrom (1985) is further reinforced by considering the limiting cases of $\delta \to 0^+$ or $c_{speed} \to 0^+$.[30] In our model, there is zero asymmetric information among the N market makers who pay the cost c_{speed} to be fast, and among players more widely the only source of asymmetric information is that some players observe the signal y_t with tiny delay δ. In the limit as $\delta \to 0^+$, all players observe the signal y_t at the same time and hence all players are symmetrically informed. In the limit as $c_{speed} \to 0^+$, the equilibrium quantity of fast market makers goes to infinity, and hence so too does the number of market makers who are symmetrically informed. Yet, there is nevertheless a strictly positive bid-ask spread in equilibrium of our model even in these limiting cases of $\delta \to 0^+$ or $c_{speed} \to 0^+$, due to sniping costs.

We summarize this discussion as follows.

[29] See also Copeland and Galai (1983) and Foucault (1999) for slightly different modeling approaches that yield the same source of bid-ask spread as in Glosten and Milgrom (1985).

[30] We thank Pete Kyle for this observation.

Proposition 2 (Positive Bid-Ask Spread)**.** *In our model there are no inventory costs (Roll, 1984), search costs (Duffie, Garleanu and Pedersen, 2005), or information asymmetries (Kyle, 1985; Glosten and Milgrom, 1985) between liquidity providers and stale-quote snipers. Nevertheless, the equilibrium bid-ask spread s^* is strictly positive. The bid-ask spread is strictly positive even in the limiting cases of $\delta \to 0^+$ (speed advantages are arbitrarily small) and $c_{speed} \to 0^+$ (speed costs are arbitrarily small).*

We wish to clarify the relationship between our result and the clear empirical evidence that bid-ask spreads are *narrower* today than in the pre-HFT era. 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 (e.g., algorithmic trading), and the speed race. Our interpretation of the empirical record is that there is considerable evidence that IT has improved bid-ask spreads – see especially Hendershott, Jones and Menkveld (2011) and the discussion in Section 4 of Jones (2013)) – which makes intuitive economic sense, as IT has lowered costs in numerous sectors throughout the economy. However, there is no empirical evidence for the proposition that the speed race per se has improved bid-ask spreads, and some recent evidence that suggests that the speed race and associated sniping widens the bid-ask spread (Foucault, Kozhan and Tham, 2013), which is consistent with our result. Our result does not imply that bid-ask spreads should be wider today than in the pre-HFT era (we are not Luddites nostalgic for 1990s information technology or market structure). Our result says that bid-ask spreads are unnecessarily wide today, i.e., they could be narrower under an alternate market design.

6.3.2 Comparative Statics of the Bid-Ask Spread

Equation (6.4) yields the following comparative statics for our source of bid-ask spread:

Proposition 3 (Comparative Statics of the Bid-Ask Spread)**.** *The equilibrium bid-ask spread s^* has the following comparative statics:*

1. *s^* is strictly decreasing in the frequency of fundamental investor demand, λ_{fund}*
2. *s^* is strictly increasing in the frequency of jumps, λ_{jump}*
3. *If jump distribution F'_{jump} is a mean-preserving spread of F_{jump}, then s^* is strictly larger under F'_{jump} than F_{jump}.*

Heuristically, s^* is widest for securities that are thinly traded (low λ_{fund}) and that are correlated to statistics that have frequent and large jumps (high λ_{jump} and high-variance F_{jump}). Examples include thinly traded small-cap stocks that are highly correlated to a small-cap stock index such as

the Russell 2000, and entrant ETFs that are thinly traded and highly correlated to an incumbent ETF.[31]

In light of the results below in Section 7, a policy implication of Proposition 3 is that the benefits of batching are especially large for such securities.

6.3.3 The Bid-Ask Spread and Arms-Race Prize Does *Not* Depend on c_{speed} and δ

It is also interesting to observe some parameters that the equilibrium bid-ask spread s^* does *not* depend on, namely, c_{speed} and δ (cf. equation (6.4)). This can be interpreted as follows. Suppose that speed technology improves each year, and we reinterpret the model so that c_{speed} is the cost of being at the cutting edge of speed technology in the current time period, and δ is the speed advantage versus other traders in the current time period. Then, each year high-frequency traders get faster, but the bid-ask spread stays the same, as does the total prize associated with the arms race, $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$.[32]

This discussion helps make sense of our findings in Section 5.2.1 on the time series evolution of the ES-SPY technical arbitrage opportunity. We found that the duration of arbitrage opportunities declined steadily from 2005-2011, but that the total pie that high frequency traders compete for has been roughly constant, fluctuating with market volatility but not exhibiting a time trend per se.

Proposition 4 (Arms Race Prize is a Constant)**.** *The equilibrium bid-ask spread, s^*, and the total prize associated with the arms race, $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$, are invariant to both the cost of speed, c_{speed}, and the magnitude of speed differences, δ $(= \delta_{slow} - \delta_{fast})$.*

Together, Proposition 4 and the empirical evidence in Section 5.2.1 suggest that the arms race is best understood as a "constant" of the continuous limit order book market design rather than as an inefficiency that is competed away over time.

6.3.4 Welfare Costs of the Arms Race: a Prisoner's Dilemma amongst Market Makers

In the equilibrium derived above market makers earn zero profits, as they simply cover their costs of speed technology. All of these expenditures on speed technology are in turn borne by

[31] For example, Vanguard ETFs initially had bid-ask spreads that were noticeably wider than incumbent ETFs for similar indices.

[32] Per (6.2), this total prize is equivalent to $N^* c_{speed}$, but it nevertheless is still invariant to c_{speed}: if c_{speed} is low, then N^* is commensurately high, and vice versa, so that in equilibrium $N^* c_{speed}$ does not vary with c_{speed} (nor with δ of course).

fundamental investors, via the bid-ask spread (cf. (6.3)). It is easy to see that this arrangement is socially inefficient – even if fundamental investors are extremely impatient.

Formally, suppose that we hold fixed the number of market makers at the equilibrium level of N^*, but eliminate the opportunity to invest in speed technology. Given this setup, there is an equilibrium that is essentially identical to that described above. The bid-ask spread is s^*, just as before, and the N^* market makers sort into 1 liquidity-provider and $N^* - 1$ stale-quote snipers, just as before. The only difference is that now all market makers – both the liquidity provider and the snipers – respond to changes in y with delay of δ. In this equilibrium, fundamental investors *still get to trade immediately*, and still pay the same bid-ask spread cost of $\frac{s}{2}$. So, the welfare of fundamental investors is unchanged. The welfare of the N^* market makers is strictly greater though, by c_{speed} per unit time.

Hence, the decision by market makers to invest in speed can be interpreted as a prisoner's dilemma.[33] The N^* market makers would each be better off if they could collectively commit not to invest in speed. But, each individual market maker has incentive to deviate and invest in speed, which ultimately results in each of them earning zero profits in equilibrium.

Proposition 5 (Prisoner's Dilemma)**.** *Social welfare would be higher by $N^* \cdot c_{speed}$ if the market makers could commit not to invest in speed technology, with these gains shared equally among the N^* market makers. But, each individual market maker has a dominant strategy incentive to 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 fundamental investors instead of market makers.

6.3.5 Relationship to the Efficient Markets Hypothesis

It is interesting to interpret the equilibrium derived above as it relates to the efficient markets hypothesis.

[33] Biais, Foucault and Moinas (2013) make a conceptually similar point in the context of an abstract rational expectations model in the style of Grossman and Stiglitz (1980). In their model, there is a single asset whose common value component has a mean of μ and an idiosyncratic shock of either $+\epsilon$ or $-\epsilon$, with equal probability; there is also a private value component, which creates a reason to trade. Investors can pay a cost C to learn the idiosyncratic shock before they engage in a single period of trading. Paying the cost also gives the investor a higher probability of finding a trading opportunity in this single period of trading. Biais, Foucault and Moinas (2013)'s key observation is that one investor's paying the cost C generates negative externalities for other investors, due to adverse selection, which can in turn create a reason for the other investors to also pay the cost C. This can lead to inefficient overinvestment. Biais, Foucault and Moinas (2013) interpret this finding as equilibrium overinvestment in speed, though one could interpret the result more broadly as equilibrium overinvestment in any source of informational advantage.

On the one hand, the market is highly efficient in the sense of instantaneously incorporating news about the value of x into prices. Formally, the measure of times $t \in [0, T]$ where the bid-ask spread for x does not contain the fundamental value y is zero.

On the other hand, a non-zero volume of trade is conducted at stale prices. Specifically, the proportion of trade that is conducted at quotes that do not contain the fundamental value y is $\frac{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*}}{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*} + \lambda_{fund}}$.

Hence, the market is highly efficient in *time space* but less so in *volume space*: a lot of volume gets transacted at incorrect prices. This volume is in turn associated with rents from public information about other securities' prices, in violation of the weak-form efficient markets hypothesis (cf. Fama, 1970).[34] 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.

Proposition 6 (Market Efficiency in Time Space but not Volume Space)**.** *In equilibrium, the midpoint of the bid-ask spread is equal to the fundamental value y_t with probability one. Nevertheless, a strictly positive proportion of trade, $\frac{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*}}{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*} + \lambda_{fund}}$, is conducted at quotes that do not contain the fundamental value y_t between the bid and the ask. These trades generate arbitrage profits for the liquidity-taking side of the trade (and commensurate losses for the liquidity-providing side of the trade), in violation of the weak-form efficient markets hypothesis. However, there is no free lunch, since in equilibrium fast traders' economic profits are zero.*

6.3.6 Canceled Orders

We briefly note that our model provides two natural explanations for the phenomenon of "canceled orders", which are limit orders that are withdrawn from the book before they are filled (see also Baruch and Glosten, 2013). First, the liquidity provider cancels his current limit orders whenever the fundamental value y_t jumps, even if the jump is tiny. Such cancellations occur at a rate of approximately $2\lambda_{jump}$, where the 2 is because the liquidity provider cancels both the bid and the ask, and the approximation is that if there is a large jump, with probability $\frac{N^*-1}{N^*}$, he will only be able to successfully cancel one of his two limit orders. Second, every time a stale-quote sniper attempts to pick off a stale quote but fails, he will cancel the associated order. Such cancellations occur at rate $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot (N^* - 2 + \frac{1}{N^*})$.

By contrast, limit orders are accepted by fundamental investors at rate λ_{fund}, and by stale-quote snipers at rate $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*}$. Some algebra yields the following result.

[34] 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.

Proposition 7 (Canceled Orders)**.** *The number of canceled limit orders, as a proportion of all submitted limit orders, is given by*

$$\frac{\lambda_{jump} \cdot (2 + \Pr(J > \frac{s^*}{2}) \cdot (N^* - 3 + \frac{2}{N^*}))}{\lambda_{jump} \cdot (2 + \Pr(J > \frac{s^*}{2}) \cdot (N^* - 2 + \frac{1}{N^*}) + \lambda_{fund}} \tag{6.5}$$

For N^* large this is approximately

$$\frac{\lambda_{jump} \cdot (2 + \Pr(J > \frac{s^*}{2}) \cdot N^*)}{\lambda_{jump} \cdot (2 + \Pr(J > \frac{s^*}{2}) \cdot N^*) + \lambda_{fund}}$$

Notice that the proportion of canceled orders is especially high for securities with low λ_{fund} relative to λ_{jump}.

Proposition 7 suggests that the high message-to-trade ratios that have been a source of concern to policy makers may simply be an equilibrium consequence of the continuous limit order book market design.

6.4 Market Thinness

Our model highlights that continuous limit order book markets create a purely technical cost of providing liquidity: liquidity providers get sniped if their quotes become stale. This occurs even if these quotes become stale in a way that the liquidity provider understands just as well as all other market participants, because the liquidity provider might lose the race to react – his message to withdraw his stale quotes might be processed after another market maker's message to exploit these same stale quotes.

In this section we highlight that this cost of providing liquidity grows linearly with the quantity of liquidity provided, whereas the benefits of providing liquidity grow less than linearly. As a result, not only is there a positive bid-ask spread for the first quoted unit in our model, but the quoted spread grows wider with quantity. That is, sniping costs cause continuous limit order book markets to be unnecessarily thin.

Consider the model of Section 6.1 but modified so that fundamental investors sometimes need to buy or sell more than 1 unit. Specifically, fundamental investors arrive to market at rate λ_{fund} as before, but now they need to transact a quantity $q \in \{1, \ldots, \bar{q}\}$, with $p_1 > 0$ the probability that they need to transact 1 unit, $p_2 > 0$ the probability that they need to transact two units, $\ldots$, $p_{\bar{q}} > 0$ the probability that they need to transact $\bar{q}$ units. As before, fundamental investors are equally divided between those needing to buy or sell, and this is orthogonal to the quantity required.

Above, we assumed that fundamental investors transact only in market orders. Here, we make

a stronger assumption, which is that fundamental investors transact in a *single* market order, i.e., a fundamental trader who needs to transact k units does so in a single market order with quantity k. We emphasize that such behavior is not optimal under the continuous limit order book market: a fundamental investor with multi-unit demand will prefer to split his order into several smaller orders (analogously to Kyle (1985); Vayanos (1999)). Instead, we view this assumption as allowing us to illustrate a mechanical point about continuous limit order book markets, which is that it is costly to provide a deep book.

There is an equilibrium of this model similar to that in Section 6.2, in which the market makers serve as both liquidity providers and stale-quote snipers, and are indifferent between the two roles in equilibrium. As above, we can assign the roles among the market makers in an arbitrary fashion. For expositional simplicity, we adopt the convention that market maker 1 serves as the lone liquidity provider, providing all $\bar{q}$ units of depth on each side of the market, with the other market makers serving as stale-quote snipers. However we note that a more realistic approach would be to have each market maker serve partly as liquidity provider and partly as stale-quote sniper: since there are now $2\bar{q}$ limit orders present in the book at any given instance, there is plenty of room for several market makers to split up the role of liquidity provider. Each such market maker will want to snipe any stale quotes that are not his own.

In equilibrium, the liquidity provider does not offer all $\bar{q}$ units of liquidity at the same bid-ask spread, but instead offers a first unit of liquidity at a spread s_1, a second unit of liquidity with a strictly wider spread $s_2 > s_1$, a third unit of liquidity with a wider spread still of $s_3 > s_2$, etc. The spread for the kth unit of liquidity, s_k, is governed by indifference between liquidity provision (LHS) and stale-quote sniping (RHS) at the kth level of the book:

$$\lambda_{fund} \cdot \sum_{i=k}^{\bar{q}} p_i \cdot \frac{s_k}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s_k}{2}) \cdot \mathbb{E}(J - \frac{s_k}{2} | J > \frac{s_k}{2}) \cdot \frac{N-1}{N} \tag{6.6}$$

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

The LHS of (6.6) represents the benefits less costs of liquidity provision in the kth level of the book. Notice that the second term on the LHS of (6.6), which describes the costs of getting sniped, is exactly the same as the second term on the LHS of (6.1). This is because, if a quote becomes stale, stale-quote snipers will attempt to pick off the liquidity provider for as much quantity as is available at an advantageous price. Similarly, the RHS of (6.6), which represents the benefits of sniping the kth level of the book, is exactly the same as the LHS of (6.2).

By contrast, except for the case of $k = 1$, the first term on the LHS of (6.6), which describes the benefits of providing liquidity, is strictly smaller than the first term on the LHS of (6.6). This

is because only proportion $\sum_{i=k}^{\bar{q}} p_i$ of investors trade the kth level of the order book.

Intuitively, the benefits of providing liquidity scale sub-linearly with the quantity offered (only some fundamental investors require a large quantity), whereas the costs of providing liquidity scale linearly with the quantity offered (snipers will exploit stale quotes in the full quantity offered).[35]

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 8 (Market Thinness)**.** *There exists a Nash equilibrium of the multi-unit demand model analogous to the Nash equilibrium of the single-unit demand model. In this equilibrium, the liquidity provider offers a single unit at bid-ask spread s_1^*, a single unit at bid-ask spread s_2^*, ..., a single unit at bid-ask spread $s_{\bar{q}}^*$, with spreads uniquely characterized by (6.6). Spreads are strictly increasing,*

$$s_1^* < s_2^* < \cdots < s_{\bar{q}}^*$$

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

The other comparative statics on bid-ask spreads are as follows. As in Proposition 3, bid-ask spreads are wider, at all levels of the book, for securities with low λ_{fund} and high λ_{jump}, and under mean-preserving spreads of F_{jump}. Additionally, bid-ask spreads are wider at the k^{th} level of the book the rarer are orders of at least size k, that is, the lower is $\sum_k^{\bar{q}} p_k$. As in Section 6.3.3, bid-ask spreads do not depend on c_{speed} or δ.

Thus, not only is there a positive bid-ask spread in our model – even in the absence of asymmetric information, inventory costs, etc. – but markets are unnecessarily thin too.

7 Frequent Batch Auctions as a Market Design Response

We propose frequent batch auctions as a market design alternative to continuous limit order books. Section 7.1 defines frequent batch auctions. Section 7.2 shows why batching eliminates the HFT arms race. Section 7.3 studies the equilibria of frequent batch auctions, and shows that batching leads to narrower spreads, deeper markets and increased social welfare. Section 7.4 makes several remarks concerning the equilibrium analysis.

[35] 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).

7.1 Frequent Batch Auctions: Definition

Informally, frequent batch auctions are uniform-price sealed-bid double auctions conducted at frequent but discrete time intervals, e.g., every 1 second. In this section we define frequent batch auctions formally.

