The Flash Crash: A New Deconstruction

Eric M. Aldrich\textsuperscript{a}, Joseph A. Grundfest\textsuperscript{b}, and Gregory Laughlin\textsuperscript{b}

\textsuperscript{a}Department of Economics, University of California, Santa Cruz, ealdrich@ucsc.edu
\textsuperscript{b}School of Law, Stanford University, grundfest@stanford.edu
\textsuperscript{b}Department of Astronomy, Yale University, gregory.laughlin@yale.edu

September 2016

Abstract

The “Flash Crash” of May 6th, 2010 comprised an unprecedented 1,000 point, five-minute decline in the Dow Jones Industrial Average that was followed by a rapid, disorderly recovery of prices. We illuminate the causes of this singular event with the first analysis that tracks the full order book activity at millisecond granularity. We document previously overlooked market data anomalies and establish that these anomalies Granger-caused liquidity withdrawal. We offer a simulation model that formalizes the process by which large sell orders, combined with widespread liquidity withdrawal, can generate Flash Crash-like events in the absence of fundamental information arrival.

Keywords: High-frequency trading, market microstructure, Flash Crash, financial market regulation.
JEL Codes: C32, C55, G14, G17, G28.

About CAFIN

The Center for Analytical Finance (CAFIN) includes a global network of researchers whose aim is to produce cutting edge research with practical applications in the area of finance and financial markets. CAFIN focuses primarily on three critical areas:

- Market Design
- Systemic Risk
- Financial Access

Seed funding for CAFIN has been provided by Dean Sheldon Kamieniecki of the Division of Social Sciences at the University of California, Santa Cruz.

*This work was supported by the Hellman Fellows Fund and the Rock Center for Corporate Governance at Stanford University.
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
Yale University

September 16, 2016

Abstract

The “Flash Crash” of May 6th, 2010 comprised an unprecedented 1,000 point, five-minute decline in the Dow Jones Industrial Average that was followed by a rapid, disorderly recovery of prices. We illuminate the causes of this singular event with the first analysis that tracks the full order book activity at millisecond granularity. We document previously overlooked market data anomalies and establish that these anomalies Granger-caused liquidity withdrawal. We offer a simulation model that formalizes the process by which large sell orders, combined with widespread liquidity withdrawal, can generate Flash Crash-like events in the absence of fundamental information arrival.

*This work was supported by the Hellman Fellows Fund and the Rock Center for Corporate Governance at Stanford University.
†Email: ealdrich@ucsc.edu.
‡Email: grundfest@stanford.edu
§Email: gregory.laughlin@yale.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 market’s 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 824 points between 1:45 and 2:45 p.m. (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 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 nearly a trillion 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, 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).
The CFTC and United States Department of Justice (DoJ) recently expanded the list of causal factors by alleging that Navinder Sarao, a small 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 generated 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 ascertain 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.

This paper attempts to address questions regarding 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 accused of displaying. Menkveld and Yueshen (2015) rely on 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 near 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 Sarao’s alleged spoofing activity.

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, this simulation model 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 present a novel examination of data feeds utilized by algorithmic traders and 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 and termination of this reporting delay correlates strongly with the inception and termination of the Flash Crash. Unlike the spoofing activity referenced above, perceived price oscillations associated with the data feed delay have a statistical impact on futures market returns, which were subsequently transmitted to equities market returns. We also document and measure the withdrawal of liquidity by market participants and further establish that the observed data feed anomalies Granger-cause the withdrawal.

This millisecond-level analysis also calls into question possibility that a 5-second CME trading halt was, alone, a cause of the subsequent market recovery, as suggested by CFTC and SEC (2010b) and Kirilenko et al. (2015). We document that prices continued to decline after the CME halt, and began their recovery only after a pause in a market data feed that exhibited extreme price oscillations. The CME halt may therefore have been a necessary but insufficient condition for the recovery if it merely contributed to the correction of the recalcitrant data feed.

These analyses support a range of conclusions. In particular, with regard to allegations against Mr. Sarao, even if one assumes 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 had 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 the statistical impact of deep order book imbalances on subsequent price movements is minuscule over the relevant time horizons.

We present a simulation model that suggests that the recurrence probability 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.
In summary, our work is consistent with a straightforward mechanism of action: corrupt data caused uncertainty among algorithmic traders, and that uncertainty rationally caused them to withdraw liquidity, particularly during a rapidly moving market. Moreover, because market prices continued to decline after the CME trading halt and began their recovery only after data feed anomalies were resolved, the public policy focus on trading halts as an appropriate response to the Flash Crash may be exaggerated. Our data suggest that improving data integrity, more than trading halts, may be necessary to prevent future Flash-Crash-type market dislocations.

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; Aldrich and Lee, 2016). 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 of 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 were also responsible (CFTC v. Sarao, 2015b).

Our analysis rests on four data sources. 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° 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 sequenced. 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 often in excess of $200 billion.

The second data source 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.

Third, 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 millisecond 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.

Finally, we obtained feed-specific data from one of the largest equities market participants for the single day of the Flash Crash. The data consist of the firm’s records of individual subscriptions to proprietary equities-market data feeds, including NYSE, NYSE Arca, Nasdaq OMX, Nasdaq BX, BATS, BATS Y, Direct Edge A (EDGA) and Direct Edge X (EDGX), aggregated at a single data server. While the proprietary quotes and transactions are effectively duplicates of those reported at the CTS, these data were transmitted directly to the market participant through prioritized infrastructure and avoided latencies associated with consolidation at the CTS. For this reason, they represent the highest quality data available for equities on the day of the Flash Crash, and allow for infrastructure latency analysis of the CTS.

3 Analysis of Messaging and Imbalances

3.1 Messaging

Figure 1 shows the price trajectory of the E-mini between 1:30 p.m. and 3:30 p.m. EDT on May 6, 2010 (blue line). Although prices had been steadily falling throughout the day, an abrupt decline commenced at 2:42:45 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. Soon after the halt, prices rebounded, experiencing a near-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. Panel (a) of Figure 1 shows that
during the 1:30 – 3:30 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 rate of messaging and trading through the day, as illustrated in panel (b) of Figure 1, which depicts messaging rates in megabits per second for the 1:30 – 3:30 p.m. period on both days. There are no particularly dramatic 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 also 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 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 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 2\( N \) 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 \( \delta_l = (\text{added}+\text{removed}) \) 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 2 depicts our estimate of $\delta_t$ 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, where 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 3 extends the analysis by computing the $\delta_t$ 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 $\delta_t$ 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
Figure 2: 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.
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 2. In particular, communica-

![Figure 3: 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.](image)

cation was well established and Nasdaq 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 communica-
tions 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 underscores the remarkable conclusion that inter-market arbitrage opportunities were available even in the presence of ultra-low latency information dissemination.

This 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 experienced 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 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 preceding high-frequency correlation analysis demonstrates that the CME unambiguously led the sharp linearly downward price movement prior to the CME trading halt. On 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 sufficient margin to take advantage of these opportunities. Several minutes were required for new market entrants 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

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

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. Because 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 4 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 add 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 2:42:45.528 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 4: Log imbalance ratios for all levels of the order book on May 6, 2010 and August 9, 2011. The two vertical red lines correspond to 2:42:45.538 p.m. EDT (the start of the Flash Crash – see Section 5) and 2:45:28 p.m. EDT (the start of the CME halt).

