Online Appendix for "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response"

Eric Budish, Peter Cramton and John Shim

Forthcoming, *Quarterly Journal of Economics*

A Backup Materials for the Empirical Analysis

A.1 Correlation Breakdown Appendix

A.1.1 Alternative Measures of the ES-SPY Correlation

In this section, we describe alternative measures of the ES-SPY correlation as a robustness check for the results presented in Section 5.1.1. These correlation measures vary along three dimensions. First, we consider both equal-weighted bid-ask midpoints and quantity-weighted bid-ask midpoints. Whereas equal-weighted midpoints place weight of $\frac{1}{2}$ on the bid and the ask, quantity-weighted midpoints place weight $\omega_t^{bid} = \frac{Q_t^{ask}}{Q_t^{ask}+Q_t^{bid}}$ on the bid and weight $\omega_t^{ask} = 1-\omega_t^{bid}$ on the ask, where Q_t^{bid} and Q_t^{ask} denote the quantity offered at the bid and ask at time t. Second, we consider correlation measures based on both simple returns and on average returns. Specifically, given a time interval τ and a time t, the simple return is the percentage change in price from time $t-\tau$ to time t, and the average return is the percentage change between the average price in the interval $(t-2\tau, t-\tau]$ and the average price in the interval $(t-\tau, t]$. Last, we consider three different ways to handle the concern that the speed-of-light travel time between New York and Chicago is roughly 4 milliseconds, which, per the theory of special relativity, represents a lower bound on the amount of time it takes information to travel between the two locations. One approach is to compute correlations based on the perspective of a trader in New York, which treats Chicago events as occurring 4ms later in New York than they do in Chicago. That is, the New York perspective treats Chicago events with time stamp t as contemporaneous with New York events with time stamp $t+4ms$. A second approach is to compute correlations based on the perspective of a trader in Chicago, in which case New York events with time stamp t are treated as contemporaneous with Chicago events with time stamp $t+4ms$. A last approach is to adjust neither dataset; this can be interpreted either as ignoring speed-of-light concerns or as taking the vantage point of a trader equidistant between Chicago and New York.

Table A.1 displays the median ES-SPY correlation for varying time intervals over all trading days in 2011 for each of our 12 ($= 2 \times 2 \times 3$) methods of computing the correlation. As can be seen from the table, the pattern depicted in Figure 5.1, which uses our main specification of equal-weighted midpoints, simple returns, and no speed of light adjustment, is robust across all of the alternative specifications – at high enough frequency the ES-SPY correlation completely breaks down.[52]

A.1.2 Equities Correlation Breakdown

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

Table A.2a displays the correlation at different time intervals between pairs of equity securities that are relatively highly correlated, for instance, the oil companies Exxon-Mobil (XOM) and Chevron (CVX). Table A.2b displays the correlation matrix for the 5 largest market capitalization US equities at a short and long time horizon. We follow the main specification used in Section 5.1, using equal-weighted midpoints and simple returns. As can be seen from the tables, the equities market correlation structure breaks down at high frequency. At human time scales such as one minute there is economically meaningful correlation between these securities, but not at high-frequency time scales such as 1ms or 100ms.

A.2 Mechanical Arbitrage Appendix

A.2.1 Computing the ES-SPY Arbitrage

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

Figure A.1 illustrates the exercise we conduct. The top portion depicts the midpoint prices of ES and SPY over the course of a fairly typical 30-minute period of trading (Panel a) and a volatile 30-minute period of trading during the financial crisis (Panel b). Notice that, while there

[52] 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 A.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 correlations several different ways. First, we use either equal-weighted or quantity-weighted midpoints in computing returns. Quantity-weighted midpoints weight the bid and ask by $\omega_t^{bid} = Q_t^{ask} / \left(Q_t^{ask} + Q_t^{bid}\right)$ and $\omega_t^{ask} = 1 - \omega_t^{bid}$, respectively, where Q_t^{bid} and Q_t^{ask} represent the quantity offered at the bid and ask. Second, we use either simple or averaged returns. Simple returns use the conventional return formula and averaged returns use the return of the average midpoint of two non-overlapping intervals. Third, we compute correlations from the perspective of a trader in New York (Chicago events occurring at time t in Chicago are treated as contemporaneous with New York events occurring at time $t + 4ms$ in New York; labeled NY), a trader in Chicago (New York events occurring at time t in New York are treated as contemporaneous with Chicago events occurring at time $t + 4ms$ in Chicago; labeled Chi), and a trader equidistant from the two locations (labeled Mid). For more details on these correlation computations, see the text of Appendix A.1.1. For more details on the data, refer to Section 4 of the main text.

