Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Flash Crash: A Cautionary Tale About Highly Fragmented Markets

Albert J. Menkveld, Bart Zhou Yueshen

To cite this article:
Albert J. Menkveld, Bart Zhou Yueshen (2018) The Flash Crash: A Cautionary Tale About Highly Fragmented Markets. Management Science

Published online in Articles in Advance 05 Nov 2018

. https://doi.org/10.1287/mnsc.2018.3040

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–19
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Flash Crash: A Cautionary Tale About Highly Fragmented Markets

Albert J. Menkveld,[a] **Bart Zhou Yueshen**[b]

[a] VU University Amsterdam, Tinbergen Institute, and Duisenberg School of Finance, FEWEB, 1081 HV, Amsterdam, Netherlands; [b] INSEAD, Singapore 138676

Contact: albertjmenkveld@gmail.com, http://orcid.org/0000-0002-9913-9242 (AJM); b@yueshen.me
http://orcid.org/0000-0002-4326-3658 (BZY)

Received: April 26, 2016
Revised: May 16, 2017; November 15, 2017
Accepted: January 4, 2018
Published Online in Articles in Advance: November 5, 2018

https://doi.org/10.1287/mnsc.2018.3040

Abstract. A breakdown of cross-market arbitrage activity could make markets more fragile and result in price crashes. We provide suggestive evidence for this novel channel based on a high-frequency analysis of the most salient crash in recent history: The Flash Crash. We further show that such an event can be extremely costly for a large seller trading in a particular venue as the seller effectively relies on local liquidity supply only. These findings highlight the vulnerability of today's highly fragmented markets.

History: Accepted by Gustavo Manso, finance.
Funding: Menkveld gratefully acknowledges VU University Amsterdam for a VU talent grant and NWO for a VIDI grant.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2018.3040.

Keywords: flash crash • large seller • electronic market • broken arbitrage

1. Introduction

On May 6, 2010, U.S. equity indices declined by 5%–6% and recovered, all in 30 minutes: an event dubbed the Flash Crash. The crash originated in the market for E-mini contracts, which are futures on the S&P 500 index. It rapidly spread not only to other index products, but also to individual stocks (CFTC and SEC 2010a, b). The crash echoed internationally. Canadian markets, for example, crashed within two minutes (IIROC 2010).

There is widespread concern that Flash Crash–type events are the result of vulnerable electronic markets. Arguably the most damaging effect is that it may scare off market participants. For example, the Investment Company Institute (ICI) claimed that, as a result of the Flash Crash, "there have been five consecutive months of U.S. equity outflows" (Zhang and Powell 2011, p. 11). A CFTC-SEC Advisory Committee[1] writes, "the net effect of that day was a challenge to investors' confidence in the markets" (Born et al. 2011, p. 2). The SEC chairperson echoed that fear when she opened a special 2012 roundtable triggered by the Flash Crash: "…our concern is not whether a single firm might fail, but whether it causes collateral damage to investors and their confidence in the integrity and stability of our markets."[2]

Several months after the crash, U.S. regulators released a study that highlighted the key role of a large seller. Reportedly, this seller initiated a sell program of 75,000 E-mini contracts worth approximately $4.1 billion (CFTC and SEC 2010a). He did so in a market where order flows grew more toxic by the hour (Easley et al. 2012). Eventually, the willingness of intermediaries to hold inventory saturated (Kirilenko et al. 2017), high-frequency traders engaged in so-called hot-potato trading (CFTC and SEC 2010a), and the E-mini price collapsed as a result. It was followed by price declines in related markets—first index-tracking exchange-traded funds (ETFs), most notably SPY, and then constituent stocks themselves (Ben-David et al. 2012).

This paper develops a novel angle to look at the Flash Crash: cross-arbitrage. In normal times, the E-mini futures market is populated with cross-market arbitrageurs who absorb shocks in this S&P 500 future and offload their positions in the broader market (e.g., by trading SPY or other ETFs tracking the index, constituent stocks, or replicating derivatives). During the Flash Crash, such cross-arbitrage activity reportedly broke as "many liquidity providers temporarily paused" (CFTC and SEC 2010a, p. 4). We analyze such cross-arbitrage activity formally and establish that it broke a half hour before the Flash Crash.

We study cross-arbitrage (i) by developing new measures to gauge the level of cross-arbitrage activity; (ii) by sequencing cross-arbitrage breakdown times and price crashes; and (iii) by offering a new narrative of the Flash Crash, weaving together the existing evidence, cross-arbitrage breakdown, and the large seller's trades. This last analysis is made possible by unique access to a data set that contains the large seller's trades.