Figure 5 depicts the same ratios for the one hour period 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 could be an artifact of stale orders that moved upward in the book as the market declined. The upshot is that the imbalance ratios appear 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 6 shows only the 1-minute moving average lines for the log order imbalance ratios on 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 elevated ratios. As noted in Hendershott’s affidavit (CFTC v. Sarao, 2015b), this is consistent with Sarao’s layering algorithm 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) E-mini contract on prior 5-second order
Figure 6: 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. The two vertical red lines correspond to 2:42:45.538 p.m. EDT (the start of the Flash Crash – see Section 5) and 2:45:28 p.m. EDT (the start of the CME halt).

book imbalance differences, \( d_t^l = N^l_{t, \text{offer}} - N^l_{t, \text{bid}} \):

\[
  r_t = \beta_0 + \beta_1 d_{t-1}^l + \varepsilon_t,
\]

where \( l \) represents a price level or group of price levels. We also fit the regression using imbalance ratios at corresponding levels, \( ir^l_t \), 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
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.

The regression for All Days and All Levels most closely corresponds to the decile analysis performed by Hendershott. While the sign and magnitude of the coefficient are

### (a) Univariate Regressions, All Days

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>2-3</th>
<th>4-7</th>
<th>8-10</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef.</td>
<td>0.0647***</td>
<td>0.0396**</td>
<td>-0.0162</td>
<td>-0.0233</td>
<td>-0.0342**</td>
<td>-0.0191</td>
<td>-0.0147</td>
<td>-0.0112</td>
<td>-0.0218</td>
<td>-0.0102</td>
<td>0.00724</td>
<td>-0.0135**</td>
<td>-0.00985</td>
<td>-0.00262</td>
</tr>
<tr>
<td>S.E.</td>
<td>0.018</td>
<td>0.0177</td>
<td>0.0167</td>
<td>0.0165</td>
<td>0.0168</td>
<td>0.017</td>
<td>0.0173</td>
<td>0.0182</td>
<td>0.018</td>
<td>0.0186</td>
<td>0.0103</td>
<td>0.00647</td>
<td>0.0087</td>
<td>0.00378</td>
</tr>
</tbody>
</table>

### (b) Univariate Regressions, All Days w/o Flash Crash

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>2-3</th>
<th>4-7</th>
<th>8-10</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef.</td>
<td>0.0698***</td>
<td>0.0387**</td>
<td>-0.0118</td>
<td>-0.0178</td>
<td>-0.0217</td>
<td>-0.00944</td>
<td>-0.00758</td>
<td>-0.005</td>
<td>-0.0156</td>
<td>-0.00634</td>
<td>0.00967</td>
<td>-0.00839</td>
<td>-0.00618</td>
<td>0.00928</td>
</tr>
<tr>
<td>S.E.</td>
<td>0.0163</td>
<td>0.0161</td>
<td>0.0152</td>
<td>0.0151</td>
<td>0.0154</td>
<td>0.0155</td>
<td>0.0158</td>
<td>0.0167</td>
<td>0.0165</td>
<td>0.017</td>
<td>0.00937</td>
<td>0.00591</td>
<td>0.00798</td>
<td>0.00346</td>
</tr>
</tbody>
</table>

*, **, and *** denote significance at the 10%, 5% and 1% levels, respectively.
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 for approximately two minutes, 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 5-second returns, which could be associated with a maximal decline of -0.324 basis points over two minutes. 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

$$r_t, of_t, d_{t}^{1}, d_{t}^{23}, d_{t}^{4567}, d_{t}^{8910} = \sum_{j=1}^{5} B_j r_{t-j}, of_{t-j}, d_{t-j}^{1}, d_{t-j}^{23}, d_{t-j}^{4567}, d_{t-j}^{8910} + u_{r,t}, u_{of,t}, u_{d^{1,t}}, u_{d^{23,t}}, u_{d^{4567,t}}, u_{d^{8910,t}} \tag{3}$$

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 imbal-
ances 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).

Panel (a) of Table 2 reports estimated cumulative impulse response functions for the data sample. We do not separately report SVAR results excluding the day of the Flash Crash since the magnitude and significance of the impulse response coefficients are nearly identical. To compute the impulse responses, we subject the estimated SVAR to a 2400 contract sell-side impulse at a specific group of levels. We report the impulse responses for 24 periods, or a total horizon of two minutes. The 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 exhibits a negative cumulative impact on subsequent returns. In all cases, as with the univariate regression results, the magnitude of the impulse responses are very small, amounting to no more than 0.2 basis points, 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 Table 1 indicate that the single day of the Flash Crash is an important 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
(a) Structural VAR Impulse Responses

<table>
<thead>
<tr>
<th>Seconds</th>
<th>Level 1 Coefficient</th>
<th>Level 1 S.E.</th>
<th>Level 2–3 Coefficient</th>
<th>Level 2–3 S.E.</th>
<th>Level 4–7 Coefficient</th>
<th>Level 4–7 S.E.</th>
<th>Level 8–10 Coefficient</th>
<th>Level 8–10 S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0968***</td>
<td>0.0218</td>
<td>-0.0116</td>
<td>0.0245</td>
<td>0.000904</td>
<td>0.0260</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.0983***</td>
<td>0.0283</td>
<td>-0.0525**</td>
<td>0.0318</td>
<td>-0.00686</td>
<td>0.0338</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.0860***</td>
<td>0.0195</td>
<td>-0.112***</td>
<td>0.0381</td>
<td>-0.0232</td>
<td>0.0403</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.0591*</td>
<td>0.0379</td>
<td>-0.131***</td>
<td>0.0429</td>
<td>-0.104**</td>
<td>0.0450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.0764***</td>
<td>0.0379</td>
<td>-0.129***</td>
<td>0.0393</td>
<td>-0.0843**</td>
<td>0.0407</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0.0756*</td>
<td>0.0476</td>
<td>-0.132***</td>
<td>0.0446</td>
<td>-0.0921**</td>
<td>0.0481</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.0784</td>
<td>0.0566</td>
<td>-0.142***</td>
<td>0.0510</td>
<td>-0.102**</td>
<td>0.0563</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.0804</td>
<td>0.0665</td>
<td>-0.0937*</td>
<td>0.0708</td>
<td>-0.163***</td>
<td>0.0647</td>
<td>-0.119*</td>
<td>0.0735</td>
</tr>
<tr>
<td>120</td>
<td>0.0814</td>
<td>0.0677</td>
<td>-0.119*</td>
<td>0.0808</td>
<td>-0.185***</td>
<td>0.0778</td>
<td>-0.135*</td>
<td>0.0893</td>
</tr>
</tbody>
</table>