(a) Equal-Weighted Midpoint Correlations

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

(b) Quantity-Weighted Midpoint Correlations

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

Table A.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 4 of the main text.

(a) Pairs of Related Companies

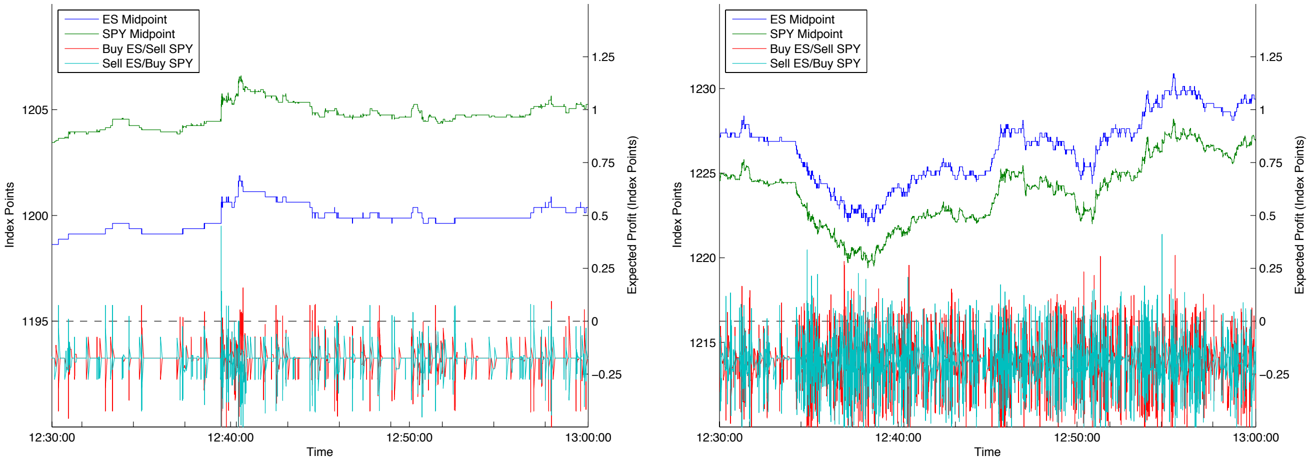
	1 ms	10 ms	100 ms	1 sec	10 sec	1min	10 min	30 min
HD-LOW	0.008	0.039	0.102	0.194	0.433	0.605	0.703	0.676
GS-MS	0.005	0.031	0.095	0.183	0.402	0.551	0.664	0.704
CVX-XOM	0.023	0.128	0.281	0.456	0.652	0.745	0.760	0.789
AAPL-GOOG	0.001	0.013	0.060	0.134	0.302	0.434	0.533	0.631

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

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

Figure A.1: Mechanical Arbitrage Illustrated

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



(a) Ordinary Day

(b) Financial Crisis

is a difference in levels between the two instruments, which is described in the main text (cf. Section 5.2.1), the two instruments' 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 instrument (at its ask) and selling the other (at its bid). Most of the time these instantaneous profits are negative, reflecting the fact that buying one instrument 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 instrument's price has just jumped a meaningful amount but the other's price has not yet changed – which we know is common from the correlation breakdown analysis in Section 5.1. At such moments, buying the cheaper instrument and selling the more expensive instrument (with cheap and expensive defined relative to the difference in levels) is sufficiently profitable to overcome bid-ask spread costs. Our exercise is to compute the frequency, duration, and profitability of such trading opportunities.

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

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

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

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

where τ^* denotes the amount of time it takes, in milliseconds, for the ES-SPY averaged-return correlation to reach 0.99, in the trailing month up to the date of time t. The high correlation between ES and SPY at intervals of length τ^* implies that prices over this time horizon produce relatively stable spreads.[53] We define a trading rule based on the presumption that, at high-frequency time horizons, deviations of S^{mid}_t from $\bar{S}_t$ are driven mostly by the correlation breakdown phenomenon we documented in Section 5.1. For instance, if ES and SPY increase in price by the same amount, but ES's price increase occurs a few milliseconds before SPY's price increase, then the instantaneous spread will first increase (when the price of ES increases) and then decrease back to its initial level (when the price of SPY increases), while $\bar{S}_t$ will remain essentially unchanged.

We consider a deviation of S^{mid}_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 instantaneous spread market according to $S^{bid}_t = P^{bid}_{ES,t} - 10 \cdot P^{ask}_{SPY,t}$ and $S^{ask}_t = P^{ask}_{ES,t} - 10 \cdot P^{bid}_{SPY,t}$. Note that $S^{bid}_t < S^{mid}_t < S^{ask}_t$ at all times t by the fact that the individual markets cannot be crossed, and that typically we will also have $S^{bid}_t < \bar{S}_t < S^{ask}_t$. 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., $\bar{S}_t < S^{bid}_t$ or $S^{ask}_t < \bar{S}_t$, 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 instrument at its ask while selling the other at its bid. As discussed in the text in Section 5.2.1, this is a conservative and simple way to account for transactions costs. Expected profits, on a per-unit spread basis, are thus:

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

If our presumption is correct that the instantaneous market crossing the moving-average is

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

due to correlation breakdown, then in the data the instantaneous market will uncross reasonably quickly, i.e., $S_t^{bid} < \bar{S}_t < S_t^{ask}$. We define the ending time of the arbitrage, t_{end}, as the first millisecond after t_{start} in which the market uncrosses, the duration of the arbitrage as $t_{end} - t_{start}$, and label the opportunity a "good arb." If the expected profitability of the arbitrage varies over the time interval $[t_{start}, t_{end}]$, i.e., the instantaneous spread takes on multiple values before it uncrosses the moving average, then we record the full time-path of expected profits and quantities and compute the quantity-weighted average profits. This requires maintaining both the actual empirical order book and a hypothetical order book which accounts for our arbitrageur's trade activity. It is common that the trades in ES and SPY that our arbitrageur makes overlap with trades in ES and SPY that someone in the data makes, and we account for this in order to avoid double counting.[54]

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

If an arbitrage opportunity lasts fewer than 4ms, the one-way speed-of-light travel time between New York and Chicago, it is not exploitable under any possible technological advances in speed. Therefore, such opportunities should not be counted as part of the prize that high-frequency trading firms are competing for, and we drop them from the analysis.[55]