The trading day is divided into equal-length discrete 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 may submit offers to buy and sell shares of stock in the form of limit orders and market orders. Just as in the continuous limit order book, a limit order is simply 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. Market orders are interpreted as limit orders with the maximum allowable bid or minimum allowable ask, both of which are assumed to be finite. In practice, price circuit breakers would determine what constitutes these maximum and minimum amounts (e.g., the price in the previous batch auction plus or minus some specified percentage). A single trader may submit multiple orders, which can be interpreted as submitting a demand function or a supply function (or both). Traders may withdraw or adjust their orders at any time during the batch interval. Orders are not visible to other market participants during the batch interval, i.e., the auction is "sealed bid", as described below. Instead, orders are announced publicly *after* the auction is conducted.

At the conclusion of each batch interval, the exchange collates all of the received orders (i.e., it "batches" the received orders), and computes the aggregate demand and supply functions out of all bids and asks, respectively. The market clears where supply equals demand, with all transactions occurring at the same price (i.e., at a "uniform price").[36] There are three possible cases to consider for market clearing. In Case 1, supply and demand intersect horizontally or at a point, which pins down a unique price p^* and a maximum possible quantity q^*. In this case, offers to buy with bids strictly greater than p^* and offers to sell with asks strictly less than p^* transact their full quantity, at price p^*, whereas for bids and asks of exactly p^* it may be necessary to ration one side of the market to enable market clearing (see Figure 7.1 for an illustration).[37,38] In Case 2, supply and

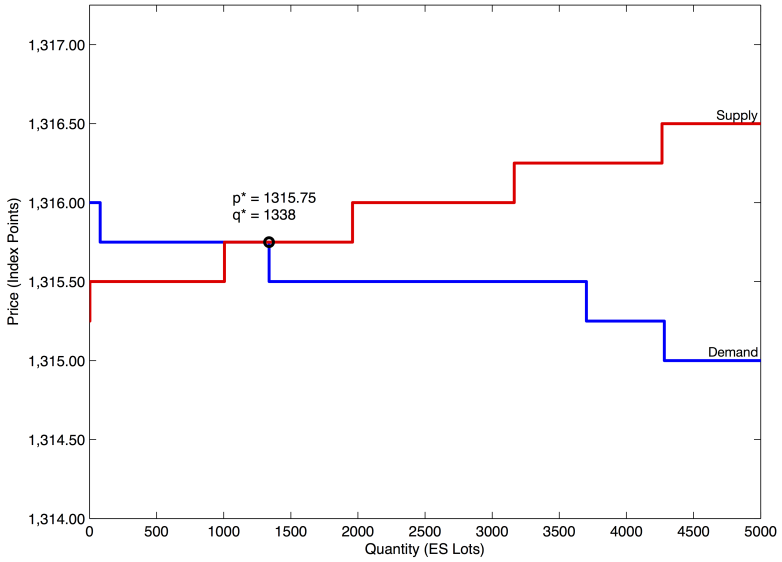
[36] Uniform-price auctions were originally proposed by Milton Friedman in the 1960s, for the sale of US Treasury bonds (Friedman, 1960).

[37] A simple rationing rule for use in practice would be to fill orders at price p^* from earlier batch intervals first and then ration pro-rata within the last batch interval filled. This encourages traders to let orders stand for longer periods, improving market depth, but without introducing a speed race. Time priority is only relevant across batch intervals, not within a batch interval.

[38] A reason to favor fine rather than coarse tick sizes is to reduce the likelihood of ties and hence the amount of rationing. Fine tick sizes also allow for more accurate preference expression. However, a tick size that is too small

Figure 7.1: Illustration of Batch Auctions

Notes: This figure illustrates batch auctions. Individual bids and asks are batched at the end of the batching interval to induce aggregate demand and supply curves. The aggregate demand and supply curves are step functions because bids and asks are for a discrete quantity at a discrete price. The market then clears where supply equals demand. If supply and demand do not intersect (the lowest ask is greater than the highest bid) then there is no trade. The example in the figure depicts illustrative supply and demand curves based on one second of order book activity in the market for ES, 9:59:28.000 to 9:59:28.999 on 2/4/2009. In the example depicted in the figure, the market clearing price is 1315.75 and the market clearing quantity is 1338 contracts. It is possible to satisfy all demand with bids weakly greater than 1315.75 and all supply with asks strictly less than 1315.75. Asks of exactly 1315.75 are rationed. This corresponds to Case 1 as described in Section 7.1; for more details, see the text.



demand intersect vertically, pinning 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 $\frac{p_L^*+p_H^*}{2}$. Finally, in Case 3, supply and demand do not intersect and the outcome is no trade.

As noted above, orders are not visible to other market participants during the batch interval. This is important to prevent gaming.[39] Instead, the exchange announces the aggregate supply and demand functions at the conclusion of each batch interval. We view this information disclosure policy as analogous to current practice under the continuous limit order book market design, under which new bids, asks, adjustments, withdrawals, etc., first are processed by the exchange, and then the updated state of the limit order book is announced publicly.

may result in needless gaming and computation to improve bids and asks by economically negligible amounts, just as in the continuous market.

[39]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.

7.2 Why and How Frequent Batch Auctions Eliminate the Arms Race

There are two reasons why frequent batch auctions eliminate (or at least substantially reduce) the high-frequency trading arms race.

First, and most centrally, frequent batch auctions reduce the value of a tiny speed advantage. Consider a situation with two market makers, one of whom is slow and observes y_t with lag δ_{slow}, and one of whom is fast and observes y_t with lag δ_{fast}. Suppose the slow market maker attempts to provide liquidity to fundamental investors, that is, to serve a role analogous to the liquidity provider in Section 6.2. A slow market maker acting as liquidity provider is vulnerable to being sniped by the fast trader if his quotes become stale. But, whereas in the continuous limit order book market he would be vulnerable to being sniped by the fast trader for *all* jumps in y, here he is only vulnerable to being sniped for jumps in y that occur at a very specific time in the batch interval. The only circumstance under which there is a jump in y that the fast trader observes but that the slow trader does not observe in time for the next batch auction is if the jump occurs in a window of time of length $\delta = \delta_{slow} - \delta_{fast}$, taking place from $(\tau - \delta_{slow}, \tau - \delta_{fast}]$. Any jumps in y that occur during the window $[0, \tau - \delta_{slow}]$ are observed by both the slow and the fast trader before they must finalize their bids for the next batch auction. Similarly, any jumps in y that occur during the window $(\tau - \delta_{fast}, \tau]$ are observed by neither the fast nor the slow trader in time for the auction at τ (both will have this information for the next auction). It is only jumps in the window $(\tau - \delta_{slow}, \tau - \delta_{fast}]$ that create asymmetric information, where the fast trader knows something about y that the slow trader does not. Hence, the proportion of the trading day during which jumps in y leave a slow liquidity provider vulnerable to being sniped is $\frac{\delta}{\tau}$, which goes to zero as τ grows large. See Figure 7.2 for an illustration. By similar reasoning, the proportion of the trading day during which jumps in y leave a *fast* liquidity provider vulnerable to being sniped is zero in our model.[40]

Second, and more subtly, frequent batch auctions change the nature of competition when there are multiple fast traders: market makers compete on price not speed. To illustrate, suppose as in the previous paragraph that there is one slow trader trying to provide liquidity to fundamental

[40]That fast traders are *never* sniped is an artifact of our stylized latency model. But, consider as well the following more realistic latency model, which will lead to a substantively similar conclusion. Fast traders 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 fast traders in response to any particular signal. Now, a fast trader 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 fast trader's 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 would be measured in milliseconds (e.g., 3 milliseconds in the Spread Networks example mentioned in the introduction), the parameter ϵ would in practice be measured in microseconds (millionths of a second). Hence, even for short batch intervals, the proportion $\frac{\epsilon}{2\tau}$ is very small. For example, if ϵ is 100 microseconds and τ is 1 second, then $\frac{\epsilon}{2\tau} = 0.00005$.

Figure 7.2: 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.

investors, but that now there are $N \geq 2$ fast traders interested in exploiting the liquidity provider's stale quotes. Suppose that there is a jump in y_t during the critical interval $(\tau - \delta_{slow}, \tau - \delta_{fast}]$ where the fast traders see the jump but the slow traders do not. Concretely, suppose that the jump is from y to y', with $y' > y + \frac{s}{2}$, where s is the liquidity provider's hypothetical bid-ask spread for a unit of x. The slow trader's quote for x is now stale: his ask price of $y + \frac{s}{2}$ is strictly lower than the new value y'. But, consider what happens when multiple fast traders attempt to exploit this stale quote. In the continuous limit order book market, when multiple fast traders attempt to exploit a stale quote, the exchange processes whichever trader's order happens to reach the exchange first. (In our model in Section 6, all orders reach the exchange at exactly the same time, and then the exchange processes them in a random order.) In the batch auction, so long as all of the orders reach the exchange by the end of the batch interval, the market processes all of the orders simultaneously – *in batch, not serial* – in its determination of the market-clearing price. But, this means that competition among fast traders drives the price of x up to the new correct level of y'. At any hypothetical market-clearing price $p < y'$, each fast trader strictly prefers to deviate and bid a tiny amount more, so the only Nash equilibrium is for the fast traders to all bid y'. In the continuous limit order book, competition drives fast traders to be ever so slightly faster than the competition, so that they can be first to accept the stale quote at $y + \frac{s}{2}$. In the batch auction, competition simply drives the price up to to the correct level of y'.