(b) Structural VAR Impulse Responses, with Price Oscillations

<table>
<thead>
<tr>
<th>Seconds</th>
<th>Level 1 Coefficient</th>
<th>Level 1 S.E.</th>
<th>Level 2–3 Coefficient</th>
<th>Level 2–3 S.E.</th>
<th>Level 4–7 Coefficient</th>
<th>Level 4–7 S.E.</th>
<th>Level 8–10 Coefficient</th>
<th>Level 8–10 S.E.</th>
<th>∆P\textsuperscript{TRF}</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0947***</td>
<td>0.0229</td>
<td>0.0927***</td>
<td>0.0217</td>
<td>-0.0104</td>
<td>0.0244</td>
<td>0.00168</td>
<td>0.0259</td>
<td>-6.63***</td>
</tr>
<tr>
<td>10</td>
<td>0.0953***</td>
<td>0.0297</td>
<td>0.0838***</td>
<td>0.0282</td>
<td>-0.0490*</td>
<td>0.0317</td>
<td>-0.00355</td>
<td>0.0336</td>
<td>-26.1***</td>
</tr>
<tr>
<td>15</td>
<td>0.0834***</td>
<td>0.0356</td>
<td>0.0220</td>
<td>0.0337</td>
<td>-0.108***</td>
<td>0.0379</td>
<td>-0.0194</td>
<td>0.0401</td>
<td>-33.9***</td>
</tr>
<tr>
<td>20</td>
<td>0.0573*</td>
<td>0.0404</td>
<td>-0.0106</td>
<td>0.0378</td>
<td>-0.126***</td>
<td>0.0427</td>
<td>-0.100**</td>
<td>0.0448</td>
<td>-31.6***</td>
</tr>
<tr>
<td>30</td>
<td>0.0771**</td>
<td>0.0460</td>
<td>0.00221</td>
<td>0.0378</td>
<td>-0.117***</td>
<td>0.0392</td>
<td>-0.0774**</td>
<td>0.0406</td>
<td>-37.7***</td>
</tr>
<tr>
<td>45</td>
<td>0.0739*</td>
<td>0.0566</td>
<td>-0.0190</td>
<td>0.0477</td>
<td>-0.110***</td>
<td>0.0449</td>
<td>-0.0828**</td>
<td>0.0483</td>
<td>-42.5***</td>
</tr>
<tr>
<td>60</td>
<td>0.0742</td>
<td>0.0627</td>
<td>-0.0414</td>
<td>0.0570</td>
<td>-0.118**</td>
<td>0.0515</td>
<td>-0.0927*</td>
<td>0.0568</td>
<td>-44.6***</td>
</tr>
<tr>
<td>90</td>
<td>0.0730</td>
<td>0.0678</td>
<td>-0.0844</td>
<td>0.0718</td>
<td>-0.139**</td>
<td>0.0656</td>
<td>-0.110*</td>
<td>0.0744</td>
<td>-46.2***</td>
</tr>
<tr>
<td>120</td>
<td>0.0732</td>
<td>0.0692</td>
<td>-0.117*</td>
<td>0.0823</td>
<td>-0.163**</td>
<td>0.0791</td>
<td>-0.126*</td>
<td>0.0907</td>
<td>-46.8***</td>
</tr>
</tbody>
</table>

*, **, 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.

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 mar-
ket 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, $p_{\text{trade}}$, she aggressively trades a single contract and with probability $p_{\text{quote}} = 1 - p_{\text{trade}}$ she passively adds an order to the order book. If trading, she aggressively sells with probability $p_{\text{tradeBid}}$ and aggressively buys with probability $1 - p_{\text{tradeBid}}$. Alternatively, if adding an order, she adds to the bid side of the book with probability $p_{\text{quoteBid}}$ and to the offer side with probability $1 - p_{\text{quoteBid}}$. When adding a passive order, she adds to level $x \in \{1, 2, \ldots, n_{\text{levels}}\}$, where the levels represent a discrete set of prices separated by the minimum price increment $\xi$. 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 $n_{\text{levels}}$. The probability $p$ is selected so that there is substantial depth at prices close to the top of book.

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 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
Algorithm 1 Non-HFT Market Participant

1: Draw independent uniform random variates $u_{\text{trade}}, u_{\text{bid}} \sim U(0, 1)$.

2: if $u_{\text{trade}} < p_{\text{trade}}$ then

3: \hspace{1em} if $u_{\text{bid}} < p_{\text{tradeBid}}$ then

4: \hspace{2em} Aggressively sell a single contract at the best bid.

5: \hspace{1em} else

6: \hspace{2em} Aggressively buy a single contract at the best offer.

7: \hspace{1em} end if

8: \hspace{1em} else

9: \hspace{2em} if $u_{\text{bid}} < p_{\text{quoteBid}}$ then

10: \hspace{3em} Passively add a contract to purchase at level $x$ in the order book, where $x \sim \text{TruncGeo}(p, n)$ and where $\text{TruncGeo}$ is a geometric distribution with parameter $p$, which has been truncated to put mass on a discrete set of values, $\{1, 2, \ldots, n\}$.

11: \hspace{2em} else

12: \hspace{3em} Passively add a contract to sell at level $x$ in the order book, where $x \sim \text{TruncGeo}(p, n)$ and where $\text{TruncGeo}$ is a geometric distribution with parameter $p$, which has been truncated to put mass on a discrete set of values, $\{1, 2, \ldots, n\}$.

13: \hspace{2em} end if

14: \hspace{1em} end if
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 $\xi$,
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

(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, $n_{\text{levels}} = 10$, 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, $p_{\text{trade}}$, 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.

26
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 per day, 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, $p_{\text{market}}$, 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 $p_{\text{market}}$, 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 $M$, $p_{\text{tradeBid}} = p_{\text{quoteBid}} = 0.5$. It is readily apparent from the table that the frequency of crashes increases as market participation declines. In particular, when $p_{\text{market}} = 0.8$, none of the simulations result in a crash, whereas all simulations result in a crash for $p_{\text{market}} = 0.001$. This is due to the fact that with a low probability of quoting activity by agent $M$, the first occurrence of two 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.
(a) Model Parameters

\[ p_0 = 4000 \quad \xi = 1 \quad n_{\text{levels}} = 10 \quad \tau = 0.85 \]

\[ p_{\text{trade}} = 0.4 \quad n_{\text{sim}} = 100,000 \quad \tau_{\text{crash}} = 6000 \quad \sigma_{\text{crash}} = 200 \]

(b) Baseline case: \( p_{\text{trade}} = 0.5 \)

<table>
<thead>
<tr>
<th>Market Participation</th>
<th>Crash Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.00</td>
</tr>
<tr>
<td>0.7</td>
<td>0.11</td>
</tr>
<tr>
<td>0.6</td>
<td>0.45</td>
</tr>
<tr>
<td>0.5</td>
<td>0.53</td>
</tr>
<tr>
<td>0.4</td>
<td>0.62</td>
</tr>
<tr>
<td>0.3</td>
<td>0.70</td>
</tr>
<tr>
<td>0.2</td>
<td>0.72</td>
</tr>
<tr>
<td>0.1</td>
<td>0.80</td>
</tr>
<tr>
<td>0.05</td>
<td>0.87</td>
</tr>
<tr>
<td>0.01</td>
<td>0.97</td>
</tr>
<tr>
<td>0.005</td>
<td>0.99</td>
</tr>
<tr>
<td>0.001</td>
<td>1.00</td>
</tr>
</tbody>
</table>

(c) Waddell & Reed case: \( p_{\text{trade}} = 0.54, p_{\text{quote}} = 0.46 \)

<table>
<thead>
<tr>
<th>Market Participation</th>
<th>Crash Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.02</td>
</tr>
<tr>
<td>0.7</td>
<td>0.22</td>
</tr>
<tr>
<td>0.6</td>
<td>0.54</td>
</tr>
<tr>
<td>0.5</td>
<td>0.68</td>
</tr>
<tr>
<td>0.4</td>
<td>0.76</td>
</tr>
<tr>
<td>0.3</td>
<td>0.81</td>
</tr>
<tr>
<td>0.2</td>
<td>0.85</td>
</tr>
<tr>
<td>0.1</td>
<td>0.92</td>
</tr>
<tr>
<td>0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>0.005</td>
<td>1.00</td>
</tr>
<tr>
<td>0.001</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

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 a 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, \( p_{\text{trade}} \), and market quote on the offer, \( 1 - p_{\text{quote}} \), 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 \( p_{\text{market}} \), 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.
5 Data Integrity Analysis

Our millisecond-level analysis of message traffic also leads to the discovery of previously unmeasured anomalies in information flows during the Flash Crash. This section reports on the nature of these anomalies, identifies their likely source, and documents their potential market impact, including their causal relationship to the withdrawal of liquidity by algorithmic traders and to the pace of the market’s price decline.