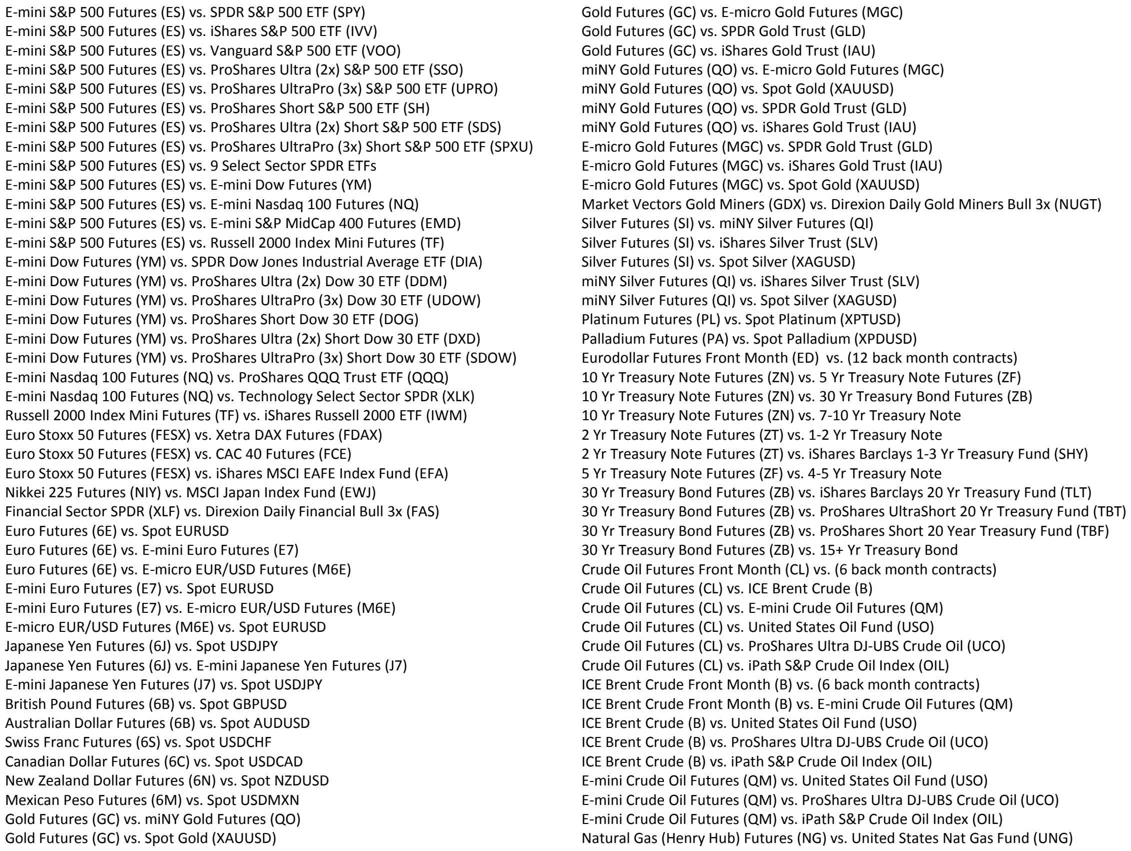
A.2.2 Illustrative List of Highly Correlated Financial Instruments

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

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

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

Figure A.2: Illustrative List of Highly Correlated Financial Instruments



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)

B Backup Materials for the Theoretical Analysis

B.1 Proofs of Theoretical Results

B.1.1 Proof of Proposition 1

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

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

Second, let us confirm that the liquidity-provider's behavior is optimal given the behavior of investors and the stale-quote snipers. If at any moment in time the liquidity provider offers a bid less than $y_t - \frac{s^*}{2}$ or an ask greater than $y_t + \frac{s^*}{2}$, then one of the other trading firms will want to undercut her. For instance, if the liquidity provider sets an ask of $y_t + \frac{s''}{2}$ with $s'' > s^*$, then one of the other trading firms will immediately respond with an ask of $y_t + \frac{s'}{2}$ with $s'' > s' > s^*$. Our analysis in Section 6.2.3 which shows that providing liquidity at a bid-ask spread of s^* is exactly as attractive as sniping when the bid-ask spread is s^* implies that providing liquidity at $s' > s^*$ is strictly preferred to sniping. Hence, a deviation which widens the bid-ask spread (either on one or both sides) is not possible in equilibrium.[56] If the liquidity provider offers a narrower bid-ask spread, $s' < s^*$, then her profits are strictly lower than they are with a spread of s^*, so this is not an attractive deviation either. Third, if the liquidity provider offers a first unit of liquidity at s^* and then additional units of liquidity on either side of the book at a spread weakly greater than s^*, her benefits of providing liquidity stay the same (as it is, she satisfies all investor demand) but her costs of getting sniped will strictly increase, since she would get sniped for the full quantity. (The exception is if she offers additional liquidity at a spread so wide that it is never sniped; this is allowed for in the proposition statement). Last, if the liquidity provider offers zero units on either side of the book, then it is attractive for other trading firms to provide liquidity, and the reasoning above implies that they will do so at a bid of $y_t - \frac{s^*}{2}$ and/or an ask of $y_t + \frac{s^*}{2}$; this is just a permutation of roles, which is allowed for in the proposition statement.