Another way to put this second point about the nature of competition under batch auctions is as follows. In the continuous limit order book market, fast traders can earn a rent for information that is widely available to other market participants – e.g., changes in the price of ES which affect the value of SPY – so long as they observe and act on the information ever so slightly faster than the other fast traders (cf. Hirshleifer, 1971). In the continuous market, *someone is always first.* In the batch auction, traders can only earn a rent from information that only they have access to – more precisely, information that they develop in time for the end of the batch interval, and

that no other traders have by the end of the batch interval. With batch intervals of, say, 1 second, there is still plenty of scope for market participants to develop genuinely asymmetric information about security values, for which they will earn a rent. But, batching eliminates paying a rent for trivial information that many market participants observe at basically the same time.[41]

We summarize this discussion as follows:

Proposition 9 (Batching Eliminates Sniping)**.** *Consider a frequent batch auction set in the model of Section 6.1.*

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 $(y_{\tau-\delta_{slow}}, y_{\tau-\delta_{fast}}]$ 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 $y_{\tau-\delta_{fast}}$. As a result, the liquidity provider does not lose money from the stale quote.*

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 he 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*

[41] This discussion relates to several recent controversies regarding the timed release of market-moving data (e.g., Fed announcements, consumer confidence reports, jobs reports, etc.). To illustrate, consider the Federal Reserve FOMC's "no taper" announcement issued on 9/18/2013 at 2:00:00.000pm. The public debate in the aftermath of this announcement concerned whether the reaction by trading algorithms to this announcement was as fast as legally possible or faster than legally possible – since the news originated in DC, and it takes information around 5 milliseconds to travel from DC to Chicago, there should not have been a reaction to the news in Chicago until 2:00:00.005pm, but there was picking-off activity sooner than that (Nanex, 2013*a*). Our point is that, whether or not the reaction was legal, this kind of public information, observable to many market participants at exactly the same time, should not earn a rent. If the next trading opportunity were a batch auction conducted at 2:00:01.000pm, the auction would have discovered a price that reflected the public "no taper" information, without any picking-off rents.

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 liquidity provider loses money from the stale quote.

7.3 Equilibrium of Frequent Batch Auctions

The discussion in Section 7.2 showed that frequent batch auctions eliminate (or at least substantially reduce) the HFT arms race, both 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 social welfare. We study the equilibria of frequent batch auctions for three cases. In the first case, the number of fast market makers is exogenous. In the second case, entry is endogenous and the batching interval is short enough that equilibrium still involves a fast liquidity provider. In the third case, entry is endogenous and the batching interval is long enough that liquidity is provided by slow market makers. We discuss the relationship among these equilibria and make some clarifying remarks in Section 7.4.

7.3.1 Model

We study the frequent batch auctions using the model of Section 6.1 that we used to study the continuous limit order book, with one modification. In the model of Section 6.1, fundamental investors arrive according to a Poisson process with arrival rate λ_{fund}. In the context of the continuous limit order book market, the Poisson process makes an implicit finiteness assumption, because the probability that more than one fundamental investor arrives at any instant is zero. Here, we need to make an explicit finiteness assumption. Specifically, we assume that fundamental investors continue to arrive according to a Poisson process, and continue to be equally likely to need to buy or sell a unit, but we assume that the net demand of fundamental investors in any batch interval – number who need to buy less number who need to sell – is bounded. Formally, let $A(\tau)$ denote the random variable describing the number of fundamental investors who arrive in a τ batch interval, and let $D(\tau)$ denote the random variable describing their net demand. We assume that there exists a $\bar{Q} < \infty$ such that the support 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, $\sqrt{\tau\lambda_{fund}}$.[42]

[42]In practice, frequent batch auctions would likely include both price and volume circuit breakers. $\bar{Q}$ can be interpreted as the maximum volume that can be traded in any one period without triggering the volume circuit breaker.

7.3.2 Exogenous Number of Fast Market Makers

We begin by considering the case where the number of fast market makers is exogenously fixed at $N \geq 2$. More precisely, there are $N \geq 2$ market makers who are exogenously constrained to pay the cost c_{speed}, and hence regard the cost as sunk. An interpretation is that this case represents the transition from continuous limit order books to frequent batch auctions – the N fast market makers are those who have invested in speed under the continuous limit order book design.

As discussed in Section 7.2, fast market makers are invulnerable to sniping in our model (cf. footnote 40). Hence, their variable cost of providing liquidity is zero, and there exists an equilibrium in which each fast market maker offers the maximum necessary depth, $\bar{Q}$, at zero bid-ask spread.

Proposition 10 (Equilibrium of Frequent Batch Auctions with Exogenous Number of Market Makers)**.** *Suppose that there are $N \geq 2$ fast market makers exogenously present in the market. Then there exists a Nash equilibrium in which each fast market maker acts as a liquidity provider, offering depth of $\bar{Q}$ at zero bid-ask spread. As compared to the equilibrium of the continuous limit order book market, the effects of batching in this equilibrium are as follows:*

1. *The bid-ask spread for the first-quoted unit is narrower: it is 0 instead of $\frac{N^* \cdot c_{speed}}{\lambda_{fund}}$.*
2. *The market is deeper: the order book has depth of $\bar{Q}$ at zero spread, whereas in the baseline model of the continuous limit order book just a single unit is offered in the order book, and in the extended model considered in Section 6.4 the bid-ask spread grows wider with the quantity traded.*

Notice that in this equilibrium fast market makers do not recoup c_{speed}. This suggests that there will be only 0 or 1 fast market makers once we allow for endogenous entry.

7.3.3 Endogenous Entry: Short Batch Intervals

We now consider endogenous entry into speed. In this section we seek an equilibrium in which there is one fast market maker who serves as liquidity provider and zero other fast market makers. We will show that such an equilibrium exists provided that the batch interval τ is small.

In this equilibrium, a single market maker pays c_{speed} and serves as a liquidity provider to fundamental investors. His role is analogous to that in the equilibrium of Section 6.2, with two key differences. First, he no longer has to worry about getting sniped. Second, while in Section 6.2 the liquidity provider could service all fundamental investor demand by maintaining a limit order book of depth one – whenever a fundamental investor arrived to market and accepted the bid or the ask, the liquidity provider immediately replenished the bid or the ask – here he will have

to provide a deeper book in order to service all fundamental investor demand. Specifically, let s_1 represent the bid-ask spread charged if the absolute value of net demand, $|D|$, is equal to 1; this corresponds to an ask for a single unit at price $y_{\tau-\delta_{fast}} + \frac{s_1}{2}$ and a bid for a single unit at a price $y_{\tau-\delta_{fast}} - \frac{s_1}{2}$, where $y_{\tau-\delta_{fast}}$ is the value of the signal y as perceived by the fast trader as of the end of the batching interval τ. Let $s_2 \geq s_1$ represent the bid-ask spread charged if $|D| = 2$, which corresponds to an ask for a single unit at price $y_{\tau-\delta_{fast}} + \frac{s_2}{2}$ and a bid for a single unit at a price $y_{\tau-\delta_{fast}} - \frac{s_2}{2}$; this way, if net demand from fundamental investors is $|D| = 2$, it is either $y_{\tau-\delta_{fast}} + \frac{s_2}{2}$ or $y_{\tau-\delta_{fast}} - \frac{s_2}{2}$ that clears the market. Similarly, for $s_3, s_4, \ldots$. The liquidity provider's benefits from providing liquidity, on a per-batch period basis, are

$$\sum_{d=1}^{\bar{Q}} Pr(|D| = d) \cdot d \cdot \frac{s_d}{2}$$

The liquidity provider's cost of providing liquidity, on a per-batch period basis, is τc_{speed}. We construct an equilibrium in which the liquidity provider recovers his costs and a strictly positive but arbitrarily small profit of $\epsilon > 0$ per unit time, i.e.,

$$\sum_{d=1}^{\bar{Q}} Pr(|D| = d) \cdot d \cdot \frac{s_d}{2} = \tau(c_{speed} + \epsilon) \tag{7.1}$$

If we consider the limit as $\tau \to 0^+$ and $\epsilon \to 0^+$, then we can obtain an instructive closed-form solution to (7.1). For τ short, the probability that there is 1 fundamental investor can be approximated as $Pr(|D| = 1) = \tau\lambda_{fund}$, and hence the benefits from providing the first unit of liquidity can be approximated as $\tau\lambda_{fund} \cdot \frac{s_1}{2}$. Hence, in the limit as $\tau \to 0^+$ and $\epsilon \to 0^+$, any solution to (7.1) has a bid-ask spread for the first unit of

$$\frac{s_1}{2} = \frac{c_{speed}}{\lambda_{fund}}. \tag{7.2}$$

In particular, a constant spread of $\frac{s_d}{2} = \frac{c_{speed}}{\lambda_{fund}}$ for all d is a solution to (7.1) in the limit.

Comparing the bid-ask spread under fast batching (7.2) to the bid-ask spread under continuous limit order books (6.1), we see that batching reduces the bid-ask spread by a term, $\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N}$, that represents the cost to the liquidity provider in continuous limit order books associated with being sniped by other fast traders. Alternatively, comparison to (6.3) shows that batching reduces the spread by $\frac{(N-1)c_{speed}}{\lambda_{fund}}$, which represents the welfare gain from reduced expenditure on c_{speed} as it manifests in the bid-ask spread.

In Appendix A.11 we show that a single fast liquidity provider offering spreads consistent with 7.1 constitutes a Nash equilibrium for τ sufficiently small. There are two key observations in the

proof. First, for τ sufficiently small, it will not be profitable for a slow market maker to enter the market and undercut the spread offered by the fast market maker. Intuitively, for τ sufficiently small, the slow market maker is nearly as vulnerable to getting sniped by the fast trader as in the continuous limit order book case, but his benefits from undercutting the fast trader are smaller than in the continuous limit order book case, since the fast trader offers a narrower bid-ask spread. Second, the fast liquidity provider will be tempted to deviate and charge a larger bid-ask spread than is prescribed by (7.1), but we can discipline against this using the off-path play of a potential entrant. The role of the strictly positive profits ϵ in (7.1) is to ensure that the incumbent finds it optimal not to provoke entry.[43]

We summarize this equilibrium as follows.

Proposition 11 (Equilibrium of Batch Auctions with Short Batch Intervals)**.** *If the batching interval τ is sufficiently small, then there exists a Nash equilibrium of the frequent batch auction market in which there is one fast market maker who serves as liquidity provider, offering bid-ask spreads consistent with (7.1), and zero other fast market makers. As compared to the equilibrium of the continuous limit order book market, the effects of batching are as follows:*

1. *The bid-ask spread for the first-quoted unit is narrower: in the limit as $\tau \to 0^+$ and $\epsilon \to 0^+$, the bid-ask spread is $\frac{c_{speed}}{\lambda_{fund}}$ instead of $\frac{N^* \cdot c_{speed}}{\lambda_{fund}}$ (see (6.3)).*

2. *The market is deeper: in the limit, there exists an equilibrium in which the order book has depth of $\bar{Q}$ at spread $\frac{c_{speed}}{\lambda_{fund}}$, whereas in the baseline model of the continuous limit order book just a single unit is offered in the order book, and in the extended model considered in Section 6.4 the bid-ask spread grows wider with the quantity traded.*

3. *Social welfare*

   (a) *benefit: expenditure on speed is reduced by $(N^* - 1) \cdot c_{speed}$ per unit time, independently of τ.*

   (b) *cost: fundamental investors pay expected delay costs of $\frac{1}{\tau} \int_0^\tau f_{delaycost}(x) \lambda_{fund} dx$ per unit time. As τ grows small, these costs go to zero per unit time.*

[43] This contestability aspect of our equilibrium has the following practical interpretation. In practice, with short batch intervals, we would expect there to be multiple fast market makers, each specializing in liquidity provision in many different markets. Entry into market making involves both the costs of speed and the costs of understanding each particular market that one enters (in our model this second cost is moot since y is a perfect signal of x's value). If observed bid-ask spreads in any one market are abnormally large, that will attract attention from HFT firms not currently specializing in that market, who then would invest in understanding that market.

7.3.4 Endogenous Entry: Long Batch Intervals

In this section we show that if the batch interval τ is sufficiently large, there is an equilibrium with zero entry by fast traders. Liquidity is provided to fundamental investors entirely by slow market makers (i.e., market makers who do not pay c_{speed}), at zero bid-ask spread.

Suppose that slow market makers in aggregate provide $\bar{Q}$ of depth for x at zero bid-ask spread. More precisely, $\bar{Q}$ slow market makers enter, and each offers a bid and an ask for a single unit at $y_{\tau-\delta_{slow}}$, where τ represents the end time of a generic batch interval, and $y_{\tau-\delta_{slow}}$ represents the best available information for a slow trader about the value of security x.

A potential entrant considers whether to invest c_{speed} to be fast, with the aim of picking off 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 fast trader will get to observe while the slow traders will not. If there are $\bar{Q}$ units of depth in the limit order book, and there is, say, a positive jump, the fast trader will wish to buy all $\bar{Q}$ units at the stale ask prices. If the imbalance D of fundamental 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 fundamental 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 market makers, 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 $E(J') \approx E(J)$. The fast trader's expected profits from exploiting the liquidity provider, on a per-unit time basis, are thus $\frac{p_{jump}}{\tau}E(J')\cdot\bar{Q} \approx \frac{\delta}{\tau}\cdot\lambda_{jump}E(J)\cdot\bar{Q}$. 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 trader's costs per unit time are c_{speed}. Hence, the fast trader will find it optimal not to enter provided that, using the approximations above,

$$\frac{\delta}{\tau}\cdot\lambda_{jump}\cdot E(J)\cdot\bar{Q} < c_{speed} \tag{7.3}$$

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 7.2), and these jumps occur at rate λ_{jump}. The LHS of

(7.3) 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 (7.3) is satisfied for sufficiently large τ. Hence, any desired market depth can be provided by slow traders at zero cost if the batch interval τ is sufficiently large. Moreover, the maximum depth $\bar{Q}$ consistent with (7.3) grows linearly with τ, whereas the expected imbalance of fundamental investor demand in a batch interval grows at rate $\sqrt{\tau}$.

We summarize the derived equilibrium as follows.

Proposition 12 (Equilibrium of Batch Auctions with Long Batch Intervals)**.** *If the batching interval τ is sufficiently large, then there exists a Nash equilibrium of the frequent batch auction market in which slow market-makers offer depth $\bar{Q}$ at zero bid-ask spread. As compared to the equilibrium of the continuous limit order book market, the effects of batching are as follows:*

1. *The bid-ask spread for the first-quoted unit is narrower: it is 0 instead of $\frac{N^* \cdot c_{speed}}{\lambda_{fund}}$.*
2. *The market is deeper: the order book has depth of $\bar{Q}$ at zero bid-ask spread, whereas in the baseline model of the continuous limit order book just a single unit is offered in the order book, and in the extended model considered in Section 6.4 the bid-ask spread grows wider with the quantity traded.*
3. *Social welfare*
   (a) *benefit: expenditure on speed is eliminated entirely, for a welfare savings of $N^* \cdot c_{speed}$ per unit time.*
   (b) *cost: fundamental investors pay expected delay costs of $\frac{1}{\tau}\int_0^\tau f_{delaycost}(x)\lambda_{fund}dx$ per unit time.*

7.4 Discussion of the Equilibria

In this section we make four remarks concerning the equilibria of frequent batch auctions.

First, it is instructive to compare the equilibrium under short batch intervals (Section 7.3.3) to the equilibrium both under continuous limit order book markets (Section 6.2) and to the equilibrium under longer batch intervals (Section 7.3.4). The first comparison indicates that moving from continuous limit order book markets to frequent batch auctions with short batch intervals has several important benefits and negligible costs. The benefits are that spreads are narrower, markets are deeper, and expenditure on speed is substantially reduced. The cost is that fundamental investors must wait a strictly positive amount of time to transact, but with

short batch intervals this cost intuitively seems small. The second comparison indicates that increasing the duration of the batch interval has additional benefits – spreads are even narrower, and expenditure on speed is eliminated altogether – but also real costs, as now fundamental investors must wait a non-negligible amount of time to transact. While stylized, we think that this analysis captures the relevant market design tradeoffs. The first comparison suggests that moving from continuous limit order books to frequent batch auctions with short τ is clearly beneficial for social welfare. The second comparison suggests that determining just how long to make τ is a more difficult market design decision, as increasing τ has real benefits but also real costs. Studying the optimal τ is an important direction for future research, and would benefit from a model tailored to study of this question.[44]

Second, the case we studied in Section 7.3.2, in which the number of fast traders is exogenously fixed, is instructive for thinking about the potential transition from continuous limit order books to frequent batch auctions. This case suggests that transitioning to frequent batch auctions will narrow spreads and improve depth for fundamental investors immediately, even if there are a large number of market making firms with substantial sunk cost investments in speed technology operating in the market. Under the continuous market, competition among fast market makers manifests in a race to snipe each other, which increases the cost of providing liquidity and ultimately the bid-ask spread. Under the batched market, at least in this simple model, competition among fast market makers manifests in a race towards narrower spreads and deeper markets for fundamental investors.