5.1 The Consolidated Tape Anomaly

The Consolidated Tape System (CTS) aggregates transacted prices for equities across SEC registered exchanges and then disseminates a uniform data feed to the market. Panel (a) of 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 panel reports volume and market 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) that are 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 that internally match orders), dark pools, and “lit” electronic communication networks (ECNs), such as Direct Edge. Although the FINRA TRFs are physically located at the NYSE and Nasdaq facilities, they represent a distinct sequence of price information.

Figure 7 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, which we refer to as the “Data Feed Divergence Point”, the price series corresponding to the FINRA TRFs and the National
### (a) SPY Volume and Share on CTA Exchanges

<table>
<thead>
<tr>
<th>Exchange</th>
<th>Volume</th>
<th>Volume Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASDAQ OMX</td>
<td>196896670</td>
<td>0.342</td>
</tr>
<tr>
<td>NYSE Arca</td>
<td>135015441</td>
<td>0.234</td>
</tr>
<tr>
<td>BATS</td>
<td>97466677</td>
<td>0.170</td>
</tr>
<tr>
<td>NASDAQ OMX BX</td>
<td>16585818</td>
<td>0.0288</td>
</tr>
<tr>
<td>Chicago Stock Exchange (CSE)</td>
<td>15044262</td>
<td>0.0261</td>
</tr>
<tr>
<td>International Securities Exchange (ISE)</td>
<td>4952069</td>
<td>0.00860</td>
</tr>
<tr>
<td>Chicago Board Options Exchange (CBOE)</td>
<td>3208600</td>
<td>0.00557</td>
</tr>
<tr>
<td>National Stock Exchange (NSX)</td>
<td>1986117</td>
<td>0.00345</td>
</tr>
<tr>
<td>Financial Industry Regulatory Authority (FINRA)</td>
<td>104663315</td>
<td>0.182</td>
</tr>
</tbody>
</table>

### (b) Notable Times During Flash Crash

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:42:45.538</td>
<td>Data Feed Divergence Point</td>
</tr>
<tr>
<td>2:45:28.000</td>
<td>CME Halt Starting Point</td>
</tr>
<tr>
<td>2:45:33.000</td>
<td>CME Halt Stopping Point</td>
</tr>
<tr>
<td>2:45:41.733</td>
<td>TRF Gap Starting Point</td>
</tr>
<tr>
<td>2:45:46.550</td>
<td>TRF Gap Stopping Point</td>
</tr>
<tr>
<td>2:46:18.114</td>
<td>Data Flag Initiation Point</td>
</tr>
<tr>
<td>2:46:34.061</td>
<td>E-mini Divergence Point</td>
</tr>
<tr>
<td>2:54:08.272</td>
<td>Data Flag Termination Point</td>
</tr>
</tbody>
</table>

### (c) High - Low Deviation Counts

<table>
<thead>
<tr>
<th>Date</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 23, 2010</td>
<td>1</td>
</tr>
<tr>
<td>May 6, 2010</td>
<td>878</td>
</tr>
<tr>
<td>Jun 7, 2010</td>
<td>1</td>
</tr>
<tr>
<td>Aug 20, 2010</td>
<td>1</td>
</tr>
<tr>
<td>Mar 10, 2011</td>
<td>1</td>
</tr>
<tr>
<td>May 6, 2011</td>
<td>1</td>
</tr>
<tr>
<td>Aug 9, 2011</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4: Panel (a): Aggregate volume and volume share for CTS participating exchanges on May 6, 2010. Panel (b): Important times during the Flash Crash period. Panel (c): 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.
Figure 7: Panel (a) shows all transacted prices reported to the CTS, excluding the Chicago Stock Exchange and Chicago Board Options Exchange, between 2:38 and 2:58 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”. The vertical red lines in panel (b) correspond to the Data Feed Divergence Point, The Data Flag Initiation Point and the Data Flag Termination Point, respectively (see panel (b) of Table 4).

Stock Exchange (NSX) 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 exchange quotations for several equities at the New York Stock Exchange. Our finding is distinct in that it focuses on reported off-exchange transactions, and, as explained at greater length below, empirically connects the evolution of data feed issues to the withdrawal of liquidity and to the rapidity of observed price declines.
Second, at approximately 2:46:34.061 p.m., which we refer to as the “E-mini Divergence Point”, the basis-adjusted E-mini price diverges from the market consensus. Because NSX accounts for less than one third of 1% of total volume, we will exclude it from subsequent analysis. The FINRA TRF prices, however, account for close to one fifth of total reported SPY volume for the day of the Flash Crash (Table 4 (a)). Significantly, these prices do not display a uniform deviation, but instead rapidly oscillate between the market consensus price and the apparent delayed price in a manner that would have been difficult or impossible for traders to interpret on a real-time basis during the Flash Crash.

Panel (b) of Figure 7 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., which we refer to as the “Data Flag Initiation Point” and “Data Flag Termination Point”, respectively, 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 the Data Feed Divergence Point to the Data Flag Initiation Point, 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. For ease of reference, we list the times above, in addition to the CME halt start and stop times, in panel (b) of Table 4.

As of May 6, 2010, FINRA Rule 6282 required that all off-exchange transactions in eligible securities (including SPY) 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 8 shows estimates of the delay in the upper portion of the FINRA TRF series, as depicted in Figure 7. To estimate the delay, we compute the maximum and minimum FINRA TRF times associated with each price $p \in [\$109.00, \$111.75]$, which spans the range of delayed prices between the Data Feed Divergence Point and Data Feed Flag Point. We then compute the median transaction time for Nasdaq OMX (not Nasdaq TRF) 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 8, aside from noise due to estimation error, the estimated delay increases monotonically and reaches 90 seconds precisely when the transactions are labeled as “out of sequence” in the FINRA TRF feed. 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.

The foregoing analysis suggests that there are three distinct time periods relevant to the analysis. Prior to the Data Feed Divergence Point documented in Figure 7, algorithmic traders would observe no divergence in reported trade prices among their various data feeds because no such divergence existed. However, from the Data Feed Divergence Point until the Data Flag Initiation Point, the FINRA TRF price series reported current prices and stale prices in an undifferentiated manner. Because of rapidly declining market prices, interspersing live trade information (showing a lower price) with stale trade information (showing a higher price), even though the stale prices were less than 90 sec-
onds old, generated a price series with significant oscillations. Only after those delayed prices became at least 90 seconds stale, at the Data Flag Initiation Point, did the FINRA TRF price feed accurately reflect out-of-sequence prices. At that point, the flagged and the unflagged feeds created two distinct price series with no oscillations. Those flags continued in place until the Data Flag Termination Point, when the 90 second delay was eliminated, even though stale prices continued to be reported as part of the FINRA TRF feed until some time approaching 2:56 p.m.

It bears emphasis that none of this analysis suggests that any market participant violated any trade reporting rule. To the contrary, late trades appear to have been properly marked in accordance with rules then in force once they were delayed by at least 90 seconds. The analysis opens the door to the possibility, however, that trades delayed by fewer than 90 seconds might have contributed to the evolution of the Flash Crash. Further, the foregoing analysis does not identify the source of the feed anomaly. However, the monotonic increase in the delay, and the demarcation as “out of sequence” at the precise time mandated by regulation appears most consistent with the hypothesis that a single entity reporting to the TRF was responsible for the observed delay and the concomitant price oscillations in the TRF data feed. The next section provides evidence supporting this hypothesis.

