Speed and Latency in Treasury and e-Mini Futures Contracts – Part 2

Raymond P. H. Fishe, Richard Haynes, and Esen Onur[*]

I. Introduction

Fast trading is a focus of regulators and many industry groups. The use of computer algorithms, co-location services, technological improvements in exchanges' matching systems, and high-speed microwave networks accelerates order entry, cancellation rates, execution speeds, and matching frequencies for equities and derivative markets. Both the Commodity Futures Trading Commission (CFTC) and the U.S. Securities and Exchange Commission (SEC) have examined faster traders and automated trading strategies to better understand their effects on regulated markets.[1] Although academic research exists on the effects of speed, latency and high frequency trading, much of this research is limited by a lack of detailed, participant level, proprietary data.[2] Many issues related to speed and latency cannot be fully addressed without individual order and trade data; this paper complements prior research by using account-level audit trail data.

The purpose of this research is to provide trader based information on both the speed of trading and message latency in the Treasury futures complex and the e-Mini futures contract. This is Part II of our analysis of speedy traders. In this study we examine message latency, defined below, in trader systems and strategies. Part I of this analysis examined the speed or rapidity of trading across a set of participant types.

Latency refers to how long it takes to reach a resolution from a particular starting point; latency is quoted in units of time and may be measured in a number of ways depending on the latency measure of interest. For financial markets, there are several important categories of latency, each of which may have implications for trading activity:

---

[*] **Fishe**: Patricia A. and George W. Wellde, Jr. Distinguished Professor of Finance, Department of Finance, Robins School of Business, University of Richmond, Richmond, VA 23173. Tel: (+1) 804-287-1269. Email: pfishe@richmond.edu. **Haynes**: U.S. Commodity Futures Trading Commission, Washington, D.C. 20581. Tel: (+1) 202-418-5000. Email: rhaynes@cftc.gov. **Onur**: U.S. Commodity Futures Trading Commission, Washington, D.C. 20581. Tel: (+1) 202-418-5000. Email: eonur@cftc.gov. Tel: (+1) 202-418-5000. The research presented in this paper was co-authored by Raymond Fishe, a CFTC limited term-consultant, and Richard Haynes and Esen Onur, who are both CFTC employees, in their official capacities with the CFTC. The Office of the Chief Economist and CFTC economists produce original research on a broad range of topics relevant to the CFTC's mandate to regulate commodity future markets, commodity options markets, and the expanded mandate to regulate the swaps markets pursuant to the Dodd-Frank Wall Street Reform and Consumer Protection Act. These papers are often presented at conferences and many of these papers are later published by peer-review and other scholarly outlets. The analyses and conclusions expressed in this paper are those of the authors and do not reflect the views of other members of the Office of Chief Economist, other Commission staff, or the Commission itself. All errors and omissions, if any, are the authors' own responsibility. First draft: September 2016.

[1] On March 25, 2015, the SEC closed an exemption that allowed proprietary trading firms—many are HFTs—to actively trade without being members of a National Securities Association and therefore subject to the regulatory arm of the industry (FINRA).

[2] Menkveld (2016) provides a recent review of HFT research as it relates to market quality metrics.

1. *Message origination latency*—how long it takes a trader or algorithm to initiate or transmit a message to a trading platform after the message process has reset; that is, the process starts *de novo* and then a signal (or set of signals) arises that is sufficient to cause a new message.
2. *Communications latency*—how long it takes the message to reach the intended trading platform after departing the originating system.
3. *Trading platform latency*—how long it takes the trading platform to process and respond to the message. The response may be dependent on message type; for instance, a confirmation-of-receipt response may be expected to have much lower latency on average than an execution response. Specifically, the latter response generally involves limit order resting times on an order book.
4. *Public broadcast latency*—how long it takes for post-message information to be broadcast from the trading platform to the trading public, assuming that such messages contain information intended for public broadcast.

Message origination latency arises because a trading strategy waits until a signal (or a marginal signal in a sequence of signals) causes it to generate a new message. For algorithmic strategies, message origination latency is expected to be well-defined and consistent across events, as the code to implement the algorithm embeds all of the necessary information to define when to originate a message. For cognitive strategies, there may not be a clear way to measure what signals cause a human trader to react and originate a message because idiosyncratic responses to signals may cause timing differences or may not coincide with sets of very similar signals.

Understanding message origination latency may also require tracking not just market signals, but signals related to a trader's prior actions—so-called feedback signals. Specifically, confirming that a previous order has executed may be a sufficient signal to initiate a new message to a trading platform. Thus, strategies that rely on market-generated data for sufficient signals will in turn generate messages whose frequency is related to specific market data frequencies or patterns.

Communications latency is generally two sided. It involves the time it takes a message to travel from its origin to the trading platform and the time to receive a response after the trading platform has confirmed the validity of the message.[3] Each side of this exchange may have different latency times as communications routes may vary. Geographical location is a key component of this latency measure. Traders with co-located facilities will have shorter communications latency times than other, more distant traders.

Trading platform latency may be decomposed into two separate types of latency.[4] The first is the time the platform takes to pre-process a message (e.g., validity checks) and initiate a confirmation to the originating entity. This is typically embedded in communications latency. In

[3] While not common, it is possible that the trading platform can reject an order instead of confirming it. One scenario in which this happens is when the trading platform prevents irrationally priced or sized limit orders from populating the order book and rejects them.