Third, we emphasize that the conclusion in Propositions 10 and 12 that bid-ask spreads are zero should not be taken literally. In particular, the reader should keep in mind that in the model of Section 6 we make several strong assumptions – no asymmetric information about fundamentals, no inventory or search costs – under which economic logic suggests that the market really should be able to provide effectively infinite depth at zero cost. In practice, of course, we would not expect frequent batch auctions to yield zero bid-ask spreads, in particular due to asymmetric information

[44]There are at least two other potential welfare benefits of batching that are outside our model. Wah and Wellman (2013) use a zero-intelligence agent-based simulation model to argue that frequent batching may lead to a more efficient match between supply and demand, if traders' valuations are heterogeneous. Considering a supply-demand diagram such as Figure 7.1, the intuition in their simulation is that batching makes it more likely that trade occurs at prices at or close to p^*, and hence that only buyers with values larger than p^* and sellers with costs less than p^* get to trade. Fuchs and Skrzypacz (2013) study a dynamic market with asymmetric information about fundamentals, and show that continuous trading can exacerbate the lemons problem. In their model, if there is asymmetric information at time 0 that will be resolved at time T, it can be socially efficient to restrict trading to occur only at times $\{0, T\}$ rather than to allow continuous trading throughout the interval $[0, T]$. If we interpret T as the duration of the information asymmetries that are exploitable by high-frequency trading firms, the Fuchs and Skrzypacz (2013) result can be interpreted as support for frequent batching. There also may be costs of batching that are outside our model. As discussed in the introduction, an important reason for keeping the batch interval short is to reduce the scope for such costs to arise.

about fundamentals (Kyle, 1985; Glosten and Milgrom, 1985). But we would expect that spreads are narrower than under the continuous limit order book case, because we have eliminated the purely technical cost of providing liquidity associated with stale quotes getting sniped.

Last, the conclusion in Proposition 11 that there is exactly one fast trader also should not be interpreted literally. Rather, we view the fact that the number of fast traders decreases from N^* to 1 as a metaphor for reduced expenditure on speed under batching (cf. footnote 43). Put differently, we encourage the reader to focus not on the reduction in the number of fast market makers from N^* to 1, but instead on the reduction in total expenditure on speed from $N^* \cdot c_{speed}$ to c_{speed}.

8 Frequent Batch Auctions and Market Stability

Our theoretical argument for batching as a response to the HFT arms race focuses on bid-ask spreads, market depth, 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.[45] In this section we outline several reasons why frequent batch auctions may enhance market stability relative to the continuous limit order book market design. These arguments are necessarily informal, but we include them due to the importance of the subject. As we note in the conclusion, we believe that market stability is an important topic for further research.

First, frequent batch auctions are computationally simple for the exchanges. Uniform-price auctions are fast to compute,[46] and exchange computers can be allocated a discrete block of time during which to perform this computation.[47] By contrast, in the continuous limit order book market design, exchange computers are not allocated a block of time during which to perform order processing, but instead process orders and other messages in serial order of their arrival. While processing any single order is computationally trivial, even a trivial operation takes non-

[45] Duncan Niederauer, 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).

[46] 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 a 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) – clears in about 10ms.

[47] For instance, with a 1 second batch interval, the first 100ms of each batch interval could be allocated to the exchange computers for computing and reporting outcomes from the previous batch interval.

zero 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 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. 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 cannot be sure what information they will have at each decision point, because of the small and somewhat random latencies involved in receiving price and trade updates from the exchanges. Additionally, in the continuous-time market, algorithmic traders are incentivized to trade off code robustness for speed, because error-checking takes time and even tiny speed advantages can matter.[48] While batching certainly cannot 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, and it makes the programming environment more natural, since code can be written with certainty about when information will arrive and by when decisions must be made.

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, and because the sequence of time stamps across exchanges may not reflect the actual sequence of events, due to varying processing delays across exchanges. It took months of analysis for regulators to understand the basic sequence of events that caused the Flash Crash (SEC and CFTC, 2010), and even today our understanding of that day's events remains incomplete.

Last, the theoretical results that show that batching leads to thicker markets (cf. Propositions 8, 10, 11 and 12) can also be interpreted as suggesting that batching enhances market stability, since thin markets are more vulnerable to what have come to be known as "mini flash crashes".[49]

[48] An interesting and analogous example is the use of microwave connections between New York and Chicago instead of high-speed fiber optic cable, as mentioned in the introduction. Microwaves are faster, shaving round-trip data transmission time from 13ms to 8.5ms, but they are less reliable, especially during adverse weather conditions (Adler, 2012).

[49] An example of what the press refers to as a mini flash crash occurred in the shares of Google on 4/22/2013. Google shares fell from $796 to $775 in roughly 0.75 seconds and then recovered to $793 within another second

In a sense, continuous markets implicitly assume that computers and communications are infinitely fast. Computers are fast but not infinitely so. The arms race for speed has made continuous markets vulnerable to instabilities that arise from the limitations of computing speed. Frequent batching in contrast 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 sealed-bid double auctions conducted e.g. every 1 second. To recap, our basic argument is as follows. First, we show empirically that continuous limit order book markets do not really "work" in continuous time: market correlations that function properly at human time scales completely break down at high-frequency time scales. Second, we show that this correlation breakdown creates technical arbitrage opportunities, available to whomever is fastest, which in turn incentivizes HFT firms to spend large sums of money on seemingly tiny speed advantages. Our empirical evidence suggests that the arms race profits should be thought of more as a constant of the continuous limit order book market design, rather than as a prize that is competed away over time. Third, we build a theoretical model guided by these empirical facts. We show that the arms race not only is intrinsically wasteful (like all arms races), but moreover that it leads to wider bid-ask spreads and thinner markets. Last, we show that discretizing the market eliminates the arms race, which in turn narrows spreads, enhances market depth and improves social welfare. Batching makes tiny speed advantages much less valuable. For example, if the batching interval is 1 second then a speed advantage of 1 millisecond is only $\frac{1}{1000}$ as useful as in the continuous market. Batching also changes the nature of competition, encouraging competition on price instead of on speed. Under the batched market, it no longer is possible to earn a rent from information that everyone in the market observes at basically the same time – a rent that ultimately comes out of the pockets of fundamental investors.

We leave open for future research several questions that relate to the practical implementation of frequent batch auctions. Most centrally, we do not attempt in this paper to calibrate the optimal batch interval. Other important practical implementation questions include the determination of optimal tick sizes, whether and in what form to include circuit breakers, and information policy.

(Russolillo, 2013). Similar incidents occurred in the shares of Symantec on 4/30/2013 (Vlastelica, 2013) and in the shares of Anadarko on 5/17/2013 (Nanex, 2013*b*). A reporter for CNNMoney wrote in March 2013 "There may not have been any major market malfunctions recently, but mini flash crashes still happen nearly every day. Stock exchanges don't publicly release data about these mini crashes – when a stock rapidly plunges then rebounds – but most active traders say there are at least a dozen a day." (Farrell, 2013)

Other things equal, we think that a useful principle to follow for practical implementation is to minimize departure from current practice, subject to realizing the benefits of batching relative to continuous limit order books.

A second important question for future research concerns the nature of competition among exchanges. Suppose that some exchanges switch to frequent batch auctions while other exchanges continue to use continuous limit order books: what is the equilibrium? We note that this question may be related to the question of the optimal batch interval; in particular, the potential threat of competition from other exchanges may be a force that suggests that batch intervals should be kept relatively short. This question may also have implications for regulatory policy.

A third important topic is to better understand issues of market stability. We discussed several reasons in Section 8 why frequent batching may enhance market stability relative to continuous limit order books; in particular, discretization respects the limits of computers and communications technology whereas continuous-time limit order books are computationally unrealistic. However, our arguments in Section 8 were speculative and informal in nature. Further research is needed, especially given the emphasis that practitioners and policy makers place 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/.

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

Bhave, Aditya, and Eric Budish. 2013. "Primary-Market Auctions for Event Tickets: Eliminating the Rents of "Bob the Broker"." *Working Paper*.

Biais, Bruno, Larry Glosten, and Chester Spatt. 2005. "Market Microstructure: A Survey of Microfoundations, Empirical Results, and Policy Implications." *Journal of Financial Markets*, 8(2): 217–264.

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

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2012. "High Frequency Trading and Price Discovery." *Working Paper*.

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

Bunge, Jacob, Jenny Strasburg, and Scott Patterson. 2013. "Nasdaq in Fresh Market Failure." *Wall Street Journal*, August 22: A1.

Cohen, Kalman J., and Robert A. Schwartz. 1989. "An Electronic Call Market: Its Design and Desirability." *in The Challenge of Information Technology for the Securities Markets: Liquidity, Volatility and Global Trading (Henry Lucas and Robert Schwartz, Eds.)*, 15–58.

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.

Ding, Shengwei, John Hanna, and Terrence Hendershott. 2013. "How Slow is the NBBO? A Comparison with Direct Exchange Feeds." *Working Paper*.

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*, Spring: 9–18.

Einstein, Albert. 1905. "Zur Elektrodynamik bewegter Körper (On the Electrodynamics of Moving Bodies)." *Annalen der Physik*, 17: 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, Doyne, and Spyros Skouras. 2012*a*. "The Value of Queue Priority." *Discussion Slides*.

Farmer, J. Doyne, and Spyros Skouras. 2012*b*. "Review of the Benefits of a Conitnuous 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.

Farrell, Maureen. 2013. "Mini Flash Crashes: A Dozen a Day." *CNN Money*, March 20. `http://money.cnn.com/2013/03/20/investing/mini-flash-crash/index.html`.

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, Roman Kozhan, and Wing Wah Tham. 2013. "Toxic Arbitrage." *Working Paper*.

Friedman, Milton. 1960. *A Program for Monetary Stability.* Fordham University Press.

Fuchs, William, and Andrzej Skrzypacz. 2013. "Costs and Benefits of Dynamic Trading in a Lemons Markets." *Working Paper*.

Glosten, Lawrence R. 1994. "Is the Electronic Open Limit Order Book Inevitable?" *Journal of Finance*, XLIX(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.

Grossman, Sanford, and Joseph Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review*, 70(3): 393–408.

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

Hasbrouck, Joel, and Gideon Saar. 2013. "Low-Latency Trading." *Johnson School Research Paper Series*.

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.

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

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

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica*, 1315–1335.

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

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

Najarian, Jon A. 2010. "The Ultimate Trading Weapon." September 21. http://www.zerohedge.com/article/guest-post-ultimate-trading-weapon.

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.

Nanex. 2013*a*. "Einstein and the Great Fed Robbery." http://www.nanex.net/aqck2/4436.html.

Nanex. 2013*b*. "How to Destroy $45 Billion in 45 Milliseconds." May. http://www.nanex.net/aqck2/4197.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. 2003. "Presidential Address: Liquidity and Price Discovery." *The Journal of Finance*, 58(4): 1335–1354.

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*.

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/.

Roll, Richard. 1984. "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market." *The Journal of Finance*, 39(4): 1127–1139.

Roth, Alvin E. 2002. "The Economist as Engineer: Game Theory, Experimentation and Computation as Tools for Design Economics." *Econometrica*.

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: 284–329.

Russolillo, Steven. 2013. "Google Suffers 'Mini Flash Crash,' Then Recovers." *Wall Street Journal*, April 22. http://blogs.wsj.com/moneybeat/2013/04/22/google-suffers-mini-flash-crash-then-recovers/.

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.

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

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.

The Government Office for Science, London. 2012. "Foresight: The Future of Computer Trading in Financial Markets." Final Project Report.

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: 219–254.

Vives, Xavier. 2010. *Information and Learning in Markets: The Impact of Market Microstructure.* Princeton University Press.

Vlastelica, Ryan. 2013. "Symantec Shares Plunge, Traders See Mini 'Flash Crash'." *Reuters*, April 30. http://www.reuters.com/article/2013/04/30/symantec-tradehalt-idUSL2N0DH1WK20130430.

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

Weller, Brian. 2013. "Intermediation Chains and Specialization by Speed: Evidence from Commodity Future Markets." *Working Paper*.

A Appendix: Proofs

A.1 Proof of Proposition 1

To complete the argument that the behavior described in Section 6.2 and Proposition 1 constitutes a Nash equilibrium, we make the following observations.

First, fundamental investors are optimizing given market-maker behavior. Fundamental investors have no benefit to delaying trade, since the bid-ask spread s^* is stationary, y is a martingale, they are risk neutral, and their costs of delay are strictly increasing. Hence, it is optimal for fundamental investors to trade immediately. Also, it is optimal for fundamental investors not to invest c_{speed} in speed. Suppose the fundamental trader arrives in the market at time t. Even though her own information about y is slightly stale – she knows $y_{t-\delta_{slow}}$ but does not know y_t – the liquidity provider's information is *not* stale, and the liquidity provider's quotes are based on y_t not $y_{t-\delta_{slow}}$ (recall that we normalized $\delta_{fast} = 0$, so $y_{t-\delta_{fast}} = y_t$). Furthermore, the liquidity provider's bid-ask spread is stationary as well.[50] Hence, the fundamental trader derives no benefit from investing in speed.

Second, let us confirm that the liquidity-provider's behavior is optimal given the behavior of fundamental investors and the stale-quote snipers. If the liquidity provider does not pay c_{speed} but otherwise acts as above, then his benefits from providing liquidity remain $\lambda_{fund} \cdot \frac{s^*}{2}$, but his costs increase to $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$, because instead of getting sniped with probability $\frac{N-1}{N}$ he is sniped with probability 1. Put differently, his costs increase by $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2}) \cdot \frac{1}{N}$. Inspection of equation (6.2) reveals that this increase in costs of getting sniped is exactly offset by the liquidity-provider's savings from not paying c_{speed}, hence the liquidity provider does not benefit from deviating to not pay c_{speed}. If at any moment in time the liquidity provider offers a wider bid-ask spread, $s' > s^*$, then one of the other market makers will want to offer a spread s'' that satisfies $s' > s'' > s^*$: the analysis above confirms that this would yield strictly positive profits. 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. Last, if the liquidity provider offers more than a single unit of quantity at the bid or ask, her benefits of providing liquidity stay the same (as it is, she satisfies all fundamental trader demand) but her costs of getting sniped will strictly increase, since she would get sniped for the full quantity. (See further discussion of this point in Section 6.4, when we generalize the model to include fundamental investors who demand multiple units).

Third, let us confirm that each stale-quote sniper's behavior is optimal given the behavior of the fundamental investors, the liquidity-provider, and the other stale-quote snipers. If a sniper does not pay c_{speed} then he will never successfully snipe, so sniping without being fast has zero benefits and zero costs. Hence, this is not an attractive deviation. 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

[50] One might expect that the liquidity provider will attempt to exploit a fundamental trader 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 fundamental investors. For instance, if y just jumped down in value, the liquidity provider might hope to sell to a fundamental trader at the old value of y (plus $\frac{s}{2}$). This is not possible in equilibrium, however, because then other market makers would no longer be indifferent between sniping and liquidity provision. They would prefer to offer more attractive quotes to fundamental investors.

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 equations (6.1) and (6.2) establish that this is not strictly preferred to the original strategy. 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 fundamental 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, we need to confirm that non-entrants cannot enter the market in a way that is profitable. If an entrant pays the cost c_{speed} to be fast and then enters the market as a stale-quote sniper, he will not recover his costs. If an entrant pays the cost c_{speed} to be fast and enters as a liquidity provider offering the same quotes as the original liquidity provider, then his quotes will reach the order book first only half the time, so he will not earn enough profits from trading with fundamental investors to recover his capital costs. The arguments above establish that he will not want to enter with a narrower or wider bid-ask spread than s. Last, if he enters as a market maker but does not pay c_{speed}, then sniping has both zero benefits and zero costs, and liquidity provision at any spread, given that there is already a liquidity provider offering s, has larger costs than benefits.

A.2 Proof of Proposition 2

The proposition follows immediately from (6.4), which characterizes s^* and does not depend on δ or c_{speed}. See also the text of Section 6.3.1.

A.3 Proof of Proposition 3

The proposition follows immediately from (6.4), noting that $\Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2})$ is strictly decreasing in s and is strictly increasing in mean-preserving spreads of F_{jump} (recall that J is the distribution of the absolute value of F_{jump}, and that F_{jump} is symmetric about zero).

A.4 Proof of Proposition 4

The claim that s^* is invariant to δ or c_{speed} follows immediately from (6.4). The claim that the total prize in the arms race is invariant to δ or c_{speed} follows from the preceding claim regarding s^* and the observation that, but for s^*, the other parameters in $\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$ are exogenous.

A.5 Proof of Proposition 5

Formally, there are N^* market makers, each of whom must choose the action *fast* or *slow*. If all N^* market makers choose slow, they each earn profits of c_{speed}, as described in the text of Section 6.2. If all N^* market makers choose fast, they each earn profits of zero, as described in Section 6.2. To show that *fast* is a dominant strategy, we make the following observations. If the number of market makers who choose fast is satisfies $1 < N < N^*$, then there is an equilibrium in which the N fast market makers play exactly as in Section 6.2, because indifference among the fast market makers between liquidity provision and stale-quote sniping (i.e., LHS of 6.1 equals LHS of 6.2) is still characterized by equation (6.4). The only difference is that each fast market maker earns larger profits than when all N^* enter, since they split the revenues from fundamental traders of $\lambda_{fund}\frac{s^*}{2}$ among N instead of splitting it among N^*. If the number of market makers who choose fast is 1, then there is an equilibrium in which the one fast market maker charges the maximum allowable bid-ask spread and is never sniped; these profits are larger than if all market makers are slow. Hence, for any number of fast market makers $0 \leq N < N^*$, any slow market maker strictly prefers to be fast than slow. Hence, fast is a dominant strategy, and we have a prisoner's dilemma.

A.6 Proof of Proposition 6

The observation that the midpoint of the bid-ask spread is equal to the fundamental value y_t 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 a fundamental investor is λ_{fund}. 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 a fundamental investor and a jump occur at the exact same time). Hence, trade at stale quotes as a proportion of all trade is $\frac{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*}}{\lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \frac{N^*-1}{N^*} + \lambda_{fund}}$.

A.7 Proof of Proposition 7

The proposition follows immediately from the equilibrium as summarized in Proposition 1 and the discussion in the text of Section 6.3.6.

A.8 Proof of Proposition 8

Equation (6.6) represents indifference between liquidity provision and stale quote sniping at the kth level of the book, for $k = 1, \ldots, \bar{q}$. The zero-profit condition for stale-quote snipers is

$$\sum_{k=1}^{\bar{q}} \lambda_{jump} \cdot \Pr(J > \frac{s_k}{2}) \cdot \mathbb{E}(J - \frac{s_k}{2} | J > \frac{s_k}{2}) \cdot \frac{1}{N} = c_{speed} \tag{A.1}$$

Notice that (A.1) sums the sniper's expected profits over all $\bar{q}$ units of the book, and asks that these total benefits equal the costs c_{speed}. Together, (6.6) and (A.1) represent $\bar{q} + 1$ equations in the $\bar{q} + 1$ unknowns, the $\bar{q}$ bid-ask spread terms and the level of entry.

To solve this system of equations, we can first use (6.6) to characterize each bid-ask spread. For the kth level of the book, the equilibrium bid-ask spread s_k^* is the unique solution to the following rearrangement of (6.6):

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

The solution to (A.2) 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. We can then plug the equilibrium bid-ask spreads $s_1^*, \ldots, s_{\bar{q}}^*$ into (A.1) to obtain the equilibrium entry quantity N^*. Given $s_1^*, \ldots, s_{\bar{q}}^*$ and N^*, the rest of the argument for equilibrium proceeds identically to that in the proof of Proposition 1.

The fact that $s_1^* < s_2^* < \cdots < s_{\bar{q}}^*$ follows from (A.2), because the probability that a fundamental investor wants to trade k units, $\sum_{i=k}^{\bar{q}} p_i$, is strictly decreasing in k. The comparative statics in each s_k^* also follow directly from (A.2), analogously to Proposition 3.

A.9 Proof of Proposition 9

The three claims for frequent batch auctions are established in the text of Section 7.2. The first claim for continuous limit order book markets is definitional. The latter two claims for continuous limit order book markets follow from the description of equilibrium of the continuous limit order book markets in Section 6.2 and Proposition 1.

A.10 Proof of Proposition 10

First, notice that it is not profitable for any player to offer liquidity at a bid-ask spread greater than zero. This follows from the fact that there are $N \geq 2$ fast market makers each already offering depth of $\bar{Q}$ at zero bid-ask spread. Any player who offers liquidity at a larger spread will never trade.

Second, as described in Section 7.2, fast market makers are not vulnerable to sniping in the batch auction. So, it is not profitable to enter as a fast market maker with the intent of sniping,

nor is it profitable for any of the N fast market makers exogenously present in the market to attempt to snipe the other liquidity providers.

Last, by assumption the $N \geq 2$ fast market makers are exogenously present in the market, each exogenously paying c_{speed}, so exit is not an option. If exit were an option this would be a profitable deviation, since the fast market makers are not recovering c_{speed}.

A.11 Proof of Proposition 11

We will show that a single fast liquidity provider offering spreads consistent with 7.1 constitutes a Nash equilibrium of the frequent batch auction for τ and ϵ sufficiently small. To do this we need to show four things.

First, we need to confirm that there is not a profitable deviation in which a slow trader enters the market with a positive bid-ask spread s' that is lower than the liquidity provider's spread s_d for some d, in an effort to profitably provide the dth unit of liquidity to fundamental investors. As $\tau \to 0^+$ and $\epsilon \to 0^+$, we have that (i) the spread s_1 implied by the zero-profit condition is converging to (7.2) and (ii) the likelihood that the absolute value of net demand $|D| \leq 1$ is converging to one. Hence, for τ and ϵ sufficiently small, the benefits from providing liquidity as a slow entrant are strictly smaller than the benefits such an entrant would have enjoyed as an entrant in the equilibrium of Section 6.2. Additionally, as $\tau \to 0^+$, the costs to a slow entrant from getting sniped converge to the same costs such an entrant would have faced in Section 6.2.[51] In the equilibrium of Section 6.2 a slow entrant was indifferent between entering and not at the equilibrium spread derived in Section 6.2.5. Hence, in the batch market, with a narrower spread, a slow entrant strictly prefers not to enter.

Second, we need to confirm that the fast trader who acts as liquidity provider does not wish to deviate by charging a higher bid-ask spread in some batch interval. This can be enforced by off-equilibrium-path play of a potential entrant. If the incumbent fast trader raises his spread in some batch interval, then the potential entrant enters beginning with the next batch interval, pays c_{speed}, and, acts as the incumbent was supposed to in equilibrium. On this path, the incumbent who deviated then exits the market, and no longer pays c_{speed}. If the incumbent does not exit, then the incumbent and the entrant engage in Bertrand competition which drives the bid-ask spread to zero, so on this path the incumbent strictly prefers to exit once he has deviated.[52] The maximum deviation payoff is finite, and there is no discounting, so we can choose τ and ϵ such that the incumbent prefers not to deviate and to instead earn $\epsilon > 0$ in perpetuity.

Third, we need to confirm that there is not a profitable deviation in which another fast trader enters the market, if he is not provoked by a deviation by the incumbent. This can be enforced off-path by assuming that the incumbent and the entrant engage in Bertrand competition in the event of such an entry, which drives the bid-ask spread to zero. Hence, the entrant cannot recover his costs of speed.

[51] That is, to enforce this equilibrium, the fast liquidity-provider threatens to pick off a slow entrant, in the off-path event that one should enter.

[52] Intuitively, there is a "token" that indicates who gets to play the role of liquidity provider, and in this equilibrium it is understood that if the current liquidity provider deviates from his prescribed play, the token is automatically passed to another player. Any player not holding the token chooses not to pay c_{speed}. See footnote 43 in the main text for a discussion of the practical interpretation of this equilibrium.

Last, we need to confirm that the fast trader does not wish to deviate by not paying the cost c_{speed} and instead providing liquidity as a slow trader. If he does so and offers a spread that is weakly less than the spread in the equilibrium of Section 6.2, then, for τ sufficiently small, another market maker profits by entering as a fast trader just to pick off his stale quotes. If he does so and offers a spread that is wider than the spread in the equilibrium of Section 6.2, then, for τ sufficiently small, another market maker profits by entering as a fast trader who both (i) acts as the fast trader is supposed to in this equilibrium (i.e., according to 7.1), and (ii) picks off the slow trader who is offering a wider spread.

A.12 Proof of Proposition 12

To complete the argument that the behavior described in Section 7.3.4 and Proposition 12 constitutes a Nash equilibrium, we make the following two observations.

First, we established in the text that it is not profitable to enter as a fast market maker. Picking off stale quotes is not sufficiently profitable, as shown by (7.3) and the surrounding discussion. Additionally, it is not profitable to enter as a fast market maker in an effort to provide liquidity, because slow market makers are already providing the maximum necessary liquidity, $\bar{Q}$, 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 market maker's activity in such event both exploits the stale quotes of the slow market makers and provides liquidity to the net demand of fundamental investors, yielding $\bar{Q}$ of total volume in expectation. As discussed, this is not sufficiently profitable to induce the fast trader to enter.

Second, each individual slow market maker has no incentive to deviate. In order to earn strictly positive profits, a slow market maker would have to charge a strictly positive bid-ask spread. But, since there are $\bar{Q}$ slow market makers in total, and the support of $D(\tau)$, the net demand of fundamental investors, is bounded by $\bar{Q} - 1$, any individual slow market maker who deviates will never get to trade. Additionally, our discussion above shows that it is not profitable for a slow market maker to pay the cost c_{speed} and play instead as a fast market maker. Hence, there is no deviation that yields strictly positive profits.