5.2 Source of the TRF Delay

To identify the source of the TRF anomalies, we obtained feed-specific data from one of the largest equities market participants on the day of the Flash Crash. According to a variety of criteria, the firm is considered a “high-frequency trader”\(^1\). The data consist of the firm’s records of individual subscriptions to proprietary equities-market data feeds, including NYSE, NYSE Arca, Nasdaq OMX, Nasdaq BX, BATS, BATS Y, Direct Edge A (EDGA) and Direct Edge X (EDGX). These separate proprietary feeds represent the highest quality market data used by firms to obtain an aggregate view of the

\(^1\)The firm has requested that it remain anonymous.
securities market for trading purposes. Indeed, as mentioned in multiple interviews that we conducted with traders\textsuperscript{2}, as well as similar interviews conducted by the CFTC/SEC staff (CFTC and SEC, 2010a), these proprietary feeds are considered to be highly superior to the CTS, both in terms of latency and overall integrity. It is important to note that in our analysis of the proprietary feed data, although we observed some delays between the source exchanges and the firm’s receipt of the messages, we did not find any delay even remotely as severe or consistent as is apparent in the TRF data. This finding, however, does not preclude the possibility that one or more of the feeds were responsible for delays at the CTS, because the path followed by the direct feeds to traders differs from the path by which information is conveyed to the FINRA TRF.

Closer examination of the EDGX data feed indicates that it is likely the sole cause of the anomalies observed in the FINRA TRF data feed. On March 12, 2010, prior to the Flash Crash, Direct Edge obtained SEC approval to convert its EDGA and EDGX ECNs to national securities exchanges\textsuperscript{3}. The transition became effective as of July, 2010, after the Flash Crash. Thus, as of May 6, 2010, both EDGA and EDGX were still reporting their transactions to the Nasdaq TRF. The total SPY volumes for EDGA and EDGX on May 6, 2010 were, 15,287,686 and 27,372,070 shares, respectively. Together, they comprised 40.75% of the 104,663,315 SPY shares reported at the TRFs. Given its central presence at the TRFs, we propose the following classification algorithm to determine the fraction of delayed TRF trades attributable to Direct Edge.

1. Select an origin data feed, $ECN \in \{\text{EDGA, EDGX}\}$, to search for TRF matches.

2. For each TRF transaction, consisting of triplet $\{Time_{TRF}, Price_{TRF}, Volume_{TRF}\}$, search for the first matching transaction pair $\{Price_{TRF}, Volume_{TRF}\}$ in $ECN$ during the 10 minutes prior to $Time_{TRF}$.

3. If a match is found, record $\{Time_{ECN}, Time_{TRF}, Price_{TRF}, Volume_{TRF}\}$.

\textsuperscript{2}Our interviews with market participants were subject to assurances of confidentiality.

\textsuperscript{3}Direct Edge was subsequently acquired by BATS Global Markets on January 31, 2014.
4. (Optional) If a match is found, delete the transaction from ECN before proceeding to the next TRF record.

We implement the search for both EDGA and EDGX ECNs. In addition, as indicated in step 4 of the classification method, we run the algorithm in two ways: destructively (iteratively deleting each matched record) and non-destructively (allowing multiple FINRA transactions could to be matched to the same Direct Edge transaction).

Table 5 reports the results of our matching algorithm on several subsets of the FINRA TRF transactions. Specifically, we separated the TRF records between the Data Feed Divergence Point (2:42:45:538) and the Data Flag Termination Point (2:54:08:272) into four distinct subsets: (1) the apparently delayed TRF transactions between the Data Feed Divergence Point and the Data Flag Initiation Point, (2) the apparently correct TRF transactions between the Data Feed Divergence Point and the Data Flag Initiation Point, (3) the out-of-sequence TRF transactions between the Data Flag Initiation Point and the Data Flag Termination Point, and (4) the in-sequence TRF transactions between the Data Flag Initiation Point and the Data Flag Termination Point. While subsets 3 and 4 are easily distinguished by the out-of-sequence code in the Daily TAQ file, we apply a linear interpolation method to classify the trades in subsets (1) and (2). We then applied our matching algorithm, in the two modes (destructive and non-destructive) described above, to each of the TRF sub-series, separately searching for matches in both the EDGX and EDGA feeds. When utilizing the destructive version of the algorithm, we first conduct the search for the apparently delayed and out-of-sequence TRF transactions.

Panel (a) of Table 5 reports matches when using EDGX as the candidate source. The destructive and non-destructive search methods both yield match rates between 90% and 95% for the period 1 delayed and period 2 out-of-sequence TRF transactions. In contrast, the match rates are much lower for the period 1 non-delayed and period 2 in-sequence transactions: the non-destructive search results in a match rates, respectively, of 42% and 58%, with those values dropping precipitously to 7% and 27% for the destructive search. Panel (b) reports equivalent results for EDGA. Here, however, the non-destructive match rates never exceed 58% and the destructive match rates are, again, substantially
Table 5: FINRA TRF and Direct EDGE ECN matching results. Panel (a) considers EDGX as the source and panel (b) considers EDGA as the source. The first column reports which subset of the TRF transactions is treated as the target and the second column reports the total number of transactions reported in the subset. The remaining columns report the number and fraction of matches in the TRF using the destructive and non-destructive versions of the algorithm.

Reduced, ranging between 25% and 42% for the different TRF subsets. These results provide strong evidence that the EDGX ECN was either largely or fully responsible for the delayed reports at the TRF.

This search procedure is susceptible to the occurrence of false positives, as demonstrated by the match rates for the non-delayed and in-sequence TRF transactions which show non-negligible probabilities that they originate in both the EDGX and EDGA feeds. To help determine the legitimacy of matches, we investigate patterns in the difference of their recorded times: $d_{ECN} = Time_{TRF} - Time_{ECN}$, for $ECN \in \{EDGA, EDGX\}$. Panel
(a) of Figure 9 depicts $\log(d_{EDGX})$ for period 1 delayed TRF transactions and period 2 out-of-sequence transactions. Panel (b) shows the same quantity for period 1 non-delayed and period 2 in-sequence transactions. The upper panel demonstrates a clear, nearly deterministic evolution of $d_{EDGX}$ through time: the delay monotonically increases from zero at the Data Feed Divergence Point until the Data Flag Initiation Point, where the value on the vertical axis corresponds to 90 seconds. After that point in time (when regulation required that out-of-sequence trades be reported as such), the delay continues to increase to a maximal value of approximately 180 seconds ($\log(d_{EDGX}) = 5.19$), before starting a steady decline through the remainder of the period. In contrast, the
Figure 10: Natural logarithm of time difference between TRF and EDGA time stamps for matched transactions. Panel (a) shows differences for the period 1 delayed and period 2 out-of-sequence transactions and panel (b) shows differences for the period 1 non-delayed and period 2 in-sequence transactions.

The lower panel does not depict a clear evolution of the delay; rather, it exhibits stochastic variation close to zero, with some departures.