[4] For a discussion of trading platform latency, see Kirilenko and Lamacie, "Latency and Asset Prices," Working paper, MIT Sloan School of Management, 2015.

times of extreme market activity, pre-processing may slow, as other messages and their related processing delays a normally near immediate confirmation response. These delays may also occur in public broadcast latency.

The second type of trading platform latency is the time the matching engine takes to act on a message. For market-type orders (e.g., marketable limit or market orders), this is the time it takes to confirm a match on the existing book, typically less than a millisecond. These times may increase if the matching process is more complicated, such as for trades with multiple legs or for trades that make use of functionality such as implied spread technology. For messages such as modifications or cancellations of existing orders, this is the time it takes to adjust or remove an existing order in the electronic order book.[5] For these order types, latency does not generally depend on the actions of other market participants.

In contrast, for a resting limit order message, matching engine latency is the time it takes to execute an order by finding a match in the market. In this case, the behavior of other participants affects latency as the order may sit on the book until sufficient execution or cancel volume moves it to the front of the queue. In effect, the latency associated with execution time may be quite long if the limit price is set away from the current market price or very short if the order is placed to replenish exhausted liquidity at the best price.[6]

Public broadcast latency focuses on how long it takes for the information produced by a matching engine to be revealed to the overall market. This is mostly a question of the difference in time between execution and post-trade transparency, but it may also include how quickly the order book updates when depth increases or decreases after new orders, modifications, or cancellations. A possible regulatory issue arises if public broadcast latency is meaningfully different between selected market participants.[7] For example, if a trader receives confirmation of an execution before the overall market, then that trader may possibly gain by reacting before the rest-of-the-market can process the new information.

These four types of latency have been affected by technology and regulatory changes during the past two decades. On net, communications, trading platform, and public broadcast latency times have decreased over this period, in many cases to a significant extent. With the growth of algorithmic trading, message origination latency also appears to have decreased, but there is more ambiguity inherent in understanding message origination times. Specifically, did message origination latency decrease because algorithms, co-location, and direct connections endowed traders with faster methods of order entry, or did it decrease because the signals generating such orders increased in frequency (possibly due to technology external to markets), and thereby increased demand for algorithms, co-location and direct connections?

[5] When orders are modified or cancelled, trading platform latency may increase due to the need to update implied spreads in the matching engine.

[6] For orders away from the market price, matching engine latency includes the possibility that the order does not execute, which may significantly lengthen average latency times if there are many such orders on the book.

[7] See Scott Patterson, Jenny Strasburg and Liam Pleven, "High-Speed Traders Exploit Loophole," The Wall Street Journal, May 1, 2013. (http://www.wsj.com/articles/SB10001424127887323798104578455032466082920).

The analysis below focuses on message origination latency—specifically, the latency tied to feedback signals, such as a trade confirmation message. Our goal is to offer empirical data that may approximate message origination latency. We say 'approximate' because this analysis is subject to the limitation that we do not know exactly what signals generate trader responses in our sample. As such the analysis embeds several assumptions about the signal sequence generating responses in order to establish start and end times for latency calculations. In addition, because we use messages from the market and do not incorporate external signals, we may do a poor job explaining the latency of traders who rely significantly on external signals.

II. Message Origination Latency

Message origination latency cannot be precisely measured without knowledge of the trader's strategy, including the timing and nature of the signals necessary to implement the strategy. In effect, we need a detailed flowchart of either the algorithms used for trading or specific instructions offered by human traders. With this caveat, we try to determine how message origination latency may be estimated from exchange-based data. Clearly, these data are only a subset of the signals available to both algorithmic and human traders, and so may be provide a bound on latency.

The essence of latency is that it is a measure of time between two, potentially related, events. In our context, a signal (or signals) begins the message origination process and then the actual message creation ends the process; this establishes some ambiguity on what "signal" should be chosen to define the beginning of this time interval. We posit that message origination latency may be approximated by analyzing how long a trader takes to act after receiving an "exit" signal from an existing order. Specifically, both execution and cancellation confirmations represent exit signals in which a trader's order is removed from the matching engine. *Our premise is that these exit signals represent an approximate starting point for the trader's (or algorithm's) strategy*. The strategy then processes other signals until sufficient information is received that causes a new order message. From this view, message origination latency is how long it takes for a new order message to be transmitted after an exit signal has been received.

Figure 1: Illustration of Message Origination Latency

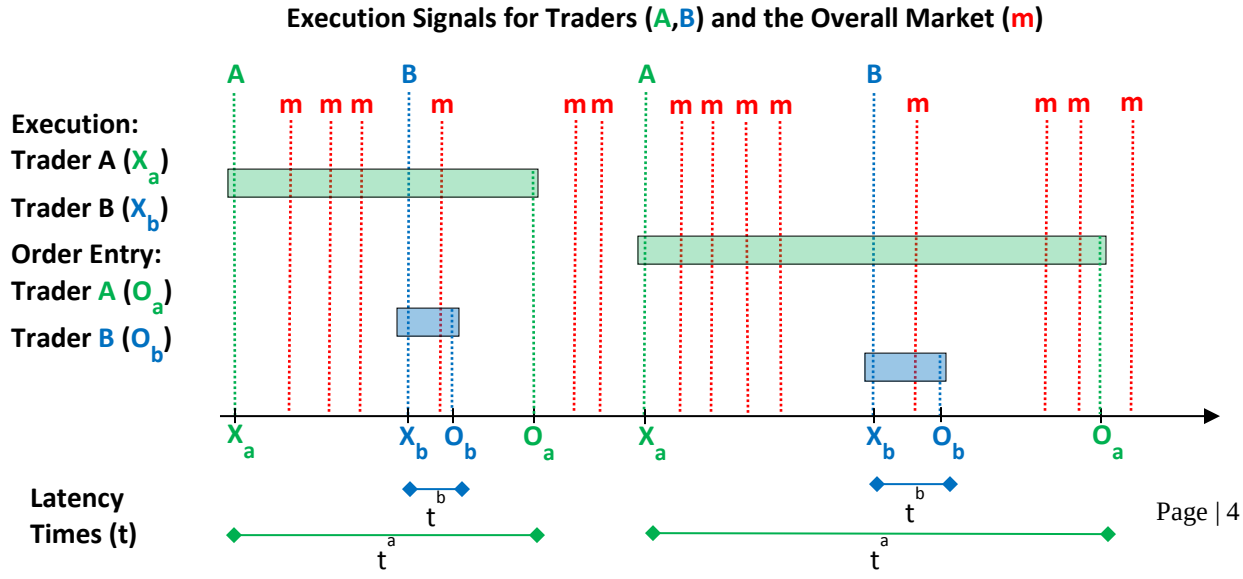


Figure 1 illustrates message origination latency based on the assumption that an execution signal starts the process that leads to a new order message. Two traders are shown in the figure, labeled "A" and "B". Execution signals provided are marked by vertical dashed lines. All of these are public signals, except that the two traders know when their trade orders are confirmed. This confirmation indicates an exit from the market for their previous order. The latency times (t) of each trader are marked depending on when they re-enter the market with a new order. We make no distinction between the buy and sell sides for this illustration, but do emphasize the differences in types of messages (i.e., order entry, cancellation, modification, and execution). The figure shows that Trader B appears to act after receiving the next (public) trade execution signal from the market, which suggests a low latency (i.e. a low threshold or simple rule for generating a new order) for this trader's message origination strategy. Trader A appears less responsive to market execution signals, initiating new orders only after five and eight execution signals, respectively. In effect, Trader A's message origination latency may be a function of other, possibly private signals, which may not be apparent in our order book-derived sample.

Note that cancellations also create exit signals and represent a significant proportion of message traffic. Traders may use cancellations for many purposes, such as to clear out stale limit orders, to search for hidden liquidity when tied to more aggressive quoting, to avoid trading with informed traders, or to avoid being adversely selected as price moves away from their resting limits. Unlike execution signals that confirm limit order matches, cancellations do not involve the actions of other traders. In effect this exit message and the new order message are both created by the trader's (or algorithm's) independent actions. Thus, message origination latency derived from cancellations may show more consistency at the trader level than the same latency using execution messages.

III. Hazard Model

Figure 1 also suggests a modeling strategy for message origination latency. Because we seek to explain how long it takes for a trader to submit a new order after an existing order has exited the market, the problem is analogous to those examined using survival analysis.[8] Cox (1992) developed a proportional risk model for survival analysis that has been shown to be flexible for many different latency applications. The Cox model assumes that covariates have a multiplicative effect on the hazard function.[9] This approach requires a specification of the hazard function (or rate). Specifically, if $t_{i,g}$ represents the time between order exit and order entry, where $g = 1,\ldots,G_i$ indexes the number of observable exit-to-reentry gaps for trader i, then a Cox-type hazard model with covariates may be specified as:

$$\lambda\left(t; x_{i,g}\right) = \lambda_0(t) exp\left(\boldsymbol{x}'_{i,g}\boldsymbol{\beta}\right), \qquad (1)$$

[8] See Kalbfleisch and Prentice (2002) for a discussion of survival time models and examples.

[9] The hazard function specifies the instantaneous failure rate at time t given that there is no failure between the initial time and t. This function equals the density of failure time divided by the probability of survival beyond time t, the latter is known as the survivor function.

where $\lambda(t; x_{i,g})$ is the hazard function for the latency time between order exit and order entry, which depends on an arbitrary baseline hazard (λ_0), and covariates ($\boldsymbol{x}'_{i,g}\boldsymbol{\beta}$) that have a multiplicative effect via the exponential specification. The covariates examined here are variables that define the trader, such as manual or algorithmic, and variables that define information (or signals) arising during the gap between exit and reentry. Some covariates are time dependent, so we adjust our estimation methods to allow for such dependence.

The coefficient vector, $\boldsymbol{\beta}$, in the Cox model is estimated using partial information maximum likelihood methods. Because of proportionality, the baseline hazard is not involved with these estimates. We use the PHREG routine in SAS to estimate these coefficients.

The covariates we examine are those that would be known to the market or closely approximated during a gap between an exit and re-entry of a new order. Specifically, traders would know the volume of trading, and from updates to the book, they may determine buy- and sell-side flows onto and off of the book. We measure these factors using both counts of messages by type as well as the quantities being adjusted by these messages. We consider these covariates as information signals received by the market participants. Time dependencies arise because the longer the gap, the greater the number of messages, on average, within the gap. To adjust for time dependence, we repeatedly sample selected covariates within the gap.

IV. Data

The data studied here are for the ten- and thirty-year treasury futures complex and the E-Mini futures contract. We analyze order book data for proprietary accounts on one trading day, August 1, 2014. On this day, there was a morning release by the BLS of July employment data. Those data showed the U.S. added 209,000 non-farm payroll jobs and the unemployment rate increased to 6.2%.[10] Expectations were for 230,000 jobs and the unemployment rate unchanged at 6.1%. This release may have added some volatility to trading during this day, so that is a caveat to our results.

The data sampled are only for the December 2014 expiration month, which was an active month at this time for all of these products. For each futures contract, participant, entry-type (algorithmic or manual), and customer type (proprietary or customer-initiated), we identified the last execution and last cancellation prior to a new order. The confirmation timestamp on this last message marks the beginning of a signal-processing gap for that participant. That is, this is the time of exit from the market and by assumption the beginning of a new strategy sequence that processes signals before re-entry. Re-entry to the market occurs when a new order is submitted, and we use the time stamp for CME receipt of the new order to mark this re-entry time. The difference between these re-entry and exit times we define as the message origination latency for each participant.

[10] See Myles Udland, "Jobs disappoint, unemployment rate rises to 6.2%," *Business Insider*, August 1, 2014.

Table 1 provides summary statistics on these sample data. This table reports information as averages and standard deviations across all message origination gaps. Thus, these data may be skewed towards latency times associated with participants who have disproportionately more gap

Table 1
Summary Statistics for "Message Origination Latency" Events

Summary data are presented for the average and standard deviation of covariates measured in the time gap between the last execution and the next new order entry (Panel A) and the last cancellation and the next new order entry (Panel B). This time gap is hypothesized to approximate message origination latency. These data only include participants who acted for their own accounts, so-called proprietary traders. The data are separated by algorithmic and manual-entry accounts for three futures contracts: E-mini, Ten-year Treasury, and 30-year Treasury.The data are for the December 2014 expiration and all orders and trades on August 1, 2014.

	E-Mini Futures Contract				Ten-Year Treasury Note Futures				Thirty-Year Treasury Bond Futures			
	Algorithmic-Entry		Manual-Entry		Algorithmic-Entry		Manual-Entry		Algorithmic-Entry		Manual-Entry	
Signal Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Panel A: Data for Message Origination Gaps Measured from Last Execution to Next New Order Entry												
Gap Latency (Seconds)	8.52	35.5	57.1	117.3	11.91	44.7	66.4	116.1	11.16	44.2	57.9	111.7
Cancel Count	436.6	1827.5	3241.8	6900.1	381.6	1490.6	2960.8	5575.6	305.9	1256.4	2110.2	4402.5
Execution Count	423.2	1777.9	3161.8	6883.8	335.1	1181.1	2413.2	4631.0	176.5	671.7	1157.8	2428.9
New Order Count	647.2	2725.5	4845.5	10437.8	524.4	2026.8	4066.6	7722.4	392.3	1597.0	2706.8	5659.0
Individual Cancel Cnt	0.30	1.08	0.15	0.64	0.26	1.11	0.12	0.48	0.26	1.10	0.13	0.48
Begin Gap Inventory	-3.23	186.2	11.3	362.2	-1.89	162.5	-29.4	468.7	0.55	58.2	-42.9	148.6
Abs(Inventory)	55.3	177.8	101.1	348.0	54.0	153.3	127.1	452.1	23.3	53.4	79.6	132.6
Sign of Inventory	0.57	0.49	0.53	0.50	0.55	0.50	0.53	0.50	0.58	0.49	0.48	0.50
Number of Gaps	268,146		7,778		122,848		2,858		88,873		1,402	
Panel B: Data for Message Origination Gaps Measured from Last Cancellation to Next New Order Entry												
Gap Latency (Seconds)	3.59	19.5	60.8	120.5	9.72	37.6	63.2	128.7	9.59	37.6	26.4	79.7
Cancel Count	168.2	1024.4	3429.4	7261.5	268.4	1230.5	2122.3	4865.7	216.6	985.6	734.4	2760.3
Execution Count	162.7	915.5	3112.9	6621.4	220.7	1027.8	1636.7	3912.6	124.9	543.3	400.6	1533.9
New Order Count	247.2	1484.1	5004.0	10630.1	358.6	1677.4	2837.0	6588.8	272.4	1261.9	927.7	3544.5
Individual Execution Cnt	0.20	0.91	0.45	3.12	0.17	0.97	0.29	2.51	0.15	0.79	0.03	0.26
Begin Gap Inventory	-2.58	136.6	-7.2	331.4	1.75	145.2	-26.0	545.7	1.39	58.9	-41.4	133.5
Abs(Inventory)	38.5	131.1	75.0	322.8	40.3	139.5	102.0	536.7	18.5	55.9	46.8	131.7
Sign of Inventory	0.62	0.49	0.63	0.48	0.65	0.48	0.75	0.44	0.66	0.47	0.73	0.44
Number of Gaps	385,656		2,645		163,156		1,563		139,769		2,159	

events than others. However, even with such skewness, no single trader represents more than 40% of these data. Panel A in the table shows results for the execution-to-new-order sequence and Panel B shows results for the cancellation-to-new-order sequence. The number of gaps analyzed is reported at the bottom of each panel. For algorithmic participants there are more cancel-to-new-order sequences for each futures contract, while this is only true for 30-year treasury futures for manual participants. Intuitively, this would imply that message origination latency for the cancel-to-new-order sequences is less than that of the execution-to-new-order sequence. The means of the gap latency measures in each panel confirm this observation.

Table 1 also shows summary results for covariates used in the hazard rate analysis (see below). These data are measured up to three times during a gap, but the data here are for totals over the entire gap length unless otherwise noted. The cancel, execution, and new order count variables show smaller averages for algorithmic- than manual-entry participants. This is consistent with the smaller average latency for algorithmic gaps. The individual data—execution and cancel counts—tend to suggest than algorithmic participants have more activity during the gap than manual participants, except for execution counts for E-mini and Ten year treasuries. Finally, the inventory data show higher absolute averages for manual- than algorithmic-entry participants.

V. Empirical Analysis

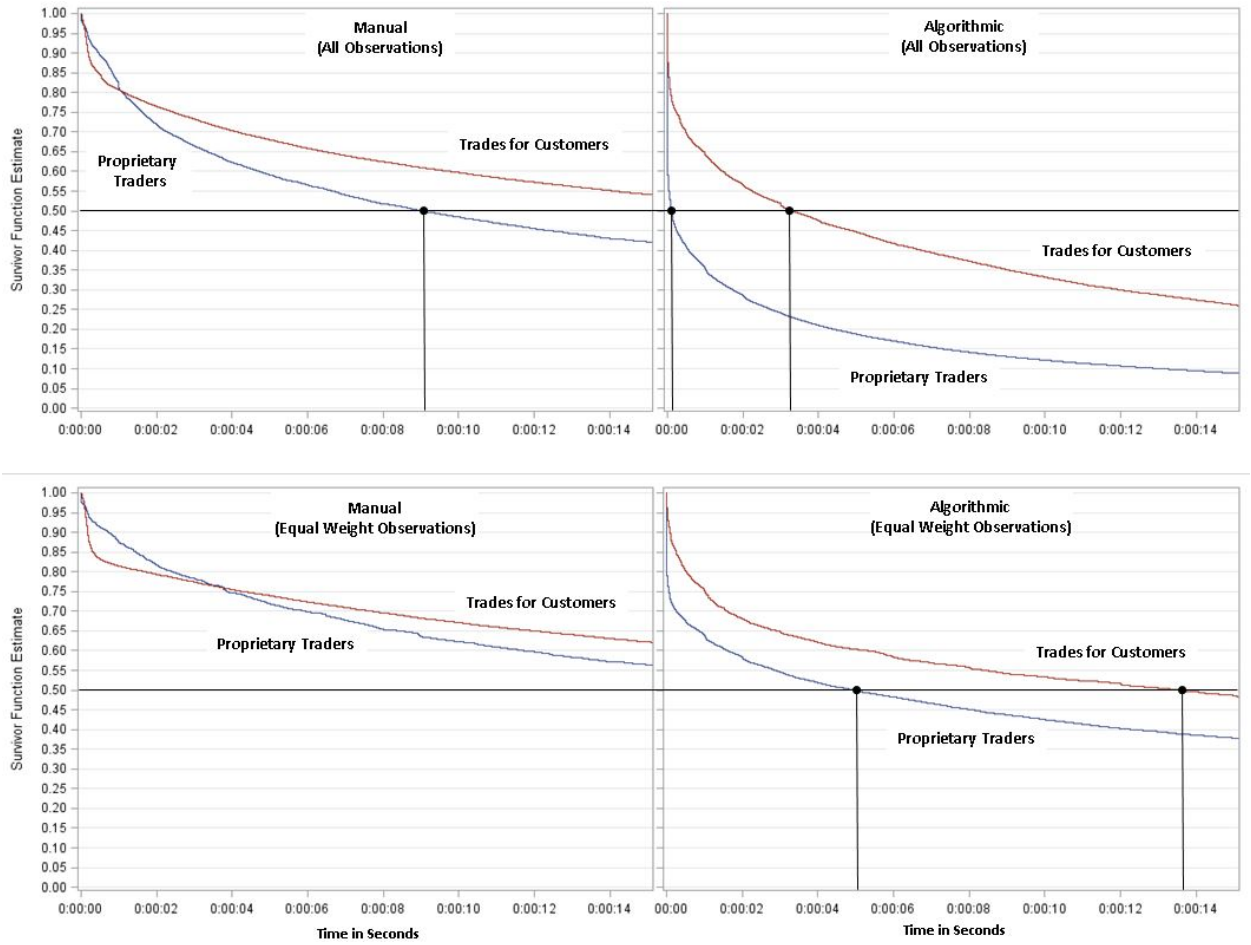
We begin by providing a more detailed analysis of how long participants in these markets wait before re-entering an order after an exit message signal. The distribution that describes this behavior is known as a survival curve. Figure 2 shows survival curves for traders in the e-Mini market, where participants are grouped by order entry (manual or algorithmic) and customer type (proprietary or customer-initiated order). These data are for message origination latency gaps measured from an execution message to a new order by participant. The upper panel shows survival curves for *all* observations while the lower panel bootstraps the data into 500 equal-weighted samples such that each participant is observed once in every sample. The bootstrap method removes the excess weight given to participants who are very active in the market; that is, high frequency traders.

The solid line crossing both graphs in Figure 2 highlights the 50% cutoff for each group. This line identifies the latency time (horizontal axis) when one-half of observed new orders have been submitted after receiving a prior execution message. Both panels show that the 50% latency cutoff implies that algorithmic-entry traders are quicker to respond versus manual-entry traders after an execution signal. The ‘all’ observation curves in the upper panel show algorithmic proprietary traders are the quickest group with 50% of the observations responding with new orders in less than 200 milliseconds. In contrast it takes manual proprietary traders over 9 seconds on average for one-half of the observations to respond.

The lower panel in Figure 2 emphasizes the problem created if all observations are collected into a single sample for analysis. This panel plots survival curves when each participant is equally weighted. There are meaningful differences in execution-to-new-order latency between participants as comparing the upper and lower panels indicates. Moreover, the groups used here

(manual versus algorithmic) offer only limited controls for this heterogeneity. This point is clear from comparing the intersections of the 50% cutoff line in the lower panel to those in the upper panel. The cutoff line does not cross either manual-entry survival curve in the lower panel, so the cutoffs for those participant groups exceed the 15 second limit on the horizontal axis. More significantly, the new cutoff for the algorithmic proprietary group increases latency time by a factor of 25 to just over 5 seconds, and by a factor of 4 to nearly 14 seconds for the customer-based algorithmic group. These results reveal how conclusions about market behavior may be meaningfully affected by disproportionate activity levels across participants. In other words, conclusions drawn from all observations will be about the average observation, not the average participant.

Figure 2: Survival Curves for E-Mini Futures: Execution-to-New-Order Latency
(All observations versus equally-weighted bootstrap sample; Data for August 1, 2014)



Tables 2 and 3 contain the estimates of our proportional hazard model. Table 2 reports results for a latency model of the time between order execution and new order entry. Table 3 shows results for the latency time between order cancelation and new order entry. The results are grouped by commodity—E-Mini, Ten-year Treasury, and 30-year Treasury futures contracts—

Table 2
Model of Latency between Order Execution and New Order Entry

Proportional hazard regression estimates are shown for a latency model of the time between order execution and new order entry. Models are estimated for E-Mini, Ten-year Treasury, and 30-year Treasury futures contracts using only participants trading for their own accounts (Proprietary traders). The models are estimated separately for algorithmic- and manual-entry participants. Covariates measured in the time gap are the number of cancellations, executions, and new orders in the entire market excluding the participant identified in the observation. The participant's cancellations within the gap are also included. All of these covariates are sampled up to three times--at approximately 20%, 50% and 75% of gap length if sufficiently long and populated--within the gap to measure time dependence. The inventory covariates are measured at the start of the gap and do not change for a given participant during the gap. The model is estimated using a partial likelihood function that takes account of time dependent covariates using a bootstrap simulation. The simulation includes 500 random samples in which a gap for each participant is drawn once for every sample. The simulation gives equal weight to all participants and removes the effect caused by only a few participants having many gaps. The table shows the average estimated hazard rate for each covariate, the average p-value of the covariate's estimated coefficient and the 95% confidence interval (95% C.I.) of the hazard ratio. The count of p-values less than 0.05 is shown for each coefficient in all models. The number of participants in each simulation and McFadden pseudo r-squared are shown at the bottom of the table for each model.

	E-Mini Futures Contract				Ten-Year Treasury Note Futures				Thirty-Year Treasury Bond Futures			
	Algorithmic-Entry		Manual-Entry		Algorithmic-Entry		Manual-Entry		Algorithmic-Entry		Manual-Entry	
Signal Variable	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05
Cancel Count	1.0111 0.0152 (1.0059, 1.0163)	474	1.0118 0.0278 (1.0054, 1.0182)	446	1.0152 0.0004 (1.0094, 1.0211)	500	1.0055 0.2880 (0.9951, 1.0161)	156	1.0248 0.0059 (1.0146, 1.0352)	491	1.0218 0.1924 (0.9981, 1.0461)	242
Execution Count	1.0027 0.1827 (0.9999, 1.0054)	241	0.9999 0.3142 (0.99626, 1.0036)	144	1.0040 0.1081 (1.0011, 1.0069)	352	0.9966 0.2723 (0.9915, 1.0017)	178	1.0118 0.0401 (1.0047, 1.0189)	431	1.0020 0.3928 (0.9849, 1.0193)	85
Individual Cancel Cnt	0.9976 <0.0001 (0.9973, 0.9978)	500	0.9770 <0.0001 (0.9729, 0.9811)	500	0.9978 <0.0001 (0.9975, 0.9980)	500	0.8771 <0.0001 (0.8572, 0.8976)	500	0.9975 <0.0001 (0.9972, 0.9979)	500	0.9443 <0.0001 (0.9305, 0.9584)	500
New Order Count	0.9922 0.0489 (0.9878, 0.9966)	419	0.9941 0.1884 (0.9885, 0.9996)	265	0.9878 0.0037 (0.9826, 0.9930)	495	1.0001 0.3375 (0.9909, 1.0093)	119	0.9780 0.0078 (0.9690, 0.9871)	487	0.9862 0.2648 (0.9656, 1.0073)	173
Inventory if Negative	1.0002 0.4866 (0.9992, 1.0009)	52	0.9991 0.2942 (0.9978, 1.0004)	166	1.0000 0.5333 (0.9992, 1.0007)	27	0.9993 0.2142 (0.9982, 1.0003)	239	1.0001 0.5200 (0.9974, 1.0028)	38	1.0016 0.3318 (0.9951, 1.0081)	151
Inventory if Positive	0.9998 0.4142 (0.9986, 1.0009)	90	1.0002 0.3574 (0.9993, 1.0012)	97	1.0000 0.5316 (0.9994, 1.0007)	22	0.9994 0.2560 (0.9973, 1.0015)	199	0.9991 0.4309 (0.9956, 1.0025)	77	1.0056 0.2984 (0.9976, 1.0136)	164
Number of participants	478		235		521		166		379		89	
McFadden R-Sqrd	10.9%		38.2%		6.6%		56.4%		6.9%		47.7%	
McFadden Adj. R-Sqrd	11.0%		38.5%		6.8%		56.9%		7.1%		48.8%	

Table 3
Model of Latency between Order Cancellation and New Order Entry

Proportional hazard regression estimates are shown for a latency model of the time between order cancellation and new order entry. Models are estimated for E-Mini, Ten-year Treasury, and 30-year Treasury futures contracts using only participants trading for their own accounts (Proprietary traders). The models are estimated separately for algorithmic- and manual-entry participants. Covariates measured in the time gap are the number of cancellations, executions, and new orders in the entire market excluding the participant identified in the observation. The participant's executions within the gap are also included. All of these covariates are sampled up to three times--at approximately 20%, 50% and 75% of gap length if sufficiently long and populated--within the gap to measure time dependence. The inventory covariates are measured at the start of the gap and do not change for a given participant during the gap. The model is estimated using a partial likelihood function that takes account of time dependent covariates using a bootstrap simulation. The simulation includes 500 random samples in which a gap for each participant is drawn once for every sample. The simulation gives equal weight to all participants and removes the effect caused by only a few participants having many gaps. The table shows the average estimated hazard rate for each covariate, the average p-value of the covariate's estimated coefficient and the 95% confidence interval (95% C.I.) of the hazard ratio. The count of p-values less than 0.05 is shown for each coefficient in all models. The number of participants in each simulation and McFadden pseudo r-squared are shown at the bottom of the table for each model.

	E-Mini Futures Contract				Ten-Year Treasury Note Futures				Thirty-Year Treasury Bond Futures			
	Algorithmic-Entry		Manual-Entry		Algorithmic-Entry		Manual-Entry		Algorithmic-Entry		Manual-Entry	
Signal Variable	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05	Average of Hazard Ratio/ p-value/95% C.I.	Count of p-Val < 0.05
Cancel Count	1.0069 0.1363 (1.0006, 1.0133)	305	1.0065 0.2598 (0.9977, 1.0154)	190	1.0105 0.0463 (1.0036, 1.0174)	413	1.0080 0.2337 (0.9977, 1.0183)	203	1.0077 0.2787 (0.9969, 1.0186)	170	1.0039 0.4841 (0.9766, 1.0319)	38
Execution Count	1.0014 0.3158 (0.9974, 1.0053)	135	1.0031 0.2549 (0.9983, 1.0078)	186	1.0011 0.3044 (0.9985, 1.0037)	145	1.0031 0.3150 (0.9974, 1.0088)	139	0.9983 0.4292 (0.9901, 1.0065)	68	0.9895 0.3710 (0.9619, 1.0181)	99
Individual Execution Cnt	0.9980 <0.0001 (0.9976, 0.9984)	500	0.9845 <0.0001 (0.9805, 0.9886)	500	0.9962 <0.0001 (0.9956, 0.9968)	500	0.9703 <0.0001 (0.9611, 0.9797)	500	0.9942 <0.0001 (0.9932, 0,9952)	500	0.8788 <0.0001 (0.8318, 0.9291)	500
New Order Count	0.9958 0.2747 (0.9899, 1.0017)	155	0.9949 0.2938 (0.9866, 1.0032)	152	0.9929 0.1021 (0.9869, 0.9988)	325	0.9941 0.2729 (0.9848, 1.0035)	178	0.9965 0.3777 (0.9866, 1.0064)	94	1.0068 0.4510 (0.9816, 1.0326)	43
Inventory if Negative	1.0007 0.3475 (0.9995, 1.0018)	123	1.0025 0.2174 (0.9997, 1.0053)	230	1.0005 0.3453 (0.9991, 1.0019)	118	0.9992 0.1959 (0.9954, 1.0030)	262	1.0014 0.3431 (0.9966, 1.0063)	116	1.0018 0.3569 (0.9798, 1.0248)	104
Inventory if Positive	0.9996 0.3710 (0.9983, 1.0009)	85	0.9978 0.2727 (0.9944, 1.0012)	200	0.9996 0.3314 (0.9985, 1.0007)	139	1.0005 0.1993 (0.9954, 1.0056)	250	0.9995 0.3540 (0.9957, 1.0032)	79	1.0264 0.3626 (0.9811, 1.0756)	95
Number of participants	378		123		460		86		353		47	
McFadden R-Sqrd	5.4%		23.9%		5.0%		28.2%		6.6%		51.0%	
McFadden Adj. R-Sqrd	5.1%		23.1%		4.8%		26.9%		6.4%		48.2%	

and by algorithmic- or manual-entry participants. These models are also only estimated using proprietary trades; there are no customer-initiated trades in these samples. Covariates measured in the time gap are the number of cancellations, executions, and new orders in the entire market excluding the participant identified in the observation. The participant's cancellations within the gap are also included for the models in Table 2, and the participant's executions within the gap are included for the models in Table 3. All of these covariates are sampled up to three times—at approximately 20%, 50% and 75% of gap length if the gap is sufficiently long and populated—within the gap to measure time dependence. The inventory covariates are measured at the start of the gap for a given participant. These are divided into positive or negative sides with the variable taking a zero value when the sign changes.

These models are estimated using a bootstrap simulation. The simulation includes 500 random samples in which a gap for each participant is drawn once for every sample. The simulation gives equal weight to all participants, which removes the effect caused by only a few participants having many message origination gaps. Both tables show the average (across samples) of the estimated hazard rates for each covariate, the average p-values of the covariate's estimated coefficient and the 95% confidence interval (95% C.I.) of the hazard ratio. The count of p-values less than 0.05 across samples is shown for each coefficient in all models. The number of participants in each simulation, and the McFadden pseudo and adjusted R-squareds are shown at the bottom of each model.

To interpret these results consider the effects of market cancellation counts on new order entry in Table 2. This covariate is measured in units of 10 contracts, as are the other market covariates (execution and new order counts). The hazard rate for the E-mini algorithmic estimates averaged 1.0111 for the algorithmic model with and average p-value of 0.0152, and a 95% confidence interval of 1.0059 to 1.0163. The fact that this covariate is significant in 474 (94.8%) of our samples strongly suggests that this variable is meaningful to participant decisions within the gap. The average hazard rate of 1.0111 implies that if the market cancellation count increases by 10 futures contracts at a given time point then proprietary algorithmic participants are 1.11% more likely to submit a new order in the next interval of time after this event. Because these market cancellations have reduced the order book queue, new orders placed after the cancellation will take less time to execute, *ceteris paribus*, which provides the greater incentive to submit a new order. In the same model, the new order count hazard rate is on average equal to 0.9922, which implies that if the new order volume increases by 10 futures contracts at a given time point, then proprietary algorithmic participants are 0.78% *less* likely to submit a new order in the next interval of time after this event. Again, the behavior of the order book provides a possibly explanation for this finding because more new orders by the market lengthens the expected resting time on the book, which makes an individual participant less likely to submit a new order, *ceteris paribus*.

The market-based results in Table 2 indicate that cancel and new order counts appear as consistently significant covariates for the message origination latency of algorithmic participants. Comparatively, however, the cancel counts of the individual participant are more significant than these market-wide covariates. In all 500 samples, this covariate was significant at less than the 0.0001 level. The estimated hazard rate here implies that participants delay entering new orders

if they have cancelled an order within the gap. This is logical, given that the reason to cancel an order may also be the reason to delay entering a new order.[11]

Except for the E-mini contract, the manual-entry results in Table 2 indicate that these traders only consider their own within gap actions as important for latency calculations. Specifically, if they have canceled an order within a message origination gap for ten-year treasuries, then in the next instant of time they are 12.3% less likely to enter a new order. Interestingly, in Table 3 where we measure the gap as starting from a cancellation message, the individual actions—in this case executions—within the gap is the only significant covariate, with one exception, in these simulations for both algorithmic- and manual-entry participants. That is, the activity of the market after a cancellation is not particularly relevant to whether a new order is submitted, except possibly the cancel count of ten-year treasuries for algorithmic participants.

What is somewhat surprising about the findings for individual-derived covariates is that such variables offer good explanatory power for manual-entry participants. We had expected manual participants to be less consistent in their behavior, leading to more difficulty explaining the variation in manual participant choices. Instead, the McFadden pseudo R-squareds range from 23% to 56% for the manual-entry models. In contrast, the algorithmic-entry models show an R-squared range of 5% to 10.9%, even with more covariates showing statistically significant results. We conclude that the manual-entry participants appear to follow a Markov-like property, where their immediate last action is a very important determinant of the probability of their next action. This property may also be part of an algorithmic-entry strategy, but the relatively low R-squared for those models suggest that we are not capturing enough of the many possible signals, or enough detail in our selected signals, that algorithms have incorporated into their market re-entry strategies.[12]

VI. Conclusion

We have defined many of the types of latency observed in financial markets and offered a model of one form: message origination latency. Our model relies on several assumptions, the most important being that the signals which generate a new order are based on data observed within the market of a specific commodity. We also assumed that the process leading up to a new order decision began after receiving an exit signal from the market. Based on these assumptions, we estimated models of message origination latency with market-based covariates for both algorithmic- and manual-entry proprietary traders. The results suggest that our market-based covariates are better at explaining the variation for manual-entry than algorithmic participants.

[11] This argument may differ by strategy. A market-making program may be more likely to enter new orders after cancelling as it tries to restore two-sided quotes, whereas directional or more aggressive accounts may wait longer.

[12] It is also possible that algorithms work in clock time, which makes event time activity less helpful in explaining variation in algorithmic actions.

VII. References

Kalbfleish, John D. and Ross L. Prentice, 2002, *The Statistical Analysis of Failure Time Data*, New Jersey: John Wiley & Sons.

Kirilenko, A. and G. Lamacie, 2015, Latency and Asset Prices, working paper, MIT Sloan School of Management.

Menkveld, Albert J., 2016, The Economics of High-Frequency Trading: Taking Stock, working paper, Finance group, Vrije Universiteit, Amsterdam.