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

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

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

B.1.2 Proof of Proposition 2

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

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

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

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

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

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

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

B.1.3 Proof of Proposition 3

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

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

Third, we show that there is no opportunity for entry by a slow trading firm for any number $N' \geq 2$ of fast trading firms. Clearly, slow trading firms will never succeed at sniping stale quotes when there are fast trading firms present. So we need to rule out the possibility of a slow trading firm providing liquidity. Suppose a slow trading firm provides liquidity at a spread of $s' < s^*$, i.e., a slow trading firm attempts to undercut the spread of the fast trading firm providing liquidity. The benefits to the slow trading firm from investor arrivals, per unit time, are $\lambda_{invest} \cdot \frac{s'}{2}$. The costs from getting sniped, per unit time, are $\lambda_{jump} \cdot \Pr(J > \frac{s'}{2}) \cdot \mathbb{E}(J - \frac{s'}{2} | J > \frac{s'}{2})$. Notice that whereas the fast trading firm is sniped with probability $\frac{N'-1}{N'}$, the slow trading firm is sniped with probability one.[57] Since $s^* > s'$, we have $\lambda_{invest} \cdot \frac{s'}{2} < \lambda_{invest} \cdot \frac{s^*}{2}$ and $\lambda_{jump} \cdot \Pr(J > \frac{s'}{2}) \cdot \mathbb{E}(J - \frac{s'}{2} | J > \frac{s'}{2}) > \lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$. But, (6.3) shows that $\lambda_{invest} \cdot \frac{s^*}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2}) = 0$ in equilibrium, so the slow trading firm's benefits less costs of liquidity provision, $\lambda_{invest} \cdot \frac{s'}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s'}{2}) \cdot \mathbb{E}(J - \frac{s'}{2} | J > \frac{s'}{2})$, are strictly negative. This shows that slow trading firms cannot profitably enter as liquidity providers with a bid-ask spread $s' < s^*$. It is obvious that they cannot profitably enter as liquidity providers with a bid-ask spread of $s' \geq s^*$ either. Hence, slow trading firms cannot profitably enter as liquidity providers.

Last, we show that the equilibrium quantity of entry by fast trading firms is N^*. Clearly, entry of $N' > N^*$ fast trading firms is not an equilibrium, since all such firms would lose money once speed costs are accounted for. Similarly, entry of $N' < N^*$ fast trading firms is not an equilibrium, since a marginal entrant could enter profitably. Hence, in any equilibrium, the number of fast trading firms is N^*.[58]

Hence, the behavior described in the text of Section 6.4.2 constitutes an equilibrium (sub-game perfect Nash at the entry stage, static Nash throughout the trading day, per fn. 18), as claimed, and the equilibrium number of fast trading firms N^* and the equilibrium bid-ask spread s^* are uniquely characterized by the zero-profit conditions (6.5)-(6.6), as claimed. The final statement of the proposition follows directly from (6.3) and (6.7).

B.1.4 Proof of Proposition 4

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

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

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

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

B.1.5 Proof of Proposition 5

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

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

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

B.1.6 Proof of Proposition 6

The three claims for frequent batch auctions are established in the text of Section 7.2. The three claims for the continuous limit order book market design follow from the description of equilibrium in Section 6.

B.1.7 Proof of Proposition 7

First, observe that it is a static Nash equilibrium for each of the N trading firms to offer depth of $\frac{\bar{Q}}{N-1}$ at zero bid-ask spread (i.e., to set bid and ask of y_τ for $\frac{\bar{Q}}{N-1}$ units), and for investors to trade at market in the batch auction immediately following their arrival. Investors clearly cannot do better since they fulfill their demand at zero cost as soon as is possible. Trading firms earn zero profits in this equilibrium. However, given the behavior of the other trading firms, each individual trading firm can do no better than to offer $\frac{\bar{Q}}{N-1}$ at zero bid-ask spread; if it sets a strictly positive spread for any unit the probability that that unit trades is zero, given the assumption that $\bar{Q}$ is an upper bound on the imbalance of investor demand in a batch interval. Similarly, it is not possible for a trading firm to snipe, since all trading firms are equally fast (cf. Proposition 6). Hence, there is no deviation that earns strictly positive profits.