Figure 10 depicts that same quantity for the TRF/EDGA matched transactions. A similar qualitative pattern emerges, but the period 1 delayed and especially the period 2 out-of-sequence transactions do not exhibit an evolution that is as clearly deterministic as of those attributed to EDGX. Taken as a whole, the matching statistics and the patterns of their attributed delays strongly suggest that EDGX is primarily responsible for the delayed SPY transactions reported to the FINRA TRF.
5.3 Impact of the TRF Delay

A clear mechanism of action links the price oscillations observed between the Data Feed Divergence Point and the Data Flag Initiation Point to the withdrawal of liquidity and the precipitation of the most rapid portion of the Flash Crash price decline. The nature of these oscillations, which were as large as 3 percent in the liquid SPY ETF (and in other symbols reported to the TRF by EDGX), is illustrated in Figure 11. Panel (a) 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 was not until approximately 15 seconds later, when the erroneous FINRA prices likewise experienced a halt of approximately equal duration that the market noticeably turned upward from its trough.

Panel (b) of Figure 11 zooms in on FINRA TRF trade prices for SPY for the five second interval following 2:45:20 p.m., a period between the Data Feed Divergence Point and the Data Flag Initiation Point, when the FINRA data feed was not flagging late trades as delayed. During this span, the higher, stale but unflagged prices of approximately $110.4 were interspersed with lower, timely prices of approximately $107.6, giving rise to an oscillation band of approximately 2.6 percent of the timely, accurate prices. However, because all of these prices were reported in an undifferentiated manner, market participants relying on the FINRA TRF feed would not be able to explain the cause of the apparent oscillation in reported prices.

It is possible that the CTS price oscillations could have caused trading algorithms to reduce activity in both equities and futures contracts, notably the E-mini. Coupled with the E-mini selling pressure from Waddell & Reed, this would have lead to an abrupt de-
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.
cline in futures prices, which would have been transmitted first to ETFs and subsequently to component stocks. However, such a transmission mechanism relies on the algorithms for large trading firms using the CTS as an input.

In accordance with the Securities and Exchange Act of 1934, all national securities markets are required to report transactions and quotations to a centralized securities information processor (SIP). The Consolidated Tape Association was created to oversee the national market system, establishing a SIP to aggregate market information. SEC regulation NMS requires the “prompt” reporting of information to the SIP, but as mentioned at the beginning of the previous section, sophisticated traders are typically able to construct their own aggregate view of the market by directly subscribing to proprietary data feeds of individual exchanges. Relying on these proprietary feeds eliminates the two-step latency of information transmission between trade origination (exchanges) and the SIP and subsequently between the SIP and financial services firms, and therefore generates more timely data. In our interviews with several low-latency trading firms, we learned that none utilize the CTS as a primary data source and that nearly all were subscribers to the EDGA and EDGX proprietary feeds, despite the fact that Direct Edge was not yet operating as a national securities exchange. CFTC and SEC (2010a) confirms that traders report that the CTS is not used as a primary data input for trading algorithms. However, many of the same firms acknowledged that the CTS is used as secondary source for data integrity checks:

Most of the firms we interviewed that are concerned with data latency in the milliseconds...subscribe directly to the proprietary feeds offered by the exchanges. These firms do not generally rely on the consolidated market data to make trading decisions and thus their trading decisions would not have been directly affected by the delay in data in this feed. However, some of these firms do use the consolidated market data feeds for data-integrity checks, and delay-induced data discrepancies certainly contributed to the general sense of unease experienced that day.
Other firms that are not concerned with data latency in the milliseconds...tend to rely on the consolidated market data feeds for trading decisions. A number of those interviewed reported pulling back from the market as general volatility increased, and those seeing delays and price-discrepancies on the consolidated market data feeds did report that was a contributing factor in their decision to curtail or halt further trading. (CFTC and SEC, 2010a, pp.36-37)

The potential mechanism of action linking FINRA price oscillations to the withdrawal of liquidity and subsequent sharp price decline is thus clear: low-latency traders who relied on the FINRA feed as a data-integrity check would have observed unexplained, sudden, and large variations in their profits and losses and in other data integrity testing regimes, in addition to the pricing discrepancies observed in a comparison of their proprietary data with the CTS feed. The presence of these unresolved oscillations could rationally cause them to exit the market until the oscillations were resolved. Similarly, to the extent that other algorithmic traders relied on the FINRA feed as a primary trading source, they would have observed unusual behavior in their trading models that could also have caused them rationally to exit the market.

A priori, we have no information as to whether price divergences in the CTS data, similar to that depicted in Figure 7, are common. To document 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 (c) 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 the Data Feed Divergence Point, with 439 occurring between the Data Feed Divergence Point and the Data Flag Initiation Point and 326 occurring between the Data Flag Initiation Point and the Data Flag Termination Point.
From these data, we conclude that price deviations of the type in Figure 7 are extremely rare in the CTS data, given the dates we examined.

Our work in Section 3.1 provides strong evidence that the Flash Crash originated in the market for the E-mini: that market was tightly coupled with and leading the equities markets all the way through the trough of the decline. Interviews conducted by the CFTC/SEC staff (CFTC and SEC, 2010a, p.17) corroborate this claim. Thus, if the TRF delay was a contributing factor in the Flash Crash, it would have first caused liquidity withdrawal in the E-mini market which would then have propagated to ETFs and individual securities. To measure the impact of the TRF price oscillations on E-mini returns, we augment the SVAR regression in Equation 3 with an additional variable: $|\Delta P_{t}^{TRF}|$, which represents the magnitude of differences of the TRF SPY price from the market consensus SPY price over 5 second time intervals. Cumulative impulse response coefficients are reported in panel (b) of Table 2. Identical to panel (a), we report return responses for 2400 contract impulses to each group of levels of the order book, however we now also report responses for an impulse of $|\Delta P_{t}^{TRF}| = $1. To put this latter impulse into context, during the period between the Data Feed Divergence Point and the Data Flag Initiation Point, the median, mean and maximum values of $|\Delta P_{t}^{TRF}|$ were, respectively, $0.64$, $1.28$ and $3.68$. While the impulse responses related to order imbalances are almost identical in terms of magnitude and statistical significance to their counterparts in panel (a), the impulse responses related to the TRF price oscillations are quite significant and large: over a two-minute horizon, they achieve a cumulative impact of nearly 47 basis points. In short, a single $1$ price oscillation over a 5-second time interval has the capacity to explain 10% of the market decline during the Flash Crash. Considering the fact that roughly 48 of these (admittedly, highly-correlated) impulses occurred over a 4-minute time window, our results suggest that the TRF price oscillations may have been a major contributing factor to the E-mini price decline.

To be sure, the TRF price oscillation variable, $|\Delta P_{t}^{TRF}|$, experiences all of its variation at the time when the dependent variable of interest, $r_{t}$, also experiences its greatest variation, and so we must be careful about resulting inference. However, we maintain
that this is no worse than including the imbalance measures in the SVAR, since they likewise experience their greatest variation at precisely the moment of the market decline. In addition, to some extent, the SVAR is designed to account for such contemporaneous relationships among the variables.

In conclusion, our analysis 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. Because this data feed anomaly was apparent to algorithmic 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 algorithmic 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).