Our proposed measure for cross-arbitrage relies on competitive cross-arbitrageurs' actively quoting for (essentially) the same security on different venues. Consider one such arbitrageur who is hit on his bid in E-mini. The arbitrageur might offload this position profitably on SPY by taking the highest bid there (if it is above the arbitrageur's bid that was just taken). Such

behavior in and of itself synchronizes prices across venues. In addition, other cross-arbitrageurs observe these aggressive sells, learn, and lower their bids and asks on all venues. This is an additional channel by which prices synchronize across venues. Guided by this insight, we use the cross-sectional variation in bids and such variation in asks as a proxy for cross-arbitrage activity. We interpret relatively high "quote dispersion" as indicative of weakened cross-arbitrage activity. Such interpretation relies on our conjecture that cross-arbitrage exercises the strongest discipline on prices across venues. Other forces, such as multi-venue execution algorithms or single-venue market makers learning from other-venue quotes, also discipline prices across markets, but we believe they are not as strong as cross-market arbitrage.

Our empirical analysis shows that cross-arbitrage severely weakened and eventually broke just ahead of the one-minute steep price drop that is referred to as the Flash Crash. To add further identification of cross-arbitrage breakdown preceding a crash, we exploit the cross section of individual stocks on May 6. In a sample of stocks that crashed most, the sequencing of cross-arbitrage breakdown correlates positively with the sequencing of price crashes. In other words, the stocks that exhibited cross-arbitrage breakdowns first were the first to crash. This relationship holds up after adding several standard control variables, such as volatility and order imbalance. We, however, are unable to identify the root cause for these breakdowns themselves. We nevertheless speculate about possible causes in Section 4.

These results suggest the following economic interpretation of the Flash Crash: If trading is fragmented across venues, then cross-arbitrage glues these venues together. Intermediaries might buy in one venue to potentially sell in another and, thus, effectively connect (end-user) buyers and sellers across these venues. Absent cross-arbitrage, however, a liquidity demander selling heavily in one venue may pay a high price for liquidity as the liquidity demander effectively relies only on local liquidity suppliers only. This intuition could be formalized by, for example, reinterpreting Grossman and Miller (1988). They parameterize liquidity supply in terms of the presence of M symmetric risk-averse market makers. If these market makers are spread out equally across two venues that become disconnected, then a liquidity demander effectively experiences liquidity supply from only $M/2$ market makers and therefore pays higher price pressure. In general, such pressure could be so high that it manifests itself as a price crash.

Such an interpretation of the Flash Crash is consistent with the empirical observation that after the cross-arbitrage breakdown, E-mini prices recovered more slowly than SPY prices. Further analysis of the large seller's trades shows that he traded most of his 75,000 contracts in the minutes *after* the breakdown. This kept selling pressure high for E-mini and, in absence of cross-arbitrage, the price differential between E-mini and SPY continued to widen. At the peak of this price wedge, the E-mini ask price was *more than 100 basis points* below the SPY bid.

Finally, the large seller's trade data allow us to quantify how much this chain of events cost him. A calibration based on Grossman and Miller (1988) suggests that the large seller overpaid for the liquidity he demanded. Our cost estimates range from \$98.6 to \$229.8 million. Such cost surely was sizeable for the large seller as it amounted to two to four times his quarterly operating income at the time.

End users paying more for liquidity at times of crashes is a disturbing finding in particular since the Flash Crash was by no means unique. Similar crashes occurred in, for example, the German DAX index (August 18, 2011, and April 17, 2013), the oil price (May 5, 2011), India's leading equity index (October 5, 2012), the 10-year U.S. Treasury (October 15, 2014), the pound sterling (October 7, 2016), and more recently in cryptocurrencies such as Ethereum (June 21, 2017) and Bitcoin (October 10, 2017).

Our paper adds to a rapidly growing literature on the Flash Crash. Kirilenko et al. (2017) use audit trail data to document that high-frequency traders (HFTs) did not withdraw from trading E-mini at the time of the crash. Kyle and Obizhaeva (2016) calibrate the Flash Crash price drop to their market microstructure invariance model and conclude that (p. 31) "the predicted price impact is smaller than the actual decline." This finding corroborates our calibration exercise. Easley et al. (2012) document that E-mini trading steadily grew more toxic in the hours before the crash. Ben-David et al. (2012) document that the E-mini price collapse was followed by price declines in related markets, first index-tracking exchange-traded funds (ETFs), then by the constituent stocks. Madhavan (2012) documents that hardest hit were ETFs that traded in fragmented markets. Borkovec et al. (2010) argue that ETFs collapsed because of extreme liquidity deterioration, not because of a technical failure of the market structure or the ETF product as such. Our paper complements this literature by its focