The Flash Crash: A New Deconstruction*

Eric M. Aldrich†
Department of Economics
University of California, Santa Cruz

Joseph A. Grundfest‡
School of Law
Stanford University

Gregory Laughlin§
Department of Astronomy and Astrophysics
University of California, Santa Cruz

January 25, 2016

Abstract

On May 6, 2010, in the span of a mere four and half minutes, the Dow Jones Industrial Average lost approximately 1,000 points. In the following fifteen minutes it recovered essentially all of its losses. This "Flash Crash" occurred in the absence of fundamental news that could explain the observed price pattern and is generally viewed as the result of endogenous factors related to the complexity of modern equity market trading. We present the first analysis of the entire order book at millisecond granularity, and not just of executed transactions, in an effort to explore the causes of the Flash Crash. We also examine information flows as reflected in a variety of data feeds provided to market participants during the Flash Crash. While assertions relating to causation of the Flash Crash must be accompanied by significant disclaimers, we suggest that it is highly unlikely that, as alleged by the United States Government, Navinder Sarao's spoofing orders, even if illegal, could have caused the Flash Crash, or that the crash was a foreseeable consequence of his spoofing activity. Instead, we find that the explanation offered by the joint CFTC-SEC Staff Report, which relies on prevailing market conditions combined with the introduction of a large equity sell order implemented in a particularly dislocating manner, is consistent with the data. We offer a simulation model that formalizes the process by which large sell orders of the sort observed in the CFTC-SEC Staff Report, combined with prevailing market conditions, could generate a Flash Crash in the absence of fundamental information. Our research also documents the emergence of heretofore unobserved anomalies in market data feeds that correlate very closely with the initiation of and recovery from the Flash Crash. Our analysis of these data feed anomalies is ongoing as we attempt to discern whether they were a symptom of the rapid trading that accompanied the Flash Crash or whether they were causal in the sense that they rationally contributed to traders' decisions to withdraw liquidity and then restore it after the anomalies were resolved.

*This work was supported by the Hellman Fellows Fund.
†Email: ealdrich@ucsc.edu.
‡Email: grundfest@stanford.edu
§Email: glaughli@ucsc.edu

1 Introduction

The "Flash Crash" of May 6, 2010, is unique in the history of American equity markets. In the span of a mere four and a half minutes, from 2:41 p.m. to 2:45:28 p.m., "the broad markets plummeted ... 5-6% to reach intraday lows of 9-10%" below the markets opening price while volumes in US equity, equity derivatives, and equity futures markets spiked (CFTC and SEC, 2010b, p.9). During this period, the Dow Jones Industrial Average "suffered the greatest one hour decline in its history," losing about 1,000 points. (Fox et al., 2015, pp.36-37). In the subsequent fifteen minutes "broad market indices recovered while ... many individual securities and ETFs experienced extreme price fluctuations and traded in a disorderly fashion." (CFTC and SEC, 2010b, p.9) "Accenture, for instance, fell from trading at $39.98 at 2:46 to one cent at 2:49 only to return to $39.51 by 2:50. Apple, on the other hand, at one moment traded for almost $100,000 per share." (Fox et al., 2015, p.37). By 3:00 p.m., "prices of most individual securities significantly recovered and trading resumed in a more orderly fashion." (CFTC and SEC, 2010b, p.9)

An observer measuring the day's activity by simply examining opening and closing prices would be oblivious to the chaos that prevailed for a short period in the late afternoon. There would be no indication that hundreds of billions of dollars of equity market capitalization had vanished and then quickly reappeared, or that this perturbation occurred in the absence of any fundamental news that could explain such a rapid and transitory change in market valuations. The Flash Crash is thus generally viewed as an endogenous event whose dynamics are attributed to the complexity of modern equity market microstructure.

This large, precipitous, and transitory price decline generated significant concern among legislators, regulators, and the investing public. The staffs of the Commodity Futures Trading Commission (CFTC) and the Securities and Exchange Commission (SEC) have attributed the event to unsettled market conditions early in the day, combined with a massive, aggressive E-mini S&P 500 futures sell order initiated by a large mutual fund complex, later identified as Waddell & Reed (CFTC and SEC, 2010b).

More recently, the CFTC and United States Department of Justice (DoJ) expanded the list of causal factors by alleging that Navinder Sarao, a London-based equity futures trader,

engaged in illegal "spoofing" activity that materially contributed to the Flash Crash (CFTC v. Sarao, 2015a,b; USA v. Sarao, 2015b,a). In separate criminal and civil proceedings, Sarao is accused of manipulating prices in the near-month E-mini S&P 500 futures contract by consistently layering the sell side of the order book with large quantities of orders at non-marketable prices, with no intention of allowing the orders to be filled in the event of price shifts. Analyses by experts on behalf of the CFTC and DoJ suggest that large order book imbalances at deep, non-marketable prices have significant effects on subsequent prices (USA v. Sarao, 2015b,a). Commentators, however, are skeptical that the actions of a relatively small trader, such as Sarao, could generate such outsized consequences (Pirrong, 2015; Clearfield and Weatherall, 2015). Indeed, if Sarao's relatively small-scale trading could in fact generate the large-scale effects asserted by the government, modern equity market structures could be viewed as alarmingly fragile.

The Flash Crash raises difficult, policy-relevant questions of causation. As is the case with most market events, the circumstances of the Flash Crash cannot be replicated. Analysts lack access to the specifications of the automated trading algorithms that were active in the markets prior to and during the crash, and cannot replicate the strategies implemented by human traders active during the relevant period. These limitations are compounded by significant identification issues attributable to complex market interactions and to the simultaneous presence of multiple potentially interactive causal factors. In this environment, correlation is easily confused for causation.

With full awareness of these limitations, this paper attempts to address questions regarding the causation of the Flash Crash through a detailed analysis of messaging data related to the E-mini S&P 500 futures contract (CME ticker ES, hereafter "E-mini") and the SPDR S&P 500 exchange traded fund (NYSE Arca ticker SPY, hereafter "SPY") on May 6, 2010, including the full order book. Relatively few academic works have analyzed the Flash Crash, and those studies typically focus exclusively on transaction data without regard to other metrics that describe market conditions prevailing at the time. Kirilenko et al. (2015) and Easley et al. (2011) are primary examples. The joint CFTC and SEC staff reports, CFTC and SEC (2010b) and CFTC and SEC (2010a), provide summary measures of order book statistics in an effort to describe the state of liquidity during the crash, but do

not accurately show the imbalances of which Sarao is being accused of displaying. Menkveld and Yueshen (2015) rely order book data subsampled at 25 millisecond intervals, in conjunction with a proprietary dataset obtained from Waddell & Reed, to investigate the impact of the large E-mini order initiated by Waddell & Reed, to quantify their accrued losses, and to establish that large-scale arbitrage opportunities existed both prior to and after the CME market stop at 2:45:28 p.m. EDT. They do not, however, analyze the full depth of the book at millisecond granularity, which our work suggests is necessary in order to assess causation in such rapidly moving markets.

Our analysis focuses on the E-mini and the SPY and in this sense tracks the analysis of the Joint CFTC and SEC Staff Report, which describes these instruments as the "two most active stock index instruments traded in the electronic futures and equity markets." (CFTC and SEC, 2010b, p.10) The E-mini is viewed as being at the epicenter of the Flash Crash and the SPY is a highly liquid near-equivalent equity market instrument.

This article presents four primary methodological contributions to the literature relating to the analysis of the Flash Crash. First, in contrast to prior analyses, we rely on all order book data to investigate market communication and synchronization at millisecond granularity in the minutes surrounding the crash. This level of granularity adds significant computational complexity, but is, we believe, necessary in the context of a market driven by high frequency trading and speed-of-light messaging between major market centers.

Second, we compute order book imbalances at all price levels provided by the CME and rigorously determine the statistical impact of deep liquidity shifts on subsequent prices for a variety of sample days. Analysis of this sort is necessary in order to study the impact, if any, of the spoofing activity alleged against Sarao.

Third, we present a simulation model that explains the evolution of the Flash Crash in a manner consistent with the observable data and that provides insight into conditions that could lead to future flash crashes. The formal rigor associated with the specification of this simulation model addresses ambiguities that remain unresolved in the CFTC-SEC staff analysis. To be sure, the simulation model presented in this paper is not the only possible explanation of the decline, but it presents a parsimonious model that is consistent with the data and that supports intuitively reasonable, policy-relevant results.

Fourth, we document an anomaly in the time stamps of trades reported to the Consolidated Tape System. This anomaly suggests that increasingly stale prices for the SPY ETF were disseminated to the market and that the inception of this reporting delay correlates strongly with the start of the Flash Crash. When the delay reached 90 seconds, these late reported trades began to be labeled as delayed (in accordance with regulations then in force). Recovery from the Crash began immediately thereafter. Our analysis of these data feed anomalies is ongoing. The objective of our ongoing research is to attempt to discern whether these anomalies were a symptom of the rapid trading that accompanied the Flash Crash or whether they were causal in the sense that they rationally contributed to traders' decisions to withdraw liquidity and then restore it after the anomalies were resolved. This further research also explores whether the 5 second CME trading halt was, in fact, a cause of the recovery, as suggested by some (CFTC and SEC, 2010b; Kirilenko et al., 2015), or whether the updated classification of the data feed anomalies, which occurred shortly after the CME halt, is better viewed as a cause of the recovery. Our current work does not preclude the possibility that the CME halt was a necessary precondition to the correction of the data feed anomalies, in which case the halt can be described as a necessary but insufficient condition for the market recovery.

These analyses support a range of conclusions. In particular, with regard to the allegations against Mr. Sarao, we find that even if we assume that Sarao's trading was a "but-for" cause of the Flash Crash (i.e., but for his presence in the market, the Flash Crash would not have occurred), it was unforeseeable to Mr. Sarao, or to anyone else in the market, that his trading would have this effect. Thus, even if it is true that Sarao engaged in illegal "spoofing" activity, it does not follow that Sarao either intended to cause the Flash Crash, or that he could have foreseen that his conduct would have such an effect.

Our analysis further suggests that Sarao's trading was likely not a "but-for" cause of the crash. Instead, consistent with the analysis of the Joint CFTC SEC Staff Report (CFTC and SEC, 2010a), we find that the Flash Crash is sufficiently explained as the result of the confluence of the unsettled market conditions that prevailed in the hours leading up to the Flash Crash combined with the size and execution strategy of the Waddell & Reed trades. We confirm that offer-side order book imbalances increased substantially in the

hour immediately prior to the crash, but only at price levels deep in the book. This order book structure is consistent with Sarao's spoofing conduct, and the aggregate effect was a heavy sell-side imbalance. Our empirical analysis suggests, however, that there is little to no relationship between deep order book imbalances and subsequent price movements. To the extent that we find mild statistical significance, the effects are minuscule over the relevant time horizons.

Our simulation model suggests that the probability of recurrence of a Flash Crash can be described as a function of the ratio of high-frequency, liquidity-consuming traders to fundamental, non-HFT traders. While the simulation shows how the Efficient Markets Hypothesis (EMH) can fail to hold for short periods after fundamental traders have left the market, prior to their exit, the analysis suggests that Sarao's layering should have had no impact on prices because his orders were sufficiently far from marketable prices so as to not display useful information. Nonetheless, if trading algorithms were searching for imbalance signals deep in the E-mini order book, and if those imbalances contributed to the exit of non-HFT traders, such a mechanism cannot be disproven *ex-post*. At most, we can conclude that Sarao was operating in an extremely complex environment, in which any of the millions of financial market actions on May 6, 2010 (including his own) could have unforeseeably precipitated a critical event and a downward cascade of prices (Clearfield and Weatherall, 2015). In this view, the price decline segment of the Flash Crash can be seen as exhibiting a form of self organized criticality (Bak et al., 1987) in which the market behaved like an equilibrium system near a critical point, and was capable of exhibiting rapid movements, extending from the minimum price tick size all the way to a cutoff established by order-unity fluctuations in the underlying asset price.

We are conducting further analyses to determine whether the anomalies in the off-exchange trades that were reported to FINRA are better viewed as symptoms of a rapidly moving market, or as causal factors. The hypothesis that these data feed anomalies are merely symptomatic would be supported by the finding that the rapid trading during the decline was sufficient to cause some parties to fall behind in their trade reporting practices, and that the delay did not contribute to traders decisions to withhold liquidity from the market. The alternative hypothesis that data feed anomalies were causal would be supported

by finding that, in response to the uncertainty generated by this informational problem, traders rationally withdrew liquidity from the market, and returned that liquidity after the informational problem was resolved. The data may also support some combination of these two hypotheses.

We emphasize that our observations regarding the implications of this data feed anomaly should be viewed as preliminary and tenuous. Additional information that is not in the public domain regarding the nature of the data feeds relied upon by algorithmic traders and the conditions under which algorithms withdraw or supply liquidity could either strengthen or refute this hypothesis. In addition, the relationship (if any) between the CME trading halt and the updated trade classifications of the FINRA data feed remains to be clarified. Thus, further research on market stops, and in particular, on their coordination among exchanges, and the relation to data feed processes, might be especially impactful.

2 Data

While several markets experienced dramatic declines on the afternoon of May 6, 2010, the near-month electronic futures contract for the S&P 500 index (known as the E-mini, commodity ticker symbol ESM0), is widely considered to be the epicenter of the crash. U.S. equity futures prices in Chicago are generally understood to lead cash prices in the U.S. equity markets themselves (Laughlin et al., 2014). Specifically, prices movements of the most liquid equities are highly correlated with the price movements of the near-month E-Mini S&P 500 futures contract, which is traded on the CME's Globex platform and valued at the numerical value of the S&P 500 Stock Index on the contract expiration date. CFTC and SEC (2010a), Kirilenko et al. (2015) and Menkveld and Yueshen (2015), suggest that a large market sell order of 75,000 E-mini contracts, initiated with an automated execution algorithm at 2:32 p.m. Eastern time, created an imbalance that triggered subsequent, market-wide events. More recently, the affidavit provided by Professor Terry Hendershott, accompanying the CFTC civil complaint against Navinder Sarao, suggests that imbalances caused by passive E-mini sell orders linked to Sarao's account may also have been responsible (CFTC v. Sarao, 2015b).

Our analysis rests on three sources of data. We emphasize that for each data set described below, we use the full sample of quotes and/or transactions, and not a throttled or subsampled version of the original data.

The first source is market depth data for the E-Mini S&P 500 Futures contract purchased from the Chicago Mercantile Exchange . These data are recorded and time stamped at the Globex matching engine, currently located in Aurora, Illinois (longitude -88.24° W, latitude 41.80 deg N). At the time of the Flash Crash on May 6, 2010, the matching engine was located at 350 E. Cermak Road in Chicago (longitude -87.62° W, latitude 41.85° N), and the relevant near-month contract was the ESM0, with expiry in June 2010. Session data are written to ASCII files using the FIX specification. Level-2 order book activity to a price depth of 10 levels on both the bid and the offer side of the order book is captured, along with trade records and other information relevant to recreating the trading session. All order book events are time-stamped to millisecond precision, with time signals propagated from GPS receivers. Events that occur within a single millisecond are time ordered. The E-mini contract trades on the March quarterly cycle (March, June, September, and December) and expires on the third Friday of the contract month. On the "roll date", eight days prior to expiry, both liquidity and price formation shift to the contract with the next-closest expiry date. During the period covered in this analysis, several million E-mini contracts were generally traded each day, corresponding to notional dollar volumes that often exceeded $200 billion.

The second source of data is the Nasdaq TotalView-ITCH historical data feed for symbol SPY (State Street Advisors S&P 500 ETF), recorded at the Nasdaq-OMX matching engine currently located in Carteret, New Jersey (longitude -74.25° W, latitude 40.58° N). These data are composed of a series of binary number-format messages employing the ITCH 4.1 specification and encompass messaging information for all displayable orders in the Nasdaq execution system. Messages are time-stamped to nanosecond precision, and, like the CME data, rely on GPS time stamps.

Finally, we obtained SPY transaction data for all participating Consolidated Tape System (CTS) exchanges from the New York Stock Exchange (NYSE) Daily Trades and Quotes Service. These data include traded price and volume information time-stamped to millisec-

ond accuracy for all public exchanges that matched orders for SPY. While the NYSE data present a lower-resolution (millisecond vs. nanosecond) view than the Nasdaq source, they permit measurement of broader equity market activity that encompasses transactions on all participating public exchanges.

3 Analysis of Messaging and Imbalances

3.1 Messaging

Figure 1 shows the price trajectory of the E-mini between 2:15 p.m. and 3:15 p.m. EDT on May 6, 2010 (blue line). Although prices had been steadily falling throughout the day, an abrupt decline commenced at approximately 2:42 p.m. and continued until 2:45:28 p.m. At that point, an automated stop logic price protection event occurred at the CME, placing the E-mini in a reserve state for 5 seconds. This trading halt lasted from 2:45:28 to 2:45:33 p.m. EDT. During the 3.5-minute period before the halt, the contract lost nearly 5% of its value. Almost immediately after the halt, prices rebounded, experiencing a full recovery to their pre-crash levels by 2:55 p.m.

To establish context, we compare events on the day of the Flash Crash with events on August 9, 2011, the day on which the CME experienced its all-time highest daily messaging volume. Aside from messaging traffic, there are further similarities in overall stressed, geopolitically induced conditions on both days. Figure 1 shows that during the 2:15 – 3:15 p.m. interval, the price trajectory was very similar on both days, with the notable exception of the rapid linear decline in price followed by the volatile price recovery on May 6, 2010.

The overall similarity between the two days is further underscored by the cumulative rate at which trading occurred through the day, as illustrated in Figure 2, which displays cumulative traded volume. The Flash crash occurred during the five minutes prior to 2:45 p.m. EDT, and consistent with that timing, the trading rate increases steeply around that time for the May 6, 2010 cumulation.

The data rates are also similar. Figure 3 charts the messaging rates in megabits per second for the 1:00 – 4:00 p.m. period on both days. There are no particularly dramatic

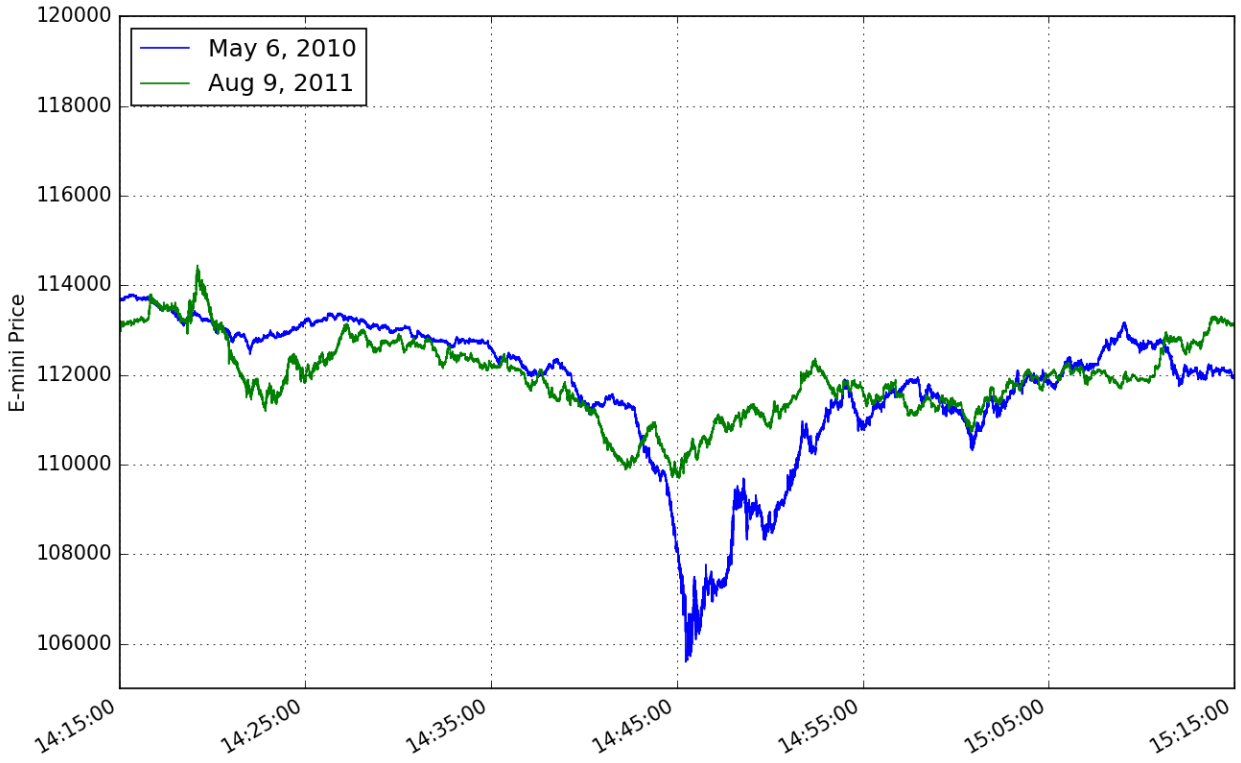

Figure 1: Comparison of E-mini S&P500 Near-month futures contract prices on May 6, 2010 (blue), and August 9, 2011 (green). Prices are shown from 2:15 p.m. through 3:15 p.m., EDT.

differences. In fact, the most prominent feature across both days is a sustained spike in messaging that occurred around 2:20 p.m. on Aug 9. Messaging during and preceding the Flash Crash was also high, but not as dramatic.

The physical separation between the futures and equities markets permits a quantitative assessment of the degree to which price formation occurs at the futures exchange in Chicago, and allows monitoring of the rate of inter-market messaging. By correlating order book activity in the equity markets with traded E-mini upticks and downticks, we can evaluate the propagation of information between the exchanges during any particular interval. We adopt the following procedures, which employ the CME and Nasdaq data described above, and are described in detail in Laughlin et al. (2014).

We first step through the CME trade and quote data that falls within a specified period when both the CME and the equity exchanges are trading. At the end of each millisecond, we screen for the occurrence of near-month E-mini futures trades in which the most recent

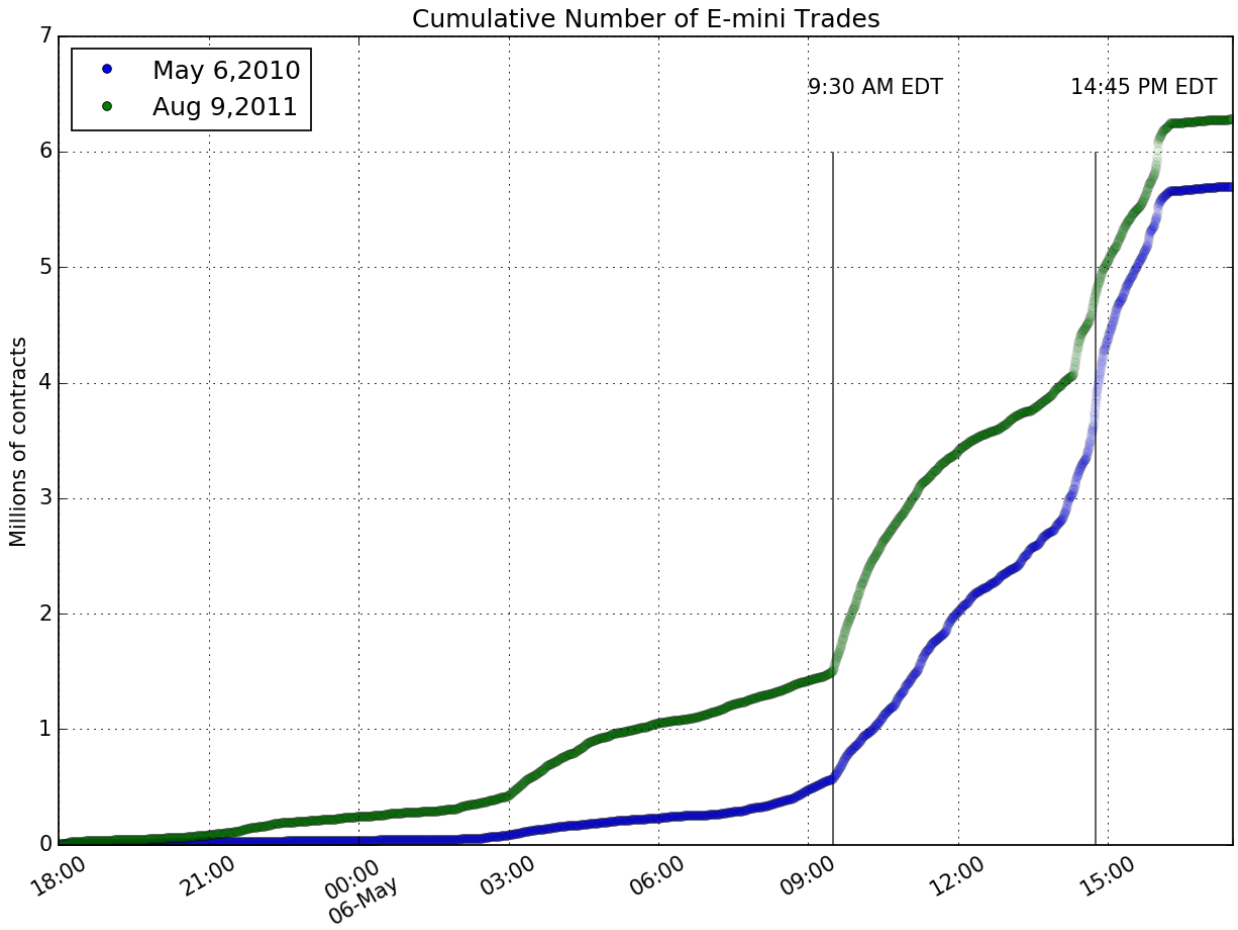

Figure 2: Comparison of the cumulative number of E-mini contracts traded for May 6, 2010 (blue) and August 9, 2011 (green). The vertical line at 2:45 p.m. marks the beginning of the recovery period from May 6, 2010's lowest intra-day traded price.

traded price at the end of a millisecond interval (which we refer to as the "in-force" trade for a given millisecond) exhibits an increase in price over the most recent in-force trade from a previous millisecond.