6 Market Liquidity

To test our hypothesis that TRF price oscillations contributed to the Flash Crash, we focus on liquidity provision during the minutes surrounding the Flash Crash and examine whether liquidity was removed at specific times associated with the TRF price oscillations. We further test whether TRF price oscillations Granger-caused liquidity removal.
6.1 Inside Depth

Figure 12 displays Nasdaq OMX SPY liquidity at fixed spreads from the market mid price. Specifically, rather than show the number of shares at the ten best price levels, the figure shows the number of SPY shares available on bid (blue) and offer (red) at the ten price levels closest to the mid-price. Due to heavy trading interest in SPY, the ten closest prices almost always represent the ten best price levels. Depth at the ten spreads is depicted as stacked bars, with the lightest shades of blue and red corresponding to a spread of 0.5 cents from the mid (typically level 1) and the darkest shades corresponding to a spread of 9.5 cents (typically level 10). Panel (a) of the figure depicts bid/offer liquidity during the period 2:40:00 – 2:48:00 p.m. EDT and demonstrates that the inside levels of the order book were heavily populated with shares until the Data Feed Divergence Point, which is represented by the first vertical red line. After that point, there was a marked decrease in liquidity, with a brief trough at 2:43:40 and subsequent momentary return of liquidity until about 2:44:20, after which liquidity at the inside market prices rapidly disappeared and only sporadically reappeared in low quantities prior to 2:48:00. The gray bars in the panel depict the magnitude (in dollars) of the TRF price oscillation, and demonstrates how the increase and decrease in the price discrepancy led the disappearance and reappearance of liquidity at the inside prices.

Panel (b) of Figure 12 shows the same data as panel (a), but focuses closely on the period 2:45:00 – 2:47:00 p.m. EDT. This panel more clearly highlights the relationship of the TRF oscillation to liquidity at inside prices: shortly after the peak oscillation at 2:45:34, inside liquidity completely disappears. Likewise, the first substantive cluster of new shares at the inside prices occurs after the TRF oscillation trough at 2:46:42. It is notable that SPY market depth, especially on the bid, persisted throughout the 5-second CME market halt.

Similar inspection of the E-mini market shows that depth at inside prices never faltered. Analogous to the figure above, Figure 13 shows number of shares on bid and offer at the four prices closest to the ES mid point (since the ES has a minimum price
Figure 12: Bid/offer order book depth at fixed spreads from the SPY mid price, sampled every 1 second. Offer depth is depicted as stacked red bars, with the lightest (darkest) shade corresponding to a spread of $0.005 ($0.095). Bids are analogously depicted in blue. The gray bars show the magnitude of the TRF price oscillation. Panel (a) isolates the time period 2:40:00 – 2:48:00 p.m. EDT and panel (b) shows the same data for the period 2:45:00 – 2:47:00 p.m. EDT. The vertical red lines correspond to the Data Feed Divergence Point, the CME Halt Starting Point and the Data Flag Initiation Point, respectively.
increment that is 2.5 times wider than that of SPY, the four inside prices correspond to the same total “distance” from the mid price as the ten SPY prices). This is most likely attributed to the fact that the E-mini is a much more heavily traded asset than SPY and also to the fact that the rapid price decline caused the inside price levels to be populated by shares that were originally intended to be at prices deep in the book. This latter point is investigated below, with aggregated E-mini order cancellations and additions during the same period of time.

Figure 13: Bid/offer order book depth at fixed spreads from the ES mid price, sampled every 1 second. Offer depth is depicted as stacked red bars, with the lightest (darkest) shade corresponding to a spread of 0.25 (1.75) index points. Bids are analogously depicted in blue. The gray bars show the magnitude of the TRF price oscillation. The vertical red lines correspond to the Data Feed Divergence Point, the CME Halt Starting Point and the Data Flag Initiation Point, respectively.

Figure 13: Bid/offer order book depth at fixed spreads from the ES mid price, sampled every 1 second. Offer depth is depicted as stacked red bars, with the lightest (darkest) shade corresponding to a spread of 0.25 (1.75) index points. Bids are analogously depicted in blue. The gray bars show the magnitude of the TRF price oscillation. The vertical red lines correspond to the Data Feed Divergence Point, the CME Halt Starting Point and the Data Flag Initiation Point, respectively.
6.2 Evidence from Order Attributions

The majority of orders in our level-2 Nasdaq OMX data are not attributed, and we therefore have no knowledge of trader identities. However, according to Nasdaq Rule 4613, designated market makers are required to maintain a continuous two-sided quote with an attribution that identifies the order originator with a Market Participant Identifier (MPID). In some cases, market makers meet this obligation by posting attributed stub quotes while quoting more aggressively with non-attributed orders. In addition, non-market makers utilize attributed messages to attract trading interest and to allow clients to verify orders placed on their behalf. Of the 6,872,223 messages on May 6, 2010, 1,071,040 (15.6%) were associated with an MPID attribution.

Despite the infrequent use of order attribution, we use the MPID field as a vehicle to obtain information about the order placement behavior of specific algorithms during the Flash Crash. Panel (a) of Figure 14 shows the net addition of attributed orders (number of shares added, less the number of shares canceled) for all tickers over 10 second intervals during the period 2:37:00 – 3:00:00 p.m. EDT. As with the previous plots, the vertical red lines correspond to the Data Feed Divergence Point, the CME Halt Starting Point and the Data Flag Initiation Point, respectively. The figure shows that the period prior to the CME Halt Starting Point was punctuated by several instances of massive deletions and that the subsequent period experienced both large-scale additions and deletions, with an overall net gain to the order book (additions). The remaining panels of the figure focus on four specific MPIDs which were primarily responsible for the deletions prior to the CME Halt Starting Point: TMBR (Timber Hill LLC), HDSN (Hudson River Trading LLC), SUSQ (Susquehanna International Group) and UBSS (UBS Securities LLC). Note that we allow the scales of the vertical axes to vary across panels in order to improve visualization of the large cancellation events. These latter panels show that millions of shares were canceled from the order book in the minutes prior to and during the Flash Crash. Further, although not easily seen at the present scales, TMBR, HDSN and SUSQ completely curtailed their attributed messaging for long stretches of time during the most
severe decline and recovery. The data thus demonstrate that sophisticated algorithmic traders withdrew large amounts of liquidity within seconds of the appearance of the TRF price oscillations in a manner consistent with the previously described mechanism of action.

Although the signature of attributed trading is not necessarily indicative of non-attributed trading, it is reasonable to assume that the exogenous event that led to these massive message flows for attributed orders would have also caused similar flows in non-attributed messages. Specifically, the lack of attributed messaging for TMBR, HDSN and SUSQ suggests that their order placement algorithms were temporarily shut off throughout much of the Flash Crash period, a fact corroborated by CFTC and SEC (2010a). This individual-scale evidence also supports the withdrawal of liquidity documented in Section 6.1, which, for these particular firms, would have been a result of using the CTS, and hence the FINRA TRF feed, as a secondary source for data integrity checks (CFTC and SEC, 2010a).

### 6.3 E-mini Order Cancellations

As already established in Section 5.3, the E-mini led the rapid market decline on May 6, 2010. However, while Figures 12 and 14 show massive order cancellations in the equities market, Figure 13 emphasizes that the market for the E-mini retained typical depth at inside prices. To understand the nature of inside depth in conjunction with plummeting prices, Figure 15 depicts the number of contracts canceled at the best bid (blue) and best offer (red) when all contracts at that price were removed from the order book. The removal of contracts at the inside market prices is either associated with trade executions, cancellations, or a combination of the two. Panels (a) and (b) isolate the portion of bid/offer removal tied to cancellations over 10-second time intervals for the periods 2:40:00 – 2:55:00 and 2:40:00 – 2:45:00 p.m. EDT, respectively. As before, the vertical red lines correspond to the Data Feed Divergence Point, the CME Halt Starting Point and the Data Flag Initiation Point, respectively, and the gray bars correspond to the magnitude of the TRF price oscillations. Panel (b) shows that E-mini cancellations were
Figure 14: Net shares added for MPID messages at Nasdaq OMX, aggregated across 10-second intervals. Panel (a) aggregates across all MPIDs and the remaining panels correspond to specific MPIDs that had a very large impact. The vertical red lines in each panel correspond to the Data Feed Divergence Point, the CME Halt Starting Point and the Data Flag Initiation Point, respectively.

rare prior to the Data Feed Divergence Point, but that they rapidly increased, especially on the bid side, during the most precipitous portion of the market decline. Notably, these cancellations appear to be preceded by the TRF oscillation. This highlights that bid
depth was being removed even though the depth at the inside spreads never disappeared (as they did in the SPY market). Subsequent to the CME Halt Starting Point and the pause in the TRF oscillations, during the recovery period, panel (a) shows a net increase in cancellations on the offer side of the order book, which would naturally be associated with the rapid market rebound.