Second, we show that this equilibrium is essentially unique. Suppose that there is an equilibrium in which trading firms earn strictly positive profits in the auction ending at τ. This means that there exists an investor imbalance quantity q (without loss consider $q > 0$) such that (i) the probability that the imbalance is q is strictly positive, and (ii) if the imbalance is q the market-clearing price is strictly greater than y_τ. Since imbalance is bounded, there exists a largest such imbalance satisfying (i) and (ii); call it $\hat{q}$, and call the price that results under this imbalance $\hat{p} > y_\tau$. Let $\hat{a}$ denote the quantity of asks in the book at price $\hat{p}$. Since $N \geq 2$ there exists at least one firm whose own quantity of asks at price $\hat{p}$ is strictly less than $\hat{a}$. This firm has a profitable deviation: keep its supply function unchanged except offer $\hat{a}$ units at a price of $\hat{p} - \epsilon$ for $\epsilon > 0$ and sufficiently small. Under this deviation the firm earns strictly higher profits if the imbalance is $\hat{q}$ (or any other imbalance that results in a market-clearing price of $\hat{p}$) which occurs with positive probability, and the same profits otherwise. A contradiction.[59]

B.1.8 Proof of Proposition 8

First, we established in the text that it is not profitable to enter as a fast trading firm given the hypothesized equilibrium behavior of slow trading firms. Picking off stale quotes is not sufficiently profitable, as shown by (7.1) and the surrounding discussion. Additionally, it is not profitable to enter as a fast trading firm in an effort to provide liquidity, because slow trading firms are already providing the maximum necessary liquidity at zero bid-ask spread. One last thing to point out is that the discussion in the text already covers the possibility of providing liquidity in the event

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

that there is a jump between times $\tau - \delta_{slow}$ and $\tau - \delta_{fast}$; the fast trading firm's activity in such an event both exploits the stale quotes of the slow trading firms and provides liquidity to the net demand of investors, yielding $\bar{Q}$ of total volume in expectation. As discussed, this is not sufficiently profitable to induce the fast trader to enter.

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

Last, we show that this equilibrium is essentially unique. We do this by ruling out four cases. First, suppose that there is an equilibrium in which there are only slow trading firms but in which some of the slow trading firms earn strictly positive profits. The argument in the second paragraph of the proof of Proposition 7 applies to show that this generates a contradiction; there will be a profitable opportunity to undercut. Second, suppose that there is an equilibrium in which two or more fast trading firms enter. The proof of Proposition 7 applies to show that if multiple fast trading firms enter they will compete spreads to zero, so this is not a profitable deviation. Third, suppose that there is an equilibrium in which a single fast trading firm enters, alongside the fringe of slow trading firms, and the fast trading firm provides all liquidity to investors. Consider a hypothetical equilibrium liquidity supply schedule $s_1 \leq s_2 \leq \cdots \leq s_{\bar{Q}}$, where s_k denotes the difference between the fast trading firm's kth best ask and kth best bid. Let $Rev(k)$ denote the hypothetical equilibrium revenue for the kth unit of liquidity provision, per unit time.[60] Let $Cost(k)$ denote the hypothetical cost to a slow trading firm, per unit time, of getting sniped by the fast trading firm if she provided the kth unit of liquidity at spread s_k instead of the fast trading firm.[61] There must exist a k for which $Rev(k) > \frac{c_{speed}}{\bar{Q}}$; otherwise, the total price of liquidity for investors could not exceed c_{speed} per unit time, which would mean that the fast trading firm cannot recover her costs. However, we know from (the exact version of) (7.1) that the sniping cost per unit of liquidity is strictly less than $\frac{c_{speed}}{\bar{Q}}$ per unit time even if the spread is zero, hence for all k we have $\frac{c_{speed}}{\bar{Q}} > Cost(k)$. Hence, there exists a k for which $Rev(k) > \frac{c_{speed}}{\bar{Q}} > Cost(k)$, i.e., there exists a unit of liquidity for which the liquidity-provision revenues strictly exceed the sniping costs.