When a millisecond interval that ends with a price-increasing in-force CME trade is located, our algorithm examines the corresponding Nasdaq data for correlated activity associated with the SPY. This ETF has very high liquidity, and is designed to closely track the S&P 500 index. In each of the N millisecond-long intervals prior to, and in each of the N millisecond-long intervals following the CME price-increasing trade,we calculate the net number of shares that have been added to the bid side of the SPY limit order book at

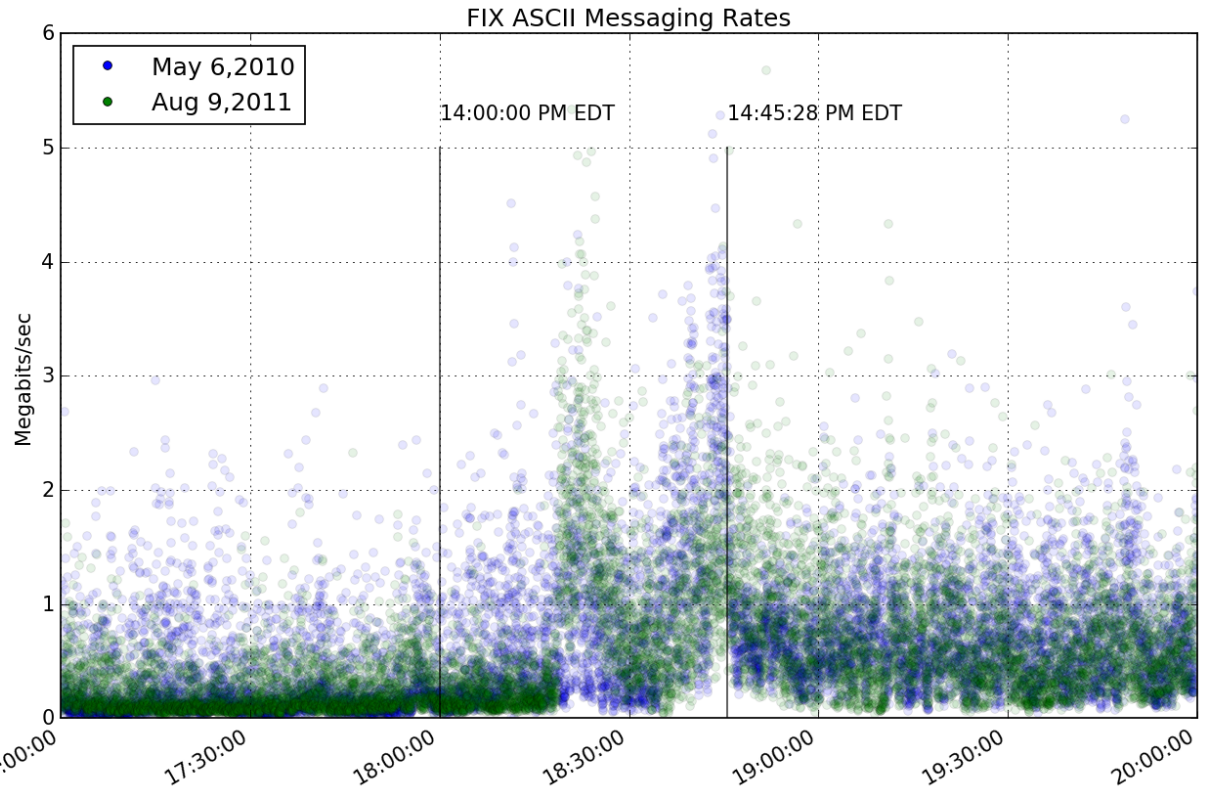

Figure 3: Data rates for the E-mini on May 6, 2010 (blue) and August 9, 2011 (green).

the three price levels corresponding to (i) the in-force Nasdaq exchange-traded SPY price at the beginning of the millisecond-long interval ("the last Nasdaq in-force price"), (ii) the last Nasdaq exchange-traded in-force SPY price + \$0.01 , and (iii) the last Nasdaq-traded in-force price - \$0.01. In addition, in each of the same $2N$ millisecond-long bins surrounding the CME event, we also calculate the net number of shares that have been removed from the three levels of the ask side of the SPY limit order book at prices corresponding to the last Nasdaq-traded SPY price and that price $\pm$ \$0.01. We then add δ_l = (added+removed) to an array that maintains cumulative sums of these deltas as a function of lag (from -N milliseconds to +N milliseconds).

The procedure is also followed for price-*decreasing* in-force trades observed in the near-month E-mini contract. In the case of these declines, however, we add $-1 \times \delta_l$ to the array that maintains the cumulative sums. This facilitates the combination of both price increases and price decreases into a single estimator, which, when divided by the total number of price-

changing E-mini trades constitutes our average "order book response" for a given interval of time.

Figure 4 depicts our estimate of δ_l for several time spans on May 6, 2010: (1) the entire day, (2) the 8 minutes prior to the CME trading halt, (3) the 8 minutes subsequent to the trading halt, (4) the 1 minute prior to the trading halt, (5) and the 1 minute subsequent to the trading halt. The panels corresponding to periods (1), (2) and (4) depict behavior that is typical of the CME/Nasdaq markets (the increasing variance of the estimator across panels is due to the decreasing sample size) – Nasdaq liquidity predictably shifts following CME price changes with a minimum lag ranging from 7 to 8 milliseconds on May 6th 2010, and a minimum lag ranging from 4 to 7 milliseconds on August 9th, 2011. The year-on-year decrease in inter-market latency can be attributed to the appearance of Spread Networks in the late summer of 2010, and to the emergence of line-of-sight microwave networks connecting Chicago and New Jersey during the first half of 2011. The CME/Nasdaq liquidity response is similarly documented in Laughlin et al. (2014) and is a byproduct of information being compounded into futures prices prior to equities. Notably, the panels corresponding to periods (3) and (5) depict no Nasdaq liquidity response on the day of the Flash Crash. This is either a result of transmission breakdown due to high messaging rates, or a choice of market participants not to coordinate actions across markets immediately after the halt. In either case, the liquidity responses in the 8 minutes following the E-mini trading halt are partially consistent with the arbitrage breakdown analysis of Menkveld and Yueshen (2015).

Figure 5 extends the analysis by computing the δ_l over non-overlapping 2-minute intervals between 2:37:28 p.m. and 3:01:33 p.m. on May 6, 2010. The figure displays cumulative values of δ_l over the span of 100 milliseconds following E-mini price changing events and uses three colors to sort responses into distinct periods: the 8 minutes prior to the E-mini halt (red), the 8 minutes after the E-mini halt (green) and the 8 minutes between 2:53:32 p.m. and 3:01:33 p.m (blue). We selected 2:53:33 p.m. as a boundary time, following the analysis of Menkveld and Yueshen, who estimate this as the time at which inter-market arbitrage opportunities disappeared (Menkveld and Yueshen, 2015). Within color group, transparency indicates responses that occurred earlier in the period. The cumulative responses support the findings of Figure 4. In particular, communication was well established and Nasdaq

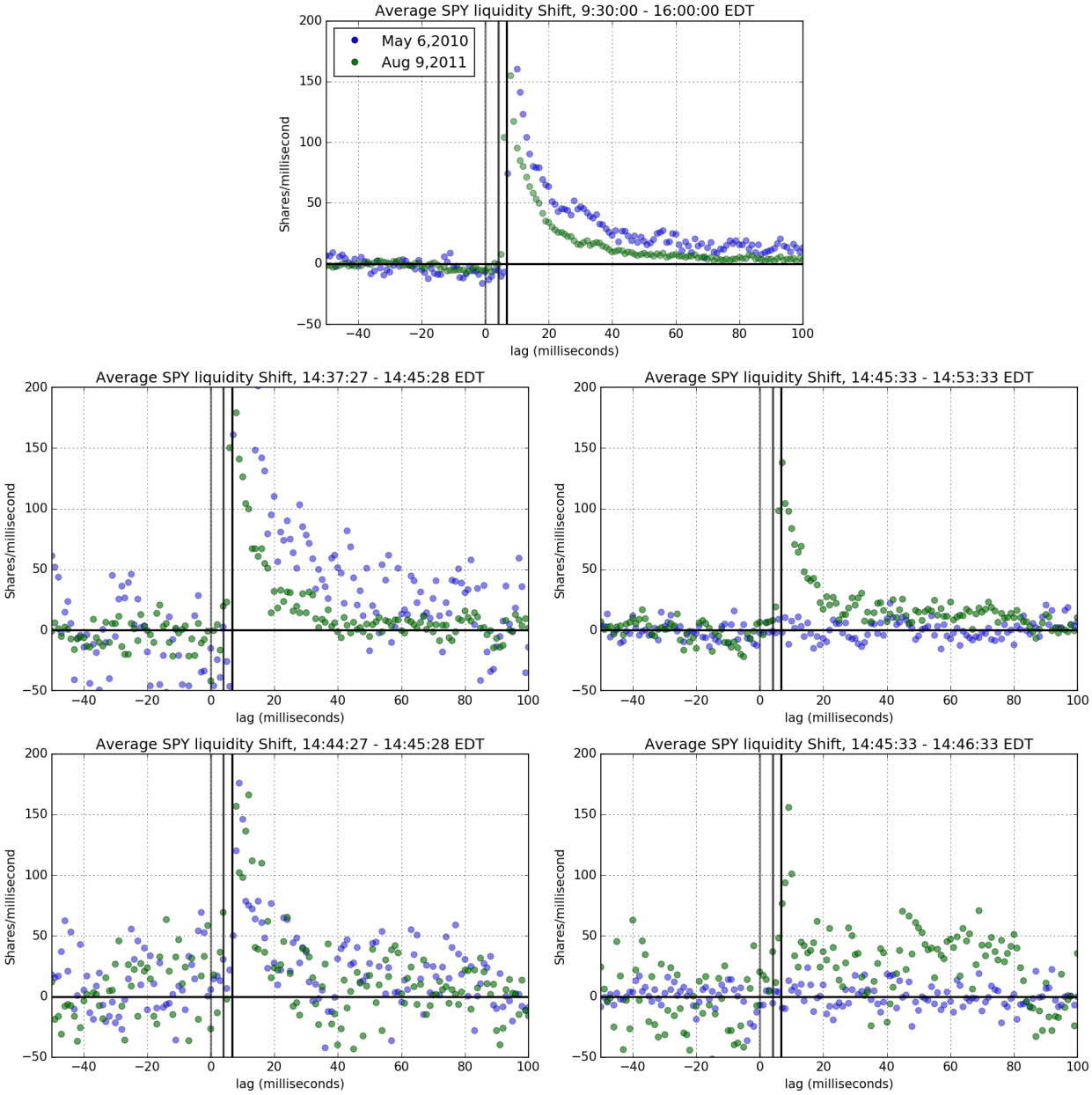

Figure 4: Nasdaq SPY order book response to CME ESM0 price changing trades, evaluated for May 6, 2010 (blue) and August 9, 2011 (green). The speed-of-light travel time between the two locations on the great circle was 3.77 ms on May 6th 2010, and 3.93 ms on August 9, 2011, due to the move of the CME match from 350 Cermak west to Aurora, IL. The three vertical lines represent $t = 0$, $t = 4.0$ and $t = 6.65$ ms. The latter represents the one-way latency provided by Spread Networks optical fiber cable that links the two markets.

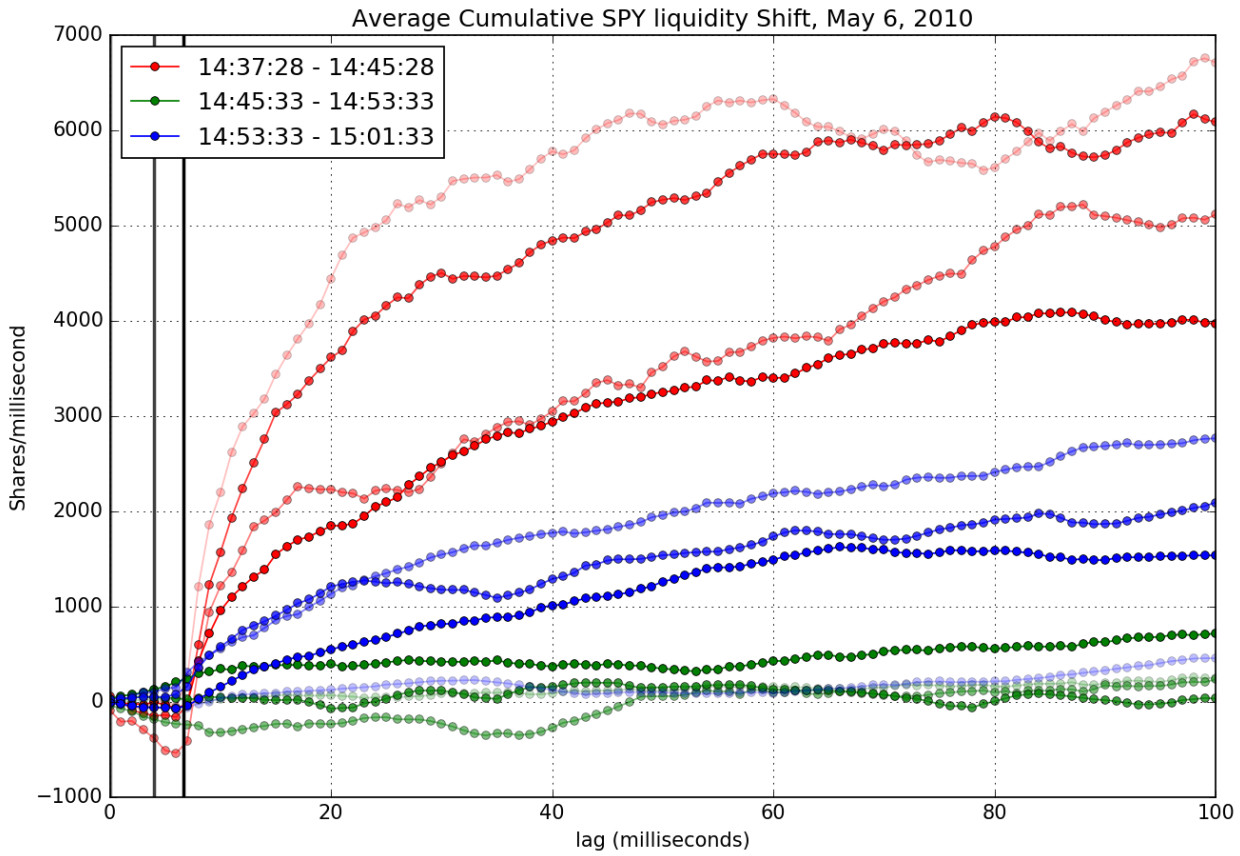

Figure 5: Nasdaq SPY cumulative order book response evaluated for May 6, 2010 at two-minute intervals during the 16 minutes surrounding the CME market stop. Transparency is an indication of earlier times within the indicated period. The three vertical lines represent $t = 0$, $t = 4.0$ and $t = 6.65$ ms.

liquidity was predictably responding to CME events in the entire period leading up to the E-mini halt. Further, during the 8 minutes following the halt, there appears to be no correlation between Nasdaq liquidity shifts and CME price events. Subsequently, the typical relationship is restored, but interestingly it does not perfectly align with the restoration of inter-market pricing estimated by Menkveld and Yueshen (2015). Rather, CME/Nasdaq correlation was not restored until roughly 2:55 p.m. It is important to emphasize that CME/Nasdaq communications were fully operational in the entire period leading up to 2:45:28, despite the fact that inter-market arbitrage had already broken down (at roughly 2:44:27, according to Menkveld and Yueshen (2015)). Thus, our measure of communication correlation across markets is not a pure surrogate for the broken prices, as it does not fully

overlap with the inter-market arbitrage period.

The foregoing analysis suggests that the CME stop logic event may not have been fully beneficial. While the rebound of market prices is typically attributed to the CME halt, both our analysis and that of Menkveld and Yueshen (2015) suggest that prices and communication between markets were not restored for a prolonged period of time.

Indeed, shortly after 2:46:30 p.m., during a period of generally upward price recovery, the largest divergences in price between the E-mini and the SPY occurred. For a period of roughly three minutes, a basis of up to five or more index points existed between SPY and the E-mini, with traders consistently paying a premium for SPY. We conjecture that these conditions arose and were maintained as a consequence of internal risk limits experience by individual HFT participants.

All market makers, at both the futures and the equity exchanges, have pre-established position limits in place with their clearing firms. If the market prices experience a sufficiently large one-way movement, arbitrageurs can enter into a large number of nominally profitable trades while ending up with large gross exposures to individual instruments at individual exchanges. Our foregoing high-frequency correlation analysis demonstrates that the CME unambiguously led the sharp linearly downward price movement prior to the CME trading halt. On the balance, during this period of rapid market decline, arbitrageurs could have repeatedly bought the E-mini and sold SPY (or equivalent baskets of equities that act as proxies for the index). In normal markets a two-sided flow emerges from point volatility which acts to keep such inventories low. In the Flash Crash event, however, the occurrence of the sustained and unprecedented one-way downward move, could have eventually been problematic for arbitrageurs. It would have thwarted their ability to unwind or reduce risk, especially if they became more aggressive as the cash-futures basis deviated from fair value. It appears plausible that the ensuing, and extended period of outright arbitrage opportunities arose because an insufficient number of participants had the margin to take advantage of these opportunities. Several minutes were required for new entrants to the market to arrive, eventually allowing disciplined, cross-exchange pricing efficiency to resume.

3.2 Imbalances

We compute a measure of order book imbalance that is level free. At any point in time we measure the imbalance ratio, IR_t, as

$$IR_t = \frac{N_{t,offer}}{N_{t,bid}}, \tag{1}$$

where $N_{t,offer}$ is the number of contracts on offer and $N_{t,bid}$ is the number of contracts on bid. The ratio can be computed for bids and offers at any single price level, but can also be computed for the aggregated numbers of shares across multiple price levels. Since this ratio is asymmetrically bounded below by zero, we employ the base 10 logarithm as our value of interest: $ir_t = \log_{10}(IR_t)$. Hence, a balanced order book, with an equal number of contracts on bid and offer would result in $ir_t = 0$. Likewise, $ir_t = 1$ indicates 10 times as many contracts on offer as on bid, an $ir_t = -1$ indicates 10 times as many on bid as on offer.

Figure 6 depicts the log imbalance ratio at 1-second intervals over the entire day on May 6, 2010 (blue points), aggregating contracts across all 10 price levels of the order book. The same measure is depicted for the highest CME volume day for the E-mini, August 9, 2011 (green points). As a visual reference, we have added 1-minute exponentially-weighted moving-average lines for each series. The first vertical red line marks the initial steep decline of the E-mini at 2:41 p.m. and the second marks the 5-second trading halt at 2:45:28, instigated by the CME Globex stop logic functionality. The figure demonstrates that the order imbalance ratios are quite similar early in the day, but that the values for May 6 become consistently elevated in the afternoon, and especially elevated some time after 2:00 p.m. After the trading halt, the May 6 ratios return to pre-crash levels before reversing sign, although in a much more diffuse manner.

Figure 7 depicts the same ratios for the one hour surrounding the Flash Crash and also disaggregates according to the 10 price levels in the order book at each second in time. The figure shows that the imbalance ratio was not uniformly elevated across price levels of the order book, but looked fairly balanced for levels 1 through 4. This is roughly consistent with the analysis of Hendershott in the CFTC complaint, who described Navinder Sarao's layering algorithm as targeting levels 4 through 7 of the sell-side of the E-mini order book (CFTC v. Sarao, 2015b). While we also observe elevated ratios for even deeper levels, this

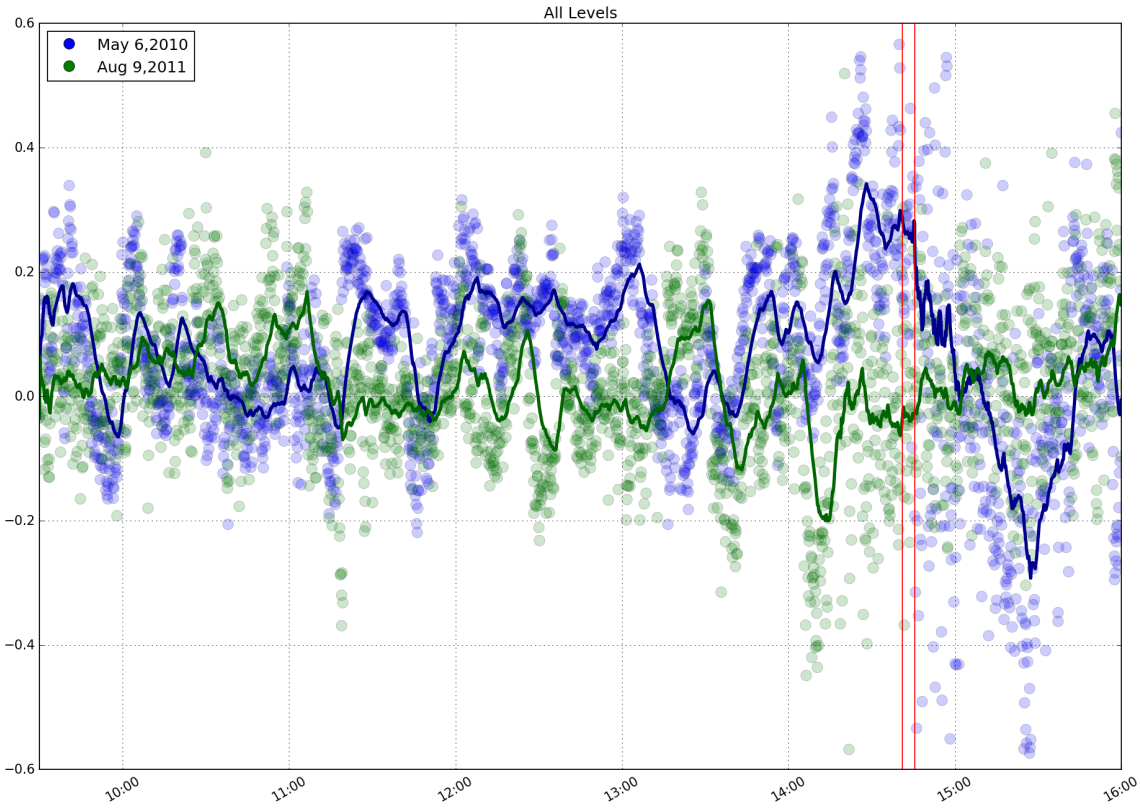

Figure 6: Log imbalance ratios for all levels of the order book on May 6, 2010 and August 9, 2011.

could be an artifact of stale orders that moved upward in the book as the market declined. The upshot is that the imbalance ratios look elevated and very similar across price levels 5 through 10, and that the resulting aggregate imbalance looks almost identical to those of the deep levels.

To understand the historical context of these measures, Figure 8 shows only the 1-minute moving average lines for the log order imbalance ratios for all trading days between April 5, 2010 and May 7, 2010. The coloring of the lines progresses from green to blue over the date range under consideration, with the lightest green corresponding to April 5, 2010 and the darkest blue corresponding to May 7, 2010. In addition, the moving averages of imbalance ratios for May 6, 2010 are depicted as a bold blue line. It is immediately obvious from the figure that the order imbalance ratio was elevated far beyond typical historical levels in the 30 minutes leading up to the crash. However, it is also clear that there were a variety of days, primarily toward the end of April and the beginning of May, that displayed similar

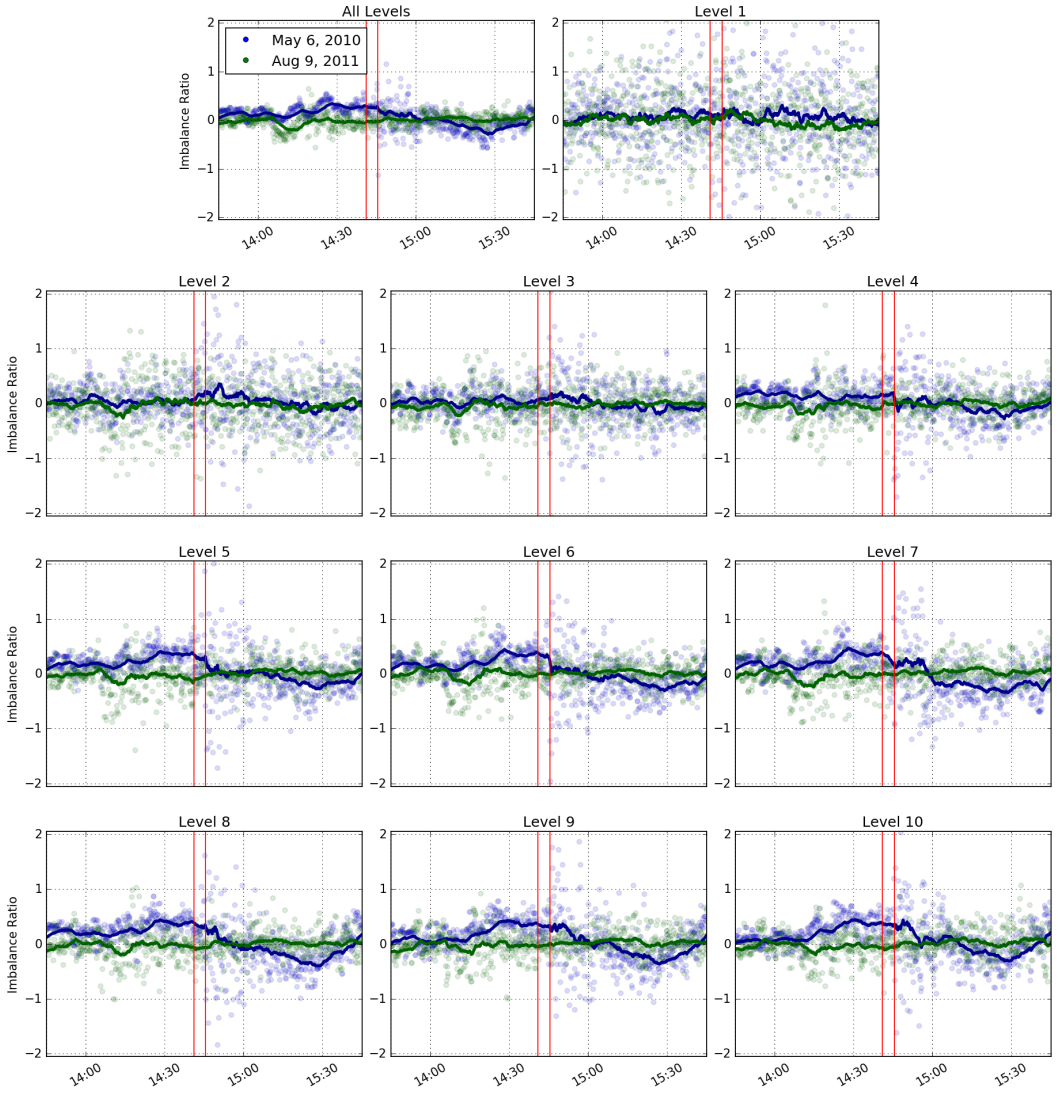

Figure 7: Log imbalance ratios between 2:15 p.m. and 3:15 p.m. on May 6, 2010 and August 9, 2011.

elevated ratios. As noted in Hendershott's affidavit, this is consistent with the fact that Sarao's layering algorithm was placing sell-side orders deep in the order book during the latter part of April.

Of primary interest is the question of causation and whether the large order imbalances on May 6, 2010 predict subsequent returns. To shed light on this question, we regress 5-second returns, r_t, for the near-month (June) ES contract on prior 5-second order book imbalance differences, $d_t^l = N_{t,\text{offer}}^l - N_{t,\text{bid}}^l$:

$$r_t = \beta_0 + \beta_1 d_{t-1}^l + \varepsilon_t, \tag{2}$$

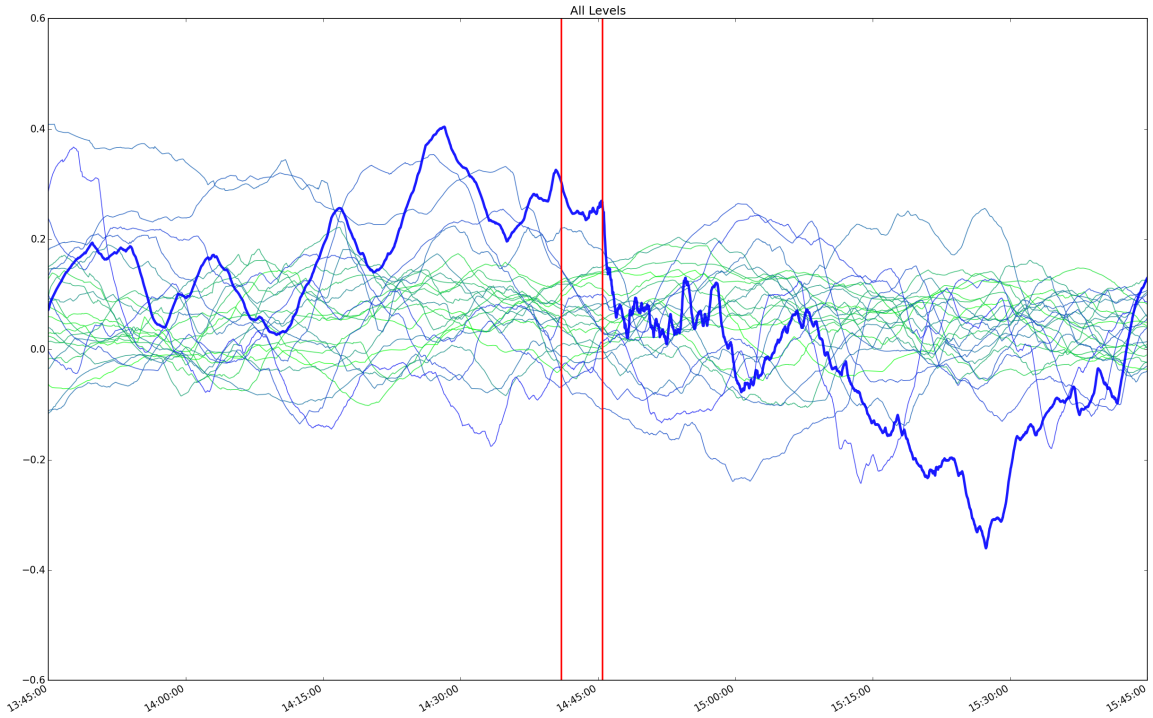

Figure 8: 1-minute exponentially-weighted moving-averages of log imbalance ratios between 1:45 p.m. and 3:45 p.m. for April 5, 2010 to May 7, 2010. The line colors progress from green to blue over the sample days, with the bold blue line representing May 6, 2010.

where l represents a price level or group of price levels. We also fit the regression using imbalance ratios at corresponding levels, ir_t^l, as the regressor, and while the results were qualitatively almost identical, we only report results for imbalance differences since this more closely aligns with the analysis performed by Hendershott (CFTC v. Sarao, 2015b). We fit the regressions for a collection of randomly selected days (2 per month, including May 6, 2010 and Aug 9, 2011) between April 2010 and August 2011 (34 total days), using imbalance differences at various levels of the book as the univariate regressor. Panel (a) of Table 1 reports the results.

The specific days in our sample include 4/6, 4/23, 5/6, 5/27, 6/7, 6/18, 7/14, 7/26, 8/9, 8/20, 9/9, 9/26, 10/13, 10/29, 11/4, 11/22, 12/8, 12/21 2010 and 1/13, 1/31, 2/15, 2/25, 3/10, 3/29, 4/8, 4/19, 5/6, 5/25, 6/13, 6/29 7/7, 7/28, 8/9, 8/25 2011. Panel (b) reports identical regression results, excluding May 6, 2010 from the sample. In all cases, the regressions are scaled so that returns are expressed in basis points and the independent

(a) Univariate Regressions, All Days														
Level	1	2	3	4	5	6	7	8	9	10	2–3	4–7	8–10	All
Coef.	0.0647***	0.0396**	-0.0162	-0.0233	-0.0342**	-0.0191	-0.0147	-0.0112	-0.0218	-0.0102	0.00724	-0.0135**	-0.00985	-0.00262
S.E.	0.018	0.0177	0.0167	0.0165	0.0168	0.017	0.0173	0.0182	0.018	0.0186	0.0103	0.00647	0.0087	0.00378
(b) Univariate Regressions, All Days w/o Flash Crash														
Level	1	2	3	4	5	6	7	8	9	10	2–3	4–7	8–10	All
Coef.	0.0698***	0.0387**	-0.0118	-0.0178	-0.0217	-0.00944	-0.00758	-0.005	-0.0156	-0.00634	0.00867	-0.00839	-0.00618	0.000288
S.E.	0.0163	0.0161	0.0152	0.0151	0.0154	0.0155	0.0158	0.0167	0.0165	0.017	0.00937	0.00591	0.00798	0.00346

*, **, and *** denote significance at the 10%, 5% and 1% levels, respectively.

Table 1: Univariate regression results. Coefficients and standard errors are reported for univariate regressions of 5-second returns on order book imbalance differences for the previous 5 seconds. Imbalance differences are measured by aggregating orders at each of the individual 10 order book levels, as well as levels 2 – 3, 4 – 7, 8 – 10, and all levels. The second and third columns report results for a selection of 34 days, including the May 6, 2010. The fourth and fifth columns report results for the same sample of days, excluding the May 6, 2010.

variable is expressed in units of 2400 E-mini contracts (the typical size of layering strategy employed by Sarao). Interestingly, for both samples, the coefficients for levels 1 and 2 (individually) and levels 2–3 (combined) demonstrate a positive effect of offer-side order imbalances on subsequent returns. This pattern is consistent with the existence of transitory jumps in imbalance ratios that occur when levels at the inside of the book are added or deleted. That is, levels 2 and 3 of the order book (on both bid and offer) typically display substantially more depth than the inside levels. Thus, when deletions occur at the inside levels, imbalance differences are momentarily computed as the difference between (deep) level 2 and (thin) level 1. Due to the reversive nature of deletions and additions at the inside of the order book (similar to bid/offer bounce), such contrarian imbalance signals are not uncommon. In contrast, deeper levels are more stable and are not subject to the same reversive effects. This observation is corroborated in Table 1 where all coefficients that exclude levels 1 and 2 are negative.

The regression for All Days and All Levels most closely corresponds to the decile analysis performed by Hendershott, and while the sign and magnitude of the coefficient are analogous to his results, we do not find the relationship to be statistically significant (CFTC v. Sarao, 2015b, see Exhibit 3). Further, when excluding the day of the Flash Crash, the sign changes

but remains statistically insignificant. Intriguingly, the imbalance differences at levels 4 – 7 display significance only when the day of the Flash Crash is included in the sample. In general, including the Flash Crash in the data sample weakly improves the significance and magnitude of the coefficients.

Setting aside questions of direction and significance, the magnitudes of the regression coefficients are extremely small in all cases. As noted above, the criminal complaint against Mr. Sarao alleges that his layering algorithm most commonly employed a strategy of adding 600 contracts to levels 4 – 7 of the E-mini order book, resulting in a 2400 contract increase in sell-side liquidity (USA v. Sarao, 2015a,b). According to the regression estimates, such an increase in liquidity at levels 4 – 7 has at most a -0.0135 basis point effect on subsequent returns. This is drastically smaller than the 500 basis point loss that occurred in the minutes prior to 2:45:28 EDT on May 6, 2010.

To account for causal relationships among simultaneous, interrelated variables, we augment the univariate regression analysis with a structural vector autoregression (SVAR) of the form

$$\begin{bmatrix} 1 & -\alpha_{1,2} & -\alpha_{1,3} & -\alpha_{1,4} & -\alpha_{1,5} & -\alpha_{1,6} \\ 0 & 1 & -\alpha_{2,3} & -\alpha_{2,4} & -\alpha_{2,5} & -\alpha_{2,6} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_t \\ of_t \\ d_t^1 \\ d_t^{23} \\ d_t^{4567} \\ d_t^{8910} \end{bmatrix} = \sum_{j=1}^{5} B_j \begin{bmatrix} r_{t-j} \\ of_{t-j} \\ d_{t-j}^1 \\ d_{t-j}^{23} \\ d_{t-j}^{4567} \\ d_{t-j}^{8910} \end{bmatrix} + \begin{bmatrix} u_{r,t} \\ u_{of,t} \\ u_{d^1,t} \\ u_{d^{23},t} \\ u_{d^{4567},t} \\ u_{d^{8910},t} \end{bmatrix}, \tag{3}$$

where of_t represents 5-second order flow, which is a signed measure of volume, defined as the total number of contracts traded on the offer minus total number of contracts traded on the bid. The contemporaneous relationship of returns, order flow and imbalances loosely follows the specification of Fleming et al. (2014) and can be construed as a constrained version of the SVAR model described by Hendershott (CFTC v. Sarao, 2015b).

Panels (a) and (b) of Table 2 report estimated cumulative impulse response functions for the data sample, including and excluding the day of the Flash Crash. In each case, we subject the estimated SVAR to a 2400 contract sell-side impulse at a specific group of

levels. We report the impulse responses for 25 periods, which corresponds to a total time interval of just over 2 minutes. Both sets of results show that a sudden increase in sell-side orders at levels $1-3$ of the order book has an initial positive impact on returns, but then reverses direction for subsequent periods, although with a net positive cumulative effect. This contrarian impact is analogous to the level 1 and 2 univariate regression coefficients reported above. Of particular importance is the fact that the impulse response for levels $4-7$ exhibit a negative cumulative impact on subsequent returns, but the statistical significance is diminished when the day of the Flash Crash is excluded from the sample. In all cases, as with the univariate regression results, the magnitude of the impulse responses are very small, amounting to more more than 2/100 of a basis point, but typically far smaller.

As a whole, these results suggest that even if Sarao's layering activity had a statistically significant effect on subsequent returns, the magnitude of the effect was extremely small. Further, the results reported in Tables 1 and 2 indicate that the single day of the Flash Crash is a major contributor to statistical significance. This pattern in the data could easily arise if Sarao's algorithm was aggressively engaging in non-market-moving behavior on a day when other market factors caused a rapid decline in prices. Correlation would then be incorrectly confused with causation.

4 Simulation

To highlight and formalize the potential mechanisms that could have caused the Flash Crash, we present a stylized model of market dynamics and simulate price trajectories under a variety of conditions.

Consider an environment populated by three agents and a single asset. Two of the agents, A and B, are characterized as high-frequency traders, deploying similar market making and trading strategies. The third, agent M, represents all other types of market participants (fundamental traders, arbitrageurs, noise traders, non-high-frequency market makers, etc.) and deploys a very different strategy relative to the other agents. Suppose that Agent M follows the strategy outlined in Algorithm 1 (displayed below). That is, with some probability, p_{trade}, she aggressively trades a single contract and with probability $p_{quote} =$

(a) Structural VAR Impulse Responses, All Days

	Level 1		Levels 2–3		Levels 4–7		Levels 8-10	
Period	Coefficient	Period	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
1	0.000965***	0.00023	0.000932***	0.000218	-0.000107	0.000245	-1.13e-07	0.00026
2	0.000981***	0.000299	0.000829***	0.000283	-0.000518*	0.000318	-0.0000755	0.000338
3	0.000859***	0.000358	0.000203	0.000339	-0.00112***	0.000381	-0.000239	0.000403
4	0.00059*	0.000405	-0.000133	0.000379	-0.0013***	0.000429	-0.00105**	0.00045
5	0.000811**	0.000417	-0.0000180	0.000349	-0.00129***	0.00038	-0.000776**	0.00039
10	0.000774*	0.000588	-0.000436	0.000507	-0.00135***	0.000467	-0.000959**	0.000507
15	0.000795	0.000651	-0.000771	0.000643	-0.00151***	0.000578	-0.00111**	0.00065
20	0.000806	0.000671	-0.00102*	0.000745	-0.00169***	0.000691	-0.00125*	0.000789
25	0.000815	0.000677	-0.00122*	0.000821	-0.00188***	0.000799	-0.00138*	0.000917

(b) Structural VAR Impulse Responses, All Days w/o Flash Crash

	Level 1		Levels 2–3		Levels 4–7		Levels 8-10	
Period	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
1	0.000941***	0.000209	0.000965***	0.0002	0.0000369	0.000226	-0.0000911	0.000241
2	0.00102***	0.000267	0.000848***	0.000255	-0.00034	0.000288	-0.000126	0.000307
3	0.000888***	0.000313	0.00024	0.000298	-0.00068**	0.000338	0.0000135	0.000359
4	0.000745**	0.00035	-0.0000661	0.00033	-0.000886***	0.000376	-0.000559*	0.000397
5	0.000892***	0.000353	-0.000132	0.000293	-0.000955***	0.00032	-0.000395	0.000331
10	0.000785*	0.000485	-0.000432	0.000421	-0.000911***	0.000389	-0.000527	0.000426
15	0.000773*	0.000533	-0.000674	0.000531	-0.000959**	0.00048	-0.000632	0.000544
20	0.000772*	0.000548	-0.000853*	0.000613	-0.00102**	0.000573	-0.000721	0.000659
25	0.000773*	0.000554	-0.000986*	0.000675	-0.0011**	0.000661	-0.000800	0.000765

*, **, and *** denote significance at the 10%, 5% and 1% levels, respectively.

Table 2: Structural VAR results. Panel (a) reports cumulative impulse response functions of the estimated structural VAR specified in Equation (3) to a 2400 contract impulse on the sell side of levels 1, 2 – 3, 4 – 7 and 8 – 10 for a selection of 34 days, including the May 6, 2010. Panel (b) reports the same results for the same selection of days, excluding the May 6, 2010.

$1 - p_{trade}$ she passively adds an order to the order book. If trading, she aggressively sells with probability $p_{tradeBid}$ and aggressively buys with probability $1 - p_{tradeBid}$. Alternatively, if adding an order, she adds to the bid side of the book with probability $p_{quoteBid}$ and to the offer side with probability $1 - p_{quoteBid}$. When adding a passive order, she adds to level $x \in \{1, 2, \ldots, n_{levels}\}$, where the levels represent a discrete set of prices separated by the minimum price increment ξ. The level, x, is selected according to a discrete probability distribution, which we parameterize as a geometric distribution with some probability p,

which is truncated to have finite support at n_{levels}. The probability p is selected so that there is substantial depth at prices close to the top of book.

Algorithm 1 Non-HFT Market Participant

Draw independent uniform random variates $u_{trade}, u_{bid} \sim U(0,1)$.
if $u_{trade} < p_{trade}$ **then**
    if $u_{bid} < p_{tradeBid}$ **then**
        Aggressively sell a single contract at the best bid.
    else
        Aggressively buy a single contract at the best offer.
    end if
else
    if $u_{bid} < p_{quoteBid}$ **then**
        Passively add a contract to purchase at level x in the order book, where $x \sim TruncGeo(p, n)$ and where $TruncGeo$ is a geometric distribution with parameter p, which has been truncated to put mass on a discrete set of values, $\{1, 2, \ldots, n\}$.
    else
        Passively add a contract to sell at level x in the order book, where $x \sim TruncGeo(p, n)$ and where $TruncGeo$ is a geometric distribution with parameter p, which has been truncated to put mass on a discrete set of values, $\{1, 2, \ldots, n\}$.
    end if
end if

Suppose that agents A and B follow the strategy outlined in Algorithm 2. In particular, if the price of the asset declines by two spreads or more over the course of two time periods, they aggressively fill all contracts at the best bid and subsequently add passive bid and offer orders at prices that move the top-of-book prices down by one increment. They employ a symmetric strategy when the asset price increases by two or more spreads over two periods. Such a strategy would induce a "hot-potato" effect (CFTC and SEC (2010a) and CFTC and SEC (2010b)) in the absence of market participation: high-frequency traders would walk the market monotonically upward or downward while passing the asset back and forth and

logging profits.

Algorithm 2 High-Frequency Trader

1: Compute $\delta = p_t - p_{t-2}$, where p_t is the current price of the asset.
2: Given a minimum price increment of ξ,
3: **if** $\delta \leq -2\xi$ **then**
4: Aggressively fill (sell) all orders at the best bid.
5: Add a passive contract on offer at the transacted price and a passive contract on bid at the price one increment below the transacted price.
6: **end if**
7: **if** $\delta \geq 2\xi$ **then**
8: Aggressively fill (buy) all orders at the best offer.
9: Add a passive contract on bid at the transacted price and a passive contract on offer at the price one increment above the transacted price.
10: **end if**

Algorithm 2 is not intended to be an accurate description of trading and market making behavior by high-frequency traders. It is, instead, a stylized representation of the behavior of agents that profit during rapid market declines. Similarly, Algorithm 1 serves as a stylized representation of non-HFT market making and trading behavior: random walking asset prices, with substantial bid/offer bounce. The simulated interaction of these agents, however, allows us to explore the circumstances under which market crashes (upward or downward) can occur as market participation shifts from a preponderance of fundamental, non-HFT agents, to a preponderance of high-frequency traders.

Panel (a) of Table 3 reports parameter values of the simulation model that remain fixed across simulations. The initial price and the minimum price increment are chosen to be proportional to the respective quantities of the ES contract at the time of the Flash Crash: roughly 1000 and 0.25 index points, respectively. The number of levels retained in the simulated order book, $n_{levels} = 10$, corresponds to the number of levels reported in the CME DataMine Market Depth data, which is the information available to traders in real time. Conditional on a market (non-HFT) event occurring, the probability that the event is a trade,

p_{trade}, is set to represent the fact that more order book activity surrounds the submission and cancellation of orders, rather than actual transactions. Finally, the probability parameter of the truncated geometric distribution, p, is set so that there is substantial depth at the best bid and best offer, but that orders are also placed with declining probability deeper in the book, allowing for occasional price movements even when HFT agents do not trade.

The remaining parameters govern the length of each simulation and the determination of crashes. In the month preceding the Flash Crash, the median number of messages sent to the CME during equities market trading hours was roughly 2,470,000, or approximately one message every 10 milliseconds. Following this approximation, we run each simulation for 100,000 periods, or just under 17 minutes, and keep track of the maximum and minimum prices over a rolling window of 6000 periods, or roughly 1 minute. If the absolute difference between maximum and minimum prices over any minute exceeds 200 spreads, or 5% of the initial price, we denote the price trajectory as a crash. These values were chosen to loosely correspond to the time periods and price movements on the day of the Flash Crash. We note that each simulation commences with a burn-in of 100 periods of market quoting to ensure adequate depth at the top of the order book at the initial time period.

Panel (b) of Table 3 reports the fraction of crash occurrences as the probability of market participation varies for our baseline calibration. The row labeled "Market Participation" denotes the probability, p_{market}, that agent M is selected at each time period to either supply a quote or transact, with the remaining probability at each period attributed to the HFT agents (who are chosen at random with equal probability). Note, however, that HFT agents only act differently from the market agent in the case of two consecutive price movements in the same direction. For each value of p_{market}, we simulated 100 price trajectories and computed the fraction of trajectories that terminated in crashes. Under the baseline setup, the market agent's quoting and trading behavior are symmetric on both sides of the order book: conditional on an action (either quote or trade) by agent M, $p_{tradeBid} = p_{quoteBid} = 0.5$. It is readily apparent from the table that the frequency of crashes increases as market participation declines. In particular, when $p_{market} = 0.8$, none of the simulations result in a crash, whereas all simulations result in a crash for $p_{market} = 0.001$. This is due to the fact that with a low probability of quoting activity by agent M, the first occurrence of two

(a) Model Parameters												
$p_0 = 4000$		$\xi = 1$			$n_{levels} = 10$			$p = 0.85$				
$p_{trade} = 0.4$		$n_{sim} = 100,000$			$\tau_{crash} = 6000$			$\sigma_{crash} = 200$				
(b) Baseline case: $p_{tradeBid} = p_{quoteBid} = 0.5$												
Market Participation	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.005	0.001
Crash Fraction	0.00	0.11	0.45	0.53	0.62	0.70	0.72	0.80	0.87	0.97	0.99	1.00
(c) Waddell & Reed case: $p_{tradeBid} = 0.54$, $p_{quoteBid} = 0.46$												
Market Participation	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.005	0.001
Crash Fraction	0.02	0.22	0.54	0.68	0.76	0.81	0.85	0.92	0.95	0.99	1.00	1.00

Table 3: Fraction of simulations resulting in crashes as market participation varies.

consecutive price movements in the same direction almost deterministically results in the HFT agents playing a game of hot potato until the asset price exceeds the crash threshold.

As mentioned in CFTC and SEC (2010b), about 13 minutes prior to the rapid ES decline that precipitated the CME market stop on May 6, 2010, a large fundamental trader initiated a sell order of 75,000 ES contracts. Menkveld and Yueshen (2015) identifies this fundamental trader as Waddell & Reed Financial, Inc. Both sources report that Waddell & Reed utilized an algorithm to implement the trade without regard to price and time, but with an volume execution target of 9% of trading volume over each previous minute. The algorithm reportedly supplied both aggressive and passive orders to the book. In an effort to understand the effect of this increased selling pressure on the probability of market crash, we repeated our simulation exercise, but with probabilities of market trade at the bid, $p_{tradeBid}$, and market quote on the offer, $1 - p_{quoteBid}$, increased by 9%. Panel (c) of Table 3 reports the fraction of simulations that result in market crashes under the revised calibration. For each value of p_{market}, the probability of market crash is higher than in the baseline case, reflecting the fact that a significant increase in selling pressure induces a greater likelihood of consecutive price declines that lead to the HFT hot potato game. We view this as a conservative result because selling pressure during the Flash Crash was actually even more skewed, as many traders other than Waddell & Reed attempted to sell their ES positions.

We recognize that this simulation model is not the only one that can replicate Flash-

Crash type behavior. It does, however, provide a rigorous and parsimonious model that is consistent with observed data, with the CFTC-SEC analysis, and with much of the intuition surrounding market behavior during the Flash Crash.

5 FINRA Trade Reporting Facilities

Our millisecond-level analysis of message traffic also leads to the discovery of previously unobserved anomalies in information flows during the Flash Crash. Our research into the implications of these information flow issues is ongoing, and further analysis is required before we can draw firm conclusions as to whether these information flow issues are symptoms of the Flash Crash, or whether they contributed to the emergence and resolution of the Flash Crash in a causal manner.

The Consolidated Tape System (CTS) aggregates transacted prices for equities across SEC registered exchanges and subsequently disseminates a uniform data feed to the market. Table 4 lists the eight exchanges that published equities transactions to the CTS on May 6, 2010, along with their total SPY volume and SPY volume share during that trading day. The last row of the table reports volume and share for over-the-counter transactions, denoted as "FINRA" in the CTS. This latter price series corresponds to transactions attributed to off-exchange entities that are required to report to the CTS via FINRA Trade Reporting Facilities (TRFs) separately established at NYSE and Nasdaq. According to CFTC and SEC (2010a), the FINRA TRF is comprised primarily of over-the-counter transactions, internalizers (institutions which internally match orders), dark pools, and ECNs, such as DirectEdge. As made clear in the table, although the FINRA TRFs are physically located at the NYSE and Nasdaq facilities, they represent a distinct sequence of price information.

Figure 9 displays CTS transacted prices at 250 millisecond intervals for SPY, between 2:38 and 2:58 p.m. EDT on May 6, 2010. Panel (a) includes all reporting facilities except the CSE and CBOE, whose trading was too thin to include. The figure also includes traded prices for the E-mini, after performing a basis adjustment that accounts for the risk-free interest rate and dividends. Two striking features emerge in panel (a). First, at approximately 2:42:45.538 p.m. EDT, the price series corresponding to the FINRA TRFs and the NSE

(a) SPY Volume and Share on CTA Exchanges		
Exchange	Volume	Volume Share
NASDAQ OMX	196896670	0.342
New York Stock Exchange (NYSE)	135015441	0.234
BATS	97466677	0.170
NASDAQ OMX BX	16585818	0.0288
Chicago Stock Exchange (CSE)	15044262	0.0261
International Securities Exchange (ISE)	4952069	0.00860
Chicago Board Options Exchange (CBOE)	3208600	0.00557
National Stock Exchange (NSE)	1986117	0.00345
Financial Industry Regulatory Authority (FINRA)	104663315	0.182
(b) High - Low Deviation Counts		
Apr 23, 2010	1	
May 6, 2010	878	
Jun 7, 2010	1	
Aug 20, 2010	1	
Mar 10, 2011	1	
May 6, 2011	1	
Aug 9, 2011	1	

Table 4: Panel (a): Aggregate volume and volume share for CTS participating exchanges on May 6, 2010. Panel (b): Number of high - low deviations exceeding \$0.50 over 250 millisecond intervals during each trading day included in the regression analysis of Tables 1 and 2.

diverge from the consensus market price and proceed to report delayed transactions until some time after 2:55 p.m. Nanex (2010) and Flood (2010) report similar delays in *quotations* for a number of equities at the New York Stock Exchange. Our finding is distinct in that it focuses on *off-exchange transactions*. Second, at approximately 2:46:34.061 p.m., the basis-adjusted E-mini price diverges from the market consensus. Because the NSE accounts for less than one third of 1% of total volume, we will not focus attention on it. The FINRA

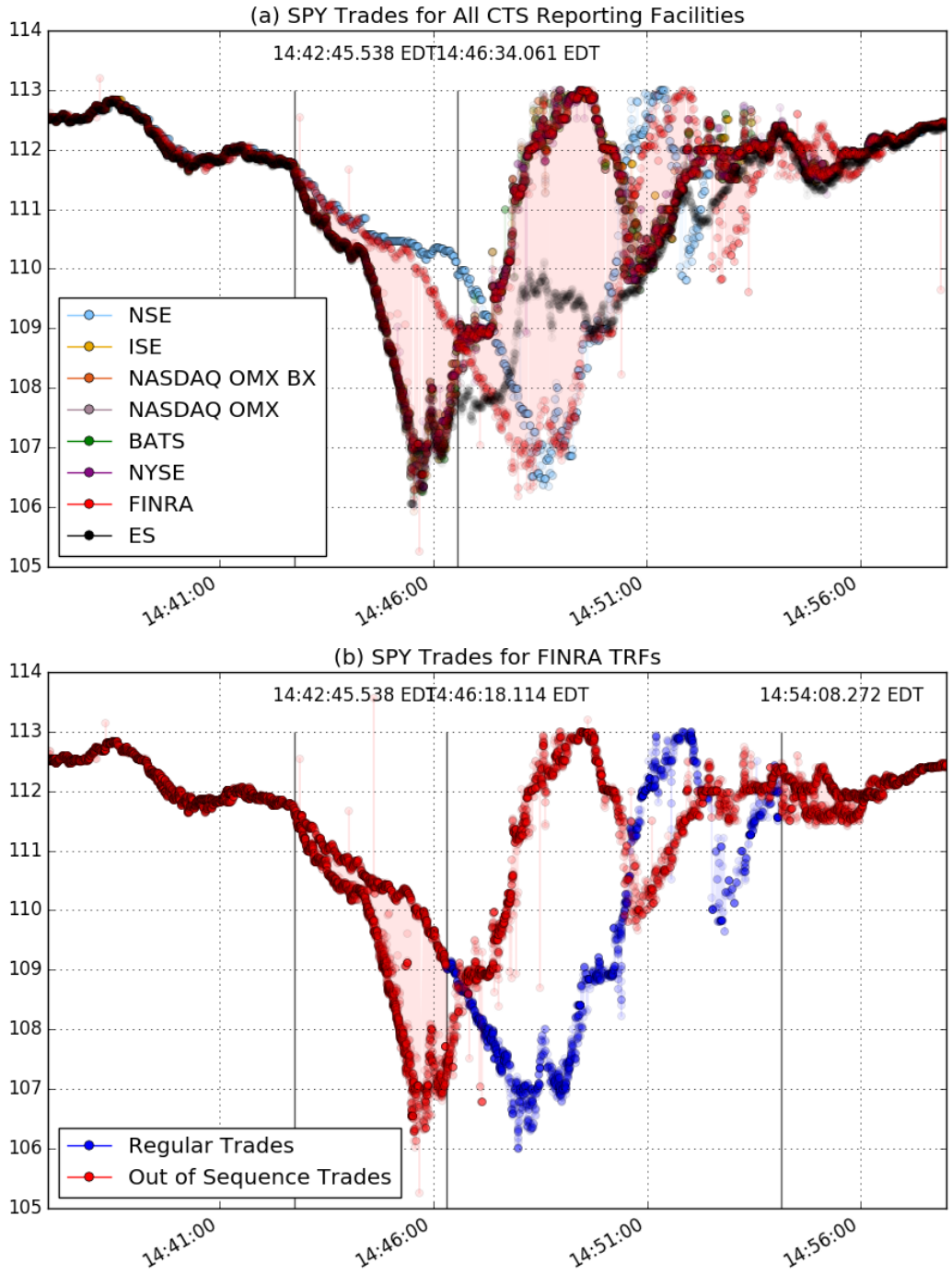

Figure 9: Panel (a) shows all transacted prices reported to the CTS, excluding the Chicago Stock Exchange and Chicago Board Options Exchange, between 2:38 and 2:53 p.m. EDT on May 6, 2010. Panel (b) isolates the FINRA TRF transactions over the same period and depicts those that were flagged as "out of sequence".

TRF prices, however, account for close to one fifth of total reported SPY volume for the day of the Flash Crash. Interestingly, these prices do not represent a uniform deviation, but instead rapidly oscillate between the market consensus price and the apparent delayed price. While we cannot identify which entity or entities were publishing delayed transactions to this data feed, it appears to be clear that one or more participants were contributing to the oscillating price series.

Panel (b) of Figure 9 isolates the FINRA TRF price series and further distinguishes prices by an "out-of-sequence" trade flag that is provided in the NYSE Daily TAQ dataset. Between 2:46:18.114 and 2:54:08.272 p.m., the delayed transactions were marked with this flag, creating two distinct price series and no oscillation. These correctly flagged points are painted blue in the figure. The upshot is that for roughly 3.5 minutes, from 2:42.45.538 to 2:46.18.114, the delayed, off-exchange prices were not labeled as such and could have been interpreted as live, marketable prices when, in fact, they reflected stale prices no longer available in the market.

As of May 6, 2010, FINRA Rule 6282 required all off-exchange transactions in eligible securities (including SPY) to be reported to the FINRA TRFs within 90 seconds of final sale. On Nov 1, 2010, and Nov 4, 2013, this rule was amended to shorten the reporting window to 30 seconds and 10 seconds, respectively. Figure 10 shows estimates of the delay in the upper portion of the FINRA TRF series, as depicted in Figure 9. To estimate the delay,

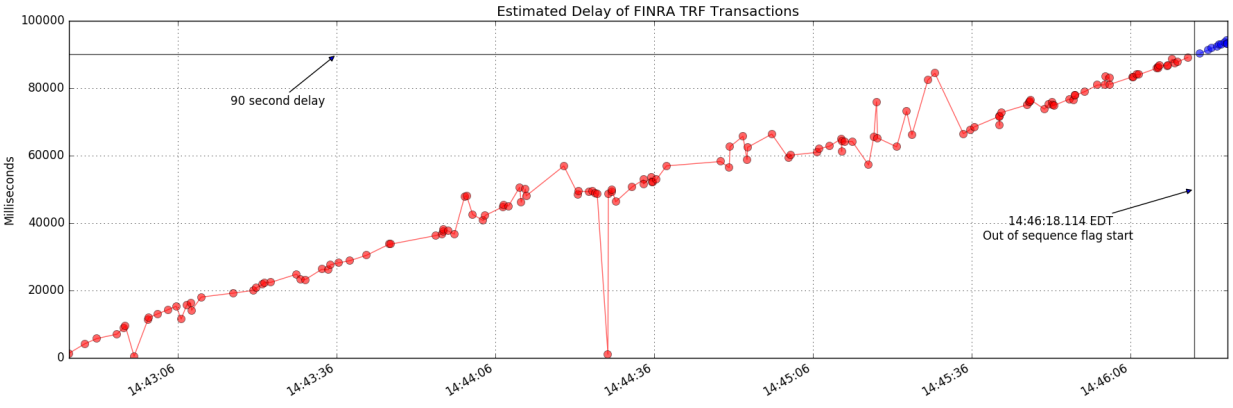

Figure 10: Estimated delay of FINRA TRF traded prices for SPY.

we compute the maximum and minimum FINRA TRF times associated with each price

$p \in [\$109.00, \$111.75]$. We then compute the median transaction time for Nasdaq trades at p, isolate any FINRA transactions that have transaction times greater than the mid point between the median Nasdaq time and the maximum FINRA time, and compute the median time of the remaining FINRA transactions. The estimated delay is the difference between the median FINRA and Nasdaq times, which we compute from 2:40:00 p.m. EDT (prior to the visual departure of the two series) to 2:46:30 p.m. EDT (after the transactions were flagged as "out of sequence"). As shown in Figure 10, aside from noise due to estimation error, the estimated delay increases monotonically and reaches 90 seconds at precisely the instant that the transactions receive the "out of sequence" attribution. Although it is theoretically possible that none of the FINRA transactions prior to 2:46:18:114 were out of sequence, our analysis strongly suggests that one or multiple FINRA members were submitting delayed transactions to the TRF, and only marked them as such when required by regulation.

It bears emphasis that none of this analysis suggests a violation of any rule by any market participant. To the contrary, it suggests that late trades were properly marked as such once they were delayed by at least 90 seconds. The analysis opens the door to the possibility, however, that trades delayed by less than 90 seconds may have contributed to the evolution of the Flash Crash.

A priori, we have no information as to whether price divergences in the CTS data, similar to that depicted in Figure 9, are common. To understand the frequency with which such divergences occur, we compute the high and low prices for SPY over 250 millisecond intervals between 9:30 am and 4:00 p.m. ET on each of the sample days included in our regression analysis reported in Tables 1 and 2, above. Panel (b) of Table 4 tabulates the number of intervals for which the difference of high and low prices exceeds \$0.50, or 50 spreads. To keep the size of the table manageable, we only report dates for which there is at least a single interval with such a deviation. Panel (b) shows that only 7 days had such a deviation. Further, aside from May 6, 2010, each of these days experienced only one such price discrepancy. In stark contrast, the day of the Flash Crash displays 878 such deviations, and all 878 occur after 2:42:45.538 p.m. EDT (the time at which we first detect a departure of the FINRA TRF price series), with 439 occurring in the interval 2:42:45.538 – 2:46:18.114 p.m. and 326 occurring in the interval 2:46:18.114 – 2:54.08.272 p.m. From these data, we

conclude that price deviations of the type in Figure 9 are extremely abnormal in the CTS data, given the dates we examined.

To further understand the behavior of FINRA TRF transactions during the time period surrounding the Flash Crash, panel (a) of Figure 11 zooms in on the interval between 2:44:00 and 2:48:00 p.m. Red and blue points again represent regular and out-of-sequence transactions. As a reference, the figure also includes Nasdaq OMX traded prices (blue line) and CME basis-adjusted traded prices for the E-mini (green line). We consider the Nasdaq transactions to represent the market consensus price, as the vast majority of CTA transactions were congruent with those values. The CME market halt is represented by the flat green line surrounding 2:45:30 p.m. Close inspection of the SPY and ES prices suggests that the market continued on a downward trajectory after E-mini trading resumed. Indeed, it wasn't until approximately 15 seconds later, when the erroneous FINRA prices likewise experienced a halt (of about equal duration), that the market noticeably turned upward from its trough.

This pattern suggests additional or alternate explanations for both the market decline and recovery on May 6, 2010. Anomalous CTS prices for the SPY, which were delayed or "out of sequence", but which were not required to be reported as such, were causing rapid oscillations in the data feed of marketable prices. If this data feed anomaly was noticed by high frequency market makers it could have caused uncertainty in their algorithms as to the correct market price, and that uncertainty could have led them (rationally) to withdraw liquidity, at least until the uncertainty was resolved. The CME trading halt would then not, in and of itself, have directly caused the market to rebound. Instead, the CME halt may have triggered a trading stop among the participants whose transactions were being reported to the FINRA TRF with a delay, resulting in the subsequent elimination of anomalous prices. This pause in oscillating prices may have been sufficient to draw high-frequency liquidity back to the market, resulting in the coincident, observed recovery. Put another way, the CME halt may have had only an accidental, indirect stabilizing effect (if any) on the market – an effect more modest than that suggested by others (CFTC and SEC, 2010b; Kirilenko et al., 2015).

Alternatively, these reporting delays may have been a symptom of the rapid trading that occurred during the Flash Crash. The CME trading halt may have provided traders with an

opportunity to catch up with their reporting obligations, and the Flash Crash might have proceeded as observed without regard to the accuracy of these particular data feeds on that day.

Additional analysis is necessary in order to reach more firm conclusions regarding the effect, if any, of these data feed issues and of the CME trading halt on the price decline and recovery that is the signature of the Flash Crash. In particular, the analysis would benefit from additional information describing the data feeds relied upon by algorithms active in the markets on May 6, 2010, and whether asynchronicities of the sort we observe could have caused the withdrawal and return of algorithm-driven liquidity. If this data feed issue explains either the price decline or recovery, it could call into question the dominant narrative regarding the cause of the Flash Crash and common perceptions regarding the role of the CME halt in stimulating the market's recovery.

6 Discussion and Conclusion

Detailed analysis of message traffic at the millisecond level of granularity for the entire order book in the E-Mini and SPY provides insights as to the evolution and causation of the Flash Crash that cannot be discerned through other modes of analysis. In particular, it allows for comparison of the dominant theories associated with Flash Crash causation. Our analysis is consistent with a narrative of the Flash Crash that is somewhat more mundane than many previous summations in the literature. The macroeconomic context, as well as the overall level of stress on the U.S. equity and futures markets from the opening bell through approximately 2:40 p.m. on May 6 2010 were high, but were by no means unprecedented. We have specifically demonstrated that trading rates, intraday volatility, and price trajectories on August 9, 2011 were extremely reminiscent of those observed during the morning and early afternoon prior to the Flash Crash. Yet on August 9, 2011, trading was orderly throughout the day, and arbitrage opportunities between fungible instruments such as the E-mini futures contract and the SPY ETF were consistently eliminated on time scales consistent with near speed-of-light messaging between the pertinent financial exchanges. Order imbalances in the futures market were consistently large on the Flash Crash day, but only for price levels

that were far from the inside levels of the order book, and we have argued that these order imbalances were quite unlikely to have had a tangible effect on the trajectory of traded prices.

The United States Government, through the joint report issued by the staffs of the CFTC and SEC (CFTC and SEC, 2010b) initially explained the Flash Crash as the result of from the interaction of unsettled market conditions with the introduction of a large sell order executed in a destabilizing manner. The Government later expanded its theory of causality to include allegations that Navinder Sarao's "numerous aggressive spoofing tactics" (CFTC v. Sarao, 2015a, para. 2) "caused artificial prices to exist" on at least twelve trading days, including the Flash Crash. (CFTC v. Sarao, 2015a, para. 1)

The CFTC asserted that Sarao's algorithms "caused the price of the E-mini S&P contract to be temporarily artificially depressed" while the algorithm was active, and that "the market price typically rebounded" once the algorithm was turned off. (CFTC v. Sarao, 2015a, para. 53) "In other words, Defendant ... introduced artificial volatility into the E-mini S&P futures market and caused artificial prices to exist." (CFTC v. Sarao, 2015a, para. 53) As for trading on the date of the Flash Crash, the CFTC alleges that Sarao's activity "represented approximately $170 million to over $200 million worth of persistent downward pressure on the E-mini S&P price" and for a period "represented 20-29% of the sell side Order Book." (CFTC v. Sarao, 2015a, para. 76) Further, "at that time, the Order Book was severely imbalanced and Defendant's [algorithmic] orders were almost equal to the entire buy-side of the Order Book." (CFTC v. Sarao, 2015a, para. 76) The Government's criminal complaint makes similar allegations (USA v. Sarao, 2015a,b).

To the extent that trading algorithms incorporate order imbalance information as a measure of potential market liquidity and direction, there is a small possibility that the elevated imbalances on May 6, 2010 influenced the precipitous decline in prices. Our empirical evidence, however, suggests that deep order book imbalances have little material effect on subsequent prices. Our empirical work, however, does not negate that Sarao engaged in illegal, manipulative conduct. Nor do we contest the allegation that Sarao's algorithms could have, on occasion, caused "artificial" prices and volatility to appear in the market, as alleged by the Government. Our challenge is, instead, to the inference that Sarao's illegal conduct

was a material contributing cause of the Flash Crash, that he intended to cause the Flash Crash, or that the Flash Crash was even a foreseeable result of his illegal trading activities. Despite this distinction, Mr. Sarao's conduct may allow him to be incarcerated, enjoined, and fined.

However, the extent to which the Flash Crash was a foreseeable consequence of Sarao's activity is a legally and policy-relevant inquiry because, among other matters, the United States Sentencing Guidelines define the "actual loss" used for sentencing purposes in mail and wire fraud cases as "the reasonably foreseeable pecuniary harm that resulted from the offense." (United States Sentencing Guidelines 2015) Thus, if the Flash Crash was not a reasonably foreseeable consequence of Sarao's illegal spoofing activity, his potential sentence, if he is convicted, should not be enhanced by the fact that his trading may have been contiguous in time or correlated with the Flash Crash.

The causal link between Sarao's trading and the Flash Crash is, however, significant from a public policy perspective. If policy makers believe that Sarao's spoofing activity materially contributed to the Flash Crash, they can then rationally conclude that increased prosecution of certain forms of trading activity is socially beneficial precisely because it decreases the probability of a future Flash Crash. Our analysis suggests that this view incorrectly conflates correlation with causation: just because Sarao's trading occurred at or around the time of the Flash Crash, does not establish that it helped cause the Flash Crash. Our analysis also suggests that the Flash Crash (as distinct from some significantly smaller price perturbation) was entirely unforeseeable to Sarao and to others in the market. Consistent with this analysis, the Government alleges that Sarao engaged in his illegal trading activities on numerous trading days, none of which experienced price movements even fractionally as dramatic as those observed during the Flash Crash. Moreover, the Government nowhere alleges that Sarao intended to cause a disruption as massive as the on the market experienced during the Flash Crash. Indeed, this paper suggests that the Flash Crash could have occurred even without Sarao's presence in the market.

Our work instead suggests that policymakers interested in reducing the probability of a future Flash Crash are better guided by the findings of the joint CFTC-SEC staff report, which does not rely on Sarao's presence in the market (CFTC and SEC, 2010b). Our analysis

reinforces the conclusion that the dynamics described by the joint CFTC-SEC staff report are consistent with the observed data and, given the current state of our knowledge, represent a sufficient explanation of the Flash Crash. Our simulation model formalizes the insights upon which the report relies, and demonstrates the existence of a market instability when liquidity thins (Easley et al., 2011). In the instability, algorithmic traders act in concert to drive a rapid, linear decline in price that is very similar to what was observed on May 6, 2010. Such declines are exacerbated by large sell orders, such as the one placed by Waddell & Reed, and are arrested by the entrance of fundamental buyers to the market. To be sure, the prosecution of illegal manipulative trading activities can play a role in the government's regulatory strategy, but the danger is that regulators will perceive this form of enforcement activity as an effective substitute for more fundamental restructuring of modern markets.

These conclusions may, however, have to be amended as a result of further analysis of anomalies discovered in the FINRA Trade Reporting Facility. We establish that these anomalies, in the form of misattributed, late reported trades, appear to have emerged virtually simultaneously with the sharpest portion of the Flash Crash price decline. These anomalies caused significant oscillations in prices reported to the consolidated tape during the rapid decline, and persisted until after the CME trading halt. The market recovery did not begin until after a halt in the FINRA feed that corrected these pricing anomalies. The FINRA halt occurred about 15 seconds after the CME halt.

This sequence of events is susceptible of two distinct interpretations. From one perspective, these data feed anomalies can be viewed as a symptom of the Flash Crash that resulted as traders fell further behind in their reporting obligations in a manner that did not violate rules or regulations in effect in the market. If traders pay no attention to this data feed, or if they understand how to interpret large oscillations of the sort we observe, the data feed anomalies might not be causal.

Alternatively, algorithmic traders observing these price oscillations could have become concerned with the overall integrity of messaging traffic and prices prevailing in the market. As our analysis suggests, anomalies of this sort are extremely rare. The combination of confusion over transactions prices combined with sharply declining markets could have led algorithmic traders to withdraw liquidity and then to restore it only after this source of

price-information uncertainty was resolved. If so, the data anomaly can be viewed as a material contributing cause of the day's precipitous price movements, and presents a rational explanation for the initiation and resolution of the Flash Crash that would be consistent to the millisecond level. This explanation is also consistent with our simulation model.

If the data feed anomalies were, in fact, a material contributing cause of the Flash Crash, then reconsideration of the role played by the CME halt may be appropriate. The CME five-second trading halt, from 2:45:28 to 2:45:33 p.m. EDT, is commonly credited with creating an opportunity for the market to recover from its intra-day lows (CFTC and SEC, 2010b; Kirilenko et al., 2015). However, our data establish that the market continued to decline after the CME halt, and that the recovery did not begin until after a halt in the FINRA data feed that began 15 seconds after the end of the CME halt and lasted for an additional 5 seconds. This pattern is consistent with the possibility that the CME halt gave the FINRA feed an opportunity to "catch up" so as to stop delivering stale and confusing prices, and that only after the confusion was resolved could a recovery take place. If so, the CME halt might be better described as a necessary condition for the correction of the FINRA feed, but an insufficient condition for the market's recovery. In contrast, the correction of the FINRA feed, for whatever reason, would constitute a necessary and sufficient condition for the market's recovery.

Additionally, we show that until the CME trading halt was implemented, quoting activity on the Nasdaq exchange was consistent with inter-exchange messaging that occurred with latencies near the physical minimum associated with the finite speed of light. We find, however, that after CME trading was restarted, correlation at the millisecond level was lost.

The policy implications of this explanation, if correct, are simple and straightforward: pay attention to data feeds. High frequency markets operate at sub-millisecond levels of precision and the introduction of sufficient noise into the market's information streams can cause traders to withdraw liquidity in a manner that can readily precipitate a Flash Crash. Regulatory actions have already reduced the time periods during which traders can report late trades without flagging them as such from the 90 second period that prevailed during the Flash Crash to the current level of 30 seconds. (FINRA Rule 6282) Further shortening of that period, if practical, may be warranted. In addition, regulators and market participants

may want to carefully consider other contingencies that can cause anomalous trade reporting or that can introduce other forms of noise into data feeds, and take measures that reduce or eliminate the possibility of such forms of confusion.

Our analysis of these data feed anomalies is continuing. It will involve additional computationally intense examinations of the sequence in which various events occurred in the market. It will also involve interviews with market participants knowledgeable about the operation of algorithms that were then present in the market and about the likely liquidity implications of anomalies of the sort we have observed.

References

Bak, P., Tang, C., and Wiesenfeld, K. (1987), "Self-Organized Criticality: An Explanation of 1/f Noise," *Physical Review Letters*, 59, 381–384.

CFTC and SEC (2010a), "Findings Regarding the Market Events of May 6, 2010," *Staff Report*.

— (2010b), "Preliminary Findings Regarding the Market Events of May 6, 2010," *Staff Report*.

CFTC v. Sarao (2015a), *Case 1:15-cv-03398*, N.D. Ill. Apr 17, 2015, Court Docket.

— (2015b), *Case 1:15-cv-03398 Appendix*, N.D. Ill. Apr 17, 2015, Court Docket.

Clearfield, C. and Weatherall, J. O. (2015), "Why the Flash Crash Really Matters," .

Easley, D., Lopez de Prado, M. M., and O'Hara, M. (2011), "The Microstructure of the "Flash Crash": Flow Toxicity, Liquidity Crashes and the Probability of Informed Trading," *Journal of Portfolio Management*, 37, 118–129.

— (2012), "Flow Toxicity and Liquidity in a High-frequency World," *Review of Financial Studies*, 25, 1457–1493.

Fleming, M. J., Mizrach, B., and Nguyen, G. (2014), "The Microstructure of a U.S. Treasury ECN: The BrokerTec Platform," *Working Paper*.

Flood, J. (2010), "NYSE Confirms Price Reporting Delays That Contributed to the Flash Crash," *http://www.ai-cio.com/channels/story.aspx?id=1490*, Retrieved Jan 6, 2015.

Fox, M. B., Glosten, L. R., and Rauterberg, G. V. (2015), "The New Stock Market: Sense and Nonsense," *Duke Law Journal*, 65, 191–277.

Hasbrouck, J. (2003), "Intraday Price Formation in U.S. Equity Index Markets," *The Journal of Finance*, LVIII, 2375–2399.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2015), "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market," *Working Paper*.

Laughlin, G., Aguirre, A., and Grundfest, J. (2014), "Information Transmission between Financial Markets in Chicago and New York," *The Financial Review*, 49, 283–312.

Menkveld, A. J. and Yueshen, B. Z. (2015), "The Flash Crash: A Cautionary Tale about Highly Fragmented Markets," *Working Paper*.

Michaels, D., Leising, M., and Mamudi, S. (2015), "Flash Crash Arrest Lays Bare Regulatory Lapses at All Levels," .

Nanex (2010), "Analysis of the "Flash Crash"," *http://www.nanex.net/20100506/FlashCrashAnalysis_Intro* Retrieved Jan 6, 2015.

Pirrong, C. (2015), "Did Spoofing Cause the Flash Crash? Not So Fast!" .

USA v. Sarao (2015a), *Case 1:15-cr-00075 Document #1*, N.D. Ill. Feb 11, 2015, Court Docket.

— (2015b), *Case 1:15-cr-00075 Document #24*, N.D. Ill. Feb 11, 2015, Court Docket.

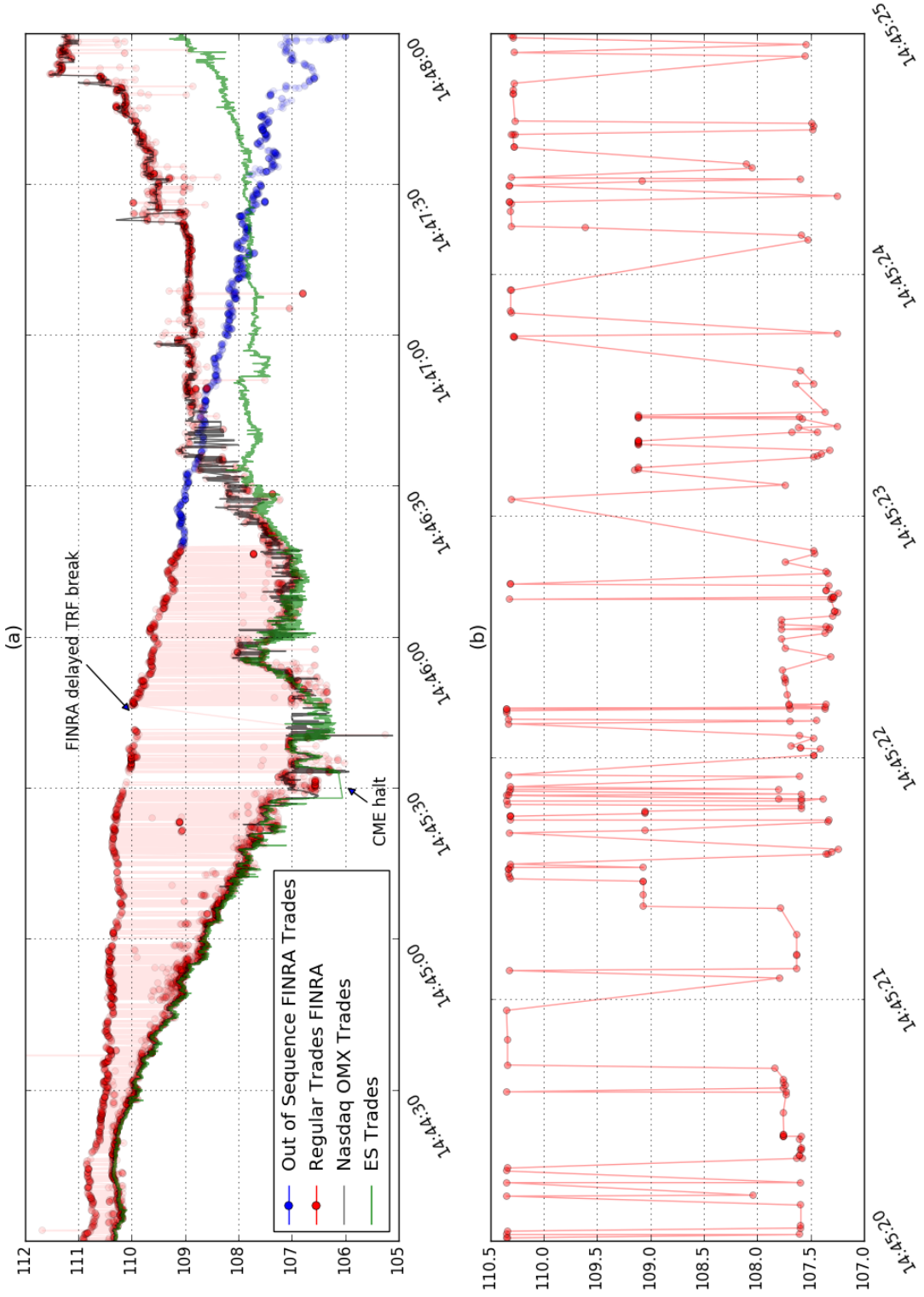

Figure 11: Panel (a): FINRA TRF traded prices for SPY (blue and red points), in addition to Nasdaq OMX traded prices for SPY (blue line) and CME traded (basis-adjusted) prices for ES (green line) for a four-minute interval on May 6, 2010. Panel (b): FINRA TRF trade prices for SPY for a five-second interval on May 6, 2010.