To place the E-mini order cancellations of Figure 15 into perspective, we compute the number of bid and offer cancellations for each of the trading days in the random sample of 34 days that we analyzed in Section 3. Only four days in the sample experienced cancellations in excess of 5000 contracts (bid and offer combined): 44,409 on May 6, 2010, 6,584 on Mar 10, 2011, 20,154 on Aug 9, 2011 and 8,096 on Aug 25, 2011. Isolating the remaining days, the mean and standard deviation of cancellations per day are 940 and 611, respectively. Further, on May 6, 2010, 39,095 (88%) occurred during the 15-minute period 2:40:00 – 2:55:00 p.m. EDT. It is clear that the period depicted in Figure 15 was exceptional in terms of E-mini cancellations.

Table 6 reports bivariate Granger causality tests for the null hypotheses that the TRF oscillations do not Granger cause E-mini bid/offer cancellations, and also that E-mini bid/offer cancellations do not Granger cause the TRF oscillations. We separately compute F-statistics for lags $p \in \{1, 2, 3, 4, 5\}$ (each lag corresponds to an interval of 10 seconds) for the data observed during 2:40:00 – 2:55:00 p.m. EDT. P-values for the F-statistics are reported in parentheses. The first row shows evidence that the TRF oscillation does Granger cause E-mini bid cancellations, especially for lags 3, 4 and 5. The same is not true of offer cancellations, where the null hypothesis is only rejected at a 10% significant level for lags 1 and 3. Testing the reverse set of hypotheses, that E-mini bid or offer cancellations Granger cause the TRF oscillation, results in no rejection at any lag.

Granger causation does not necessarily establish true causation. In their weakest form, our results merely show that the TRF price oscillation precedes and has predictive power for E-mini bid cancellations. If the forces that caused the TRF delay were exogenous to other factors that created downward pressure on the market (e.g., an independent
Figure 15: E-mini contract cancellations aggregated across 10-second intervals. Red bars correspond to offer cancellations and blue bars to bid cancellations. The gray bars show the magnitude of the TRF price oscillation. Panel (a) isolates the time period 2:40:00 – 2:55:00 p.m. EDT and panel (b) shows the same date for the period 2:40:00 – 2:45:20 p.m. EDT. The vertical red lines correspond to the Data Feed Divergence Point, the CME Halt Starting Point and the Data Flag Initiation Point, respectively.
Table 6: F-statistics (p-values in parentheses) for Granger causality tests involving the TRF price oscillation and E-mini bid and offer cancellations. The data comprise the period 2:40:00 – 2:55:00 p.m. EDT, sampled at 10-second intervals.

<table>
<thead>
<tr>
<th>Test</th>
<th>$p = 1$</th>
<th>$p = 2$</th>
<th>$p = 3$</th>
<th>$p = 4$</th>
<th>$p = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRF $\rightarrow$ Bids</td>
<td>3.03 (0.0855)</td>
<td>2.35 (0.1013)</td>
<td>3.89 (0.0119)</td>
<td>2.68 (0.0378)</td>
<td>2.46 (0.0405)</td>
</tr>
<tr>
<td>TRF $\rightarrow$ Offers</td>
<td>2.87 (0.0938)</td>
<td>2.25 (0.1113)</td>
<td>2.47 (0.0681)</td>
<td>1.65 (0.1708)</td>
<td>1.56 (0.1832)</td>
</tr>
<tr>
<td>Bids $\rightarrow$ TRF</td>
<td>2.09 (0.1523)</td>
<td>1.30 (0.2776)</td>
<td>1.08 (0.3623)</td>
<td>0.963 (0.4328)</td>
<td>0.975 (0.4389)</td>
</tr>
<tr>
<td>Offers $\rightarrow$ TRF</td>
<td>1.31 (0.2557)</td>
<td>1.82 (0.1692)</td>
<td>1.43 (0.2412)</td>
<td>1.22 (0.3073)</td>
<td>1.06 (0.3873)</td>
</tr>
</tbody>
</table>

6.4 Summary

The estimated SVAR of Section 5 suggests that the oscillation in TRF prices had a statistically significant impact on E-mini prices. The work of this section, in contrast, has pointed to a relationship between the oscillation and liquidity provision. Clearly, prices and liquidity provision are intimately related, as the withdrawal of liquidity, especially on one side of the market, could rationally cause strong directional moves in prices.

From first-hand interviews, we know that many trading firms curtailed their activity due to data integrity problems. We believe that delay in the TRF feed was a substantial contributor to these failed data integrity checks. Our work above documents that subsequent to the emergence of the price oscillation, a number of firms withdrew from the market, and that liquidity at inside market prices in the SPY collapsed. While a similar collapse did not occur in the E-mini, primarily due to its greater depth of market, bid-side cancellations increased dramatically, which, joint with the massive selling pressure (both passive and aggressive) from Waddell & Reed, caused the price to plummet. This rapid price decline was then transmitted to the equities market, which, with very thin liquidity, caused the SPY and many other equities prices to likewise experience massive
movements.

It is important to note that the reappearance of liquidity at the inside of the SPY market coincides with the disappearance of the TRF oscillation: when the delayed price series in Figure 11 crosses the consensus price series at roughly 2:46:45, the inside prices of the order book are again populated with orders. The addition of liquidity at this moment can be seen both in Figure 12 and Figure 11, where the variability of the traded price series for Nasdaq (gray line) suddenly reduces.

7 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 that is somewhat more mundane than many previous summations in the literature. The macroeconomic environment was uncertain, and 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 was high, but the level of concern generated by these two factors was by no means unprecedented. We have specifically demonstrated that trading rates, intraday volatility, and price trajectories on August 9, 2011 were extremely similar to 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 far from the inside levels of the order book, and we have demonstrated 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 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, or fined.

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, as we suggest, 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. 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 one 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. 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.

Our work instead suggests that policymakers interested in reducing the probability of
a future Flash Crash should place greater emphasis on challenges related to the integrity of data feed information. 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 during the Flash Crash. Such declines can be exacerbated by large sell orders, such as the one placed by Waddell & Reed, and can be arrested by the entrance of fundamental buyers to the market.

The close relationship between the data feed anomalies established in our analysis and the observed withdrawal of liquidity, combined with the fact that prices continued to decline after the CME halt and reversed only after the data feed anomaly was resolved, suggests that the lack of integrity in information flows contributed to the market’s steep decline. Trading halts alone may therefore be insufficient to address extreme market volatility if data feeds are generating unreliable information. Greater attention to the integrity of data – particularly in a market dominated by algorithmic and low-latency trading – may be central component of an optimal policy response to the Flash Crash of 2010. Regulatory actions have already reduced the time periods during which traders can report undesignated late trades from the 90-second period that prevailed during the Flash Crash to the current window of 10 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. Greater investment in data integrity may therefore be a prudent, if low-visability, response to the Flash Crash.
References


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