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

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

Hence, a slow trading firm can profitably undercut the fast trading firm, a contradiction. Fourth, suppose that there is an equilibrium in which a single fast trading firm enters, alongside the fringe of slow trading firms, and liquidity is provided by both the fast trading firm and slow trading firms. Since sniping is zero sum, for it to be the case that neither the fast trading firm wishes to undercut a slow trading firm providing liquidity nor that a slow trading firm wishes to undercut either the fast trading firm or another slow trading firm, we must have $Rev(k) = Cost(k)$ for all $k = 1, \ldots, \bar{Q}$. But, per the argument for the previous case, for the fast trading firm to recover her costs her profits per unit time must exceed $\frac{c_{speed}}{\bar{Q}}$ for at least one level of the book. Hence, for some k we have $Cost(k) = Rev(k) > \frac{c_{speed}}{\bar{Q}}$, which generates a contradiction since, as described in the previous case, (7.1) implies that $\frac{c_{speed}}{\bar{Q}} > Cost(k)$ for all k.

B.2 Supporting Materials for Section (8): Alternative Responses to the HFT Arms Race

B.2.1 Tobin Tax Analysis

Introduce a Tobin tax of $\theta > 0$ per unit traded to the endogenous entry model of Section (6.4). For simplicity, assume that the tax is paid by the liquidity taking side of the trade; the analysis is economically identical if the two sides split the tax or the liquidity providing side pays the tax.

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

The results described informally in the text are formally stated as follows:

Proposition 9 (Effect of Tobin Tax)**.** *In the model of the continuous limit order book with endogenous entry as studied in Section 6.4, adding a Tobin tax of $\theta > 0$ per unit traded has the following effects:*

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

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

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

Proof of Proposition 9 Equilibrium with the Tobin tax is still governed by zero-profit conditions for liquidity provision and stale-quote sniping, as in Proposition 3, which also encode indifference between the two roles. The zero-profit conditions are now

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

for the liquidity provider and

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

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

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

B.2.2 Random Message Delay Analysis

To use our model to analyze the random message delay, we add an assumption for tractability that there are not multiple jumps in y or investor arrivals within the horizon of the random delay.

Proposition 10 (Effect of Random Message Delays). *Consider the model of the continuous limit order book with endogenous entry as studied in Section 6.4. Incorporate a random message delay that is a uniform draw from $[0, \epsilon]$, for some $\epsilon > 0$. Assume that, after any jump in y or investor arrival, both Poisson processes pause for 2ϵ time, the maximum time it takes to submit and then cancel an order. There is an equilibrium with the following features:*

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

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

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

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

Proof of Proposition 10 If is straightforward to see that there is an equilibrium with the same structure as in Section 6.4, with N^* trading firms of whom 1 provides liquidity and $N^* - 1$ snipe, with the bid-ask spread s^*, and equilibrium characterized as in Section 6.4 by the zero-profit conditions (6.5)-(6.6). The only difference is that after each jump in y that is larger than $\frac{s^*}{2}$, all N^* trading firms send their desired message (either a snipe or a cancel) as often as possible, namely $\bar{M} \to \infty$ times. It is random which of the $N^* \cdot \bar{M}$ messages reaches the exchange first, so just as in Section 6.4 each fast trading firm has a $\frac{1}{N^*}$ probability of winning the race. The 2ϵ pause assumption ensures that the liquidity provider can maintain depth of exactly one at all times when investors might arrive to market and ensures that stale-quote snipers can cancel unsuccessful snipes without risk that they themselves get sniped in the event of a subsequent jump in the opposite direction. The $\bar{M} \to \infty$ limit ensures that the probability that a slow trading firm can win the race to snipe goes to zero because of the δ speed disadvantage. Hence there is no role for slow trading firms in this equilibrium, just as in Section 6.4.

B.3 Supporting Materials for Section 7.3.3: How Long is Long Enough to Stop the Speed Race?

B.3.1 Calibration of Equation (7.1)

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

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

λ_{jump} and $\mathbb{E}(J)$ represent the frequency and size of sniping opportunities, or more precisely jumps that would be sniping opportunities if they occur during the correct δ interval of the batch interval. In our ES/SPY data, there are roughly 200,000 sniping opportunities per year (800 per day times 250 days), averaging roughly 0.01 per share. $\bar{Q}$ represents the depth of the order book in the auction. In our SPY data, top-of-book depth averages about 40,000 shares.[62] Our theory predicts that frequent batch auctions should narrow spreads – which both increases the likelihood that a jump creates a sniping opportunity and increases the profits of any given sniping opportunity – and increase depth. We thus double the product of these figures from ES/SPY, i.e., we use $\lambda_{jump}\mathbb{E}(J)\bar{Q} = 2 * 200,000 * 0.01 * 40,000 = 160,000$ annualized.

c_{speed} represents the cost of speed. Data on speed expenditures by high-frequency trading firms are mostly proprietary. An exception is that the high-frequency trading firm GETCO's financial data was released publicly when GETCO merged with Knight Capital Group, because the merger filing detailed the financials of each of the merging firms separately (KCG, 2013). In 2012, the last year for which standalone GETCO data are available, GETCO spent \$84M on colocation and data line expenses, \$31M on capital expenditures, and \$161M on employee compensation. For each of these expenses it is not possible to know how much of the expense was related to speed per se, so to be conservative suppose that \$100M of the total relates to speed. This \$100M figure then represents the annual cost of speed for a single firm, for all of its trading activity, not just ES/SPY. If we assume that ES/SPY represents 1% of the speed race – under this assumption, our \$75M estimate for the total prize in ES/SPY would imply a total prize overall of \$7.5bn per year – then the annual cost of speed that should be attributed to ES/SPY is 1% of the \$100mm, for c_{speed} of \$1 million per year.

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

As should be clear from the discussion of each of the inputs above, these figures should be interpreted as giving no more than an extremely rough sense of magnitudes. One can easily tinker

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

with each of the inputs above to get a bound for τ that is an order of magnitude larger (e.g., if depth $\bar{Q}$ is considerably higher, ES/SPY is considerably less than 1% of the speed race, or c_{speed} is lower), or an order of magnitude smaller.

Below we analyze a modification of our model in which, under frequent batch auctions, information arrives in discrete time rather than continuous time. The idea of this modification is that, to the extent that information y about the value of security x is information about other prices, then the use of frequent batch auctions would cause information to arrive in discrete time at frequency τ. Under this modification we obtain an equilibrium analogous to that in Section 7.3.3 but with a simpler and less stringent sufficient condition under which frequent batch auctions stop the speed race: $\tau > \delta_{slow}$. For New York - Chicago trades, δ_{slow} is about 4ms under the first interpretation of slow traders and about 7ms under the second interpretation. For equities only trades, δ_{slow} would be less than 1ms under the first interpretation of slow traders and at most a few milliseconds under the second interpretation.

B.3.2 Modification to the Model: Endogenous Entry with Discrete-Time Information Arrival

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

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

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

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

The following proposition describes the equilibrium.

Proposition 11 (Equilibrium of Frequent Batch Auctions with Endogenous Entry and Discrete–Time Information Arrival)**.** *Consider a modification of the model of Section 7.3.3 in which in-*

formation y evolves in discrete time, with updates occurring as one batch interval ends and the next begins. Let τ satisfy (B.4)*, a sufficient condition for which is* (B.5)*. Then any equilibrium is analogous to the equilibrium in Proposition 8:*

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

Proof. Observe that given the assumption about information arrival there is zero benefit to speed if τ satisfies (B.4); both slow trading firms and fast trading firms have exactly the same information at the conclusion of each batch interval. Hence, there is no reason to pay the cost c_{speed} and there are only slow trading firms in equilibrium. Given this observation, the result that there is at least $\bar{Q}$ of depth at zero bid-ask spread, in any equilibrium, follows directly from the arguments in the proof for the exogenous entry case, Proposition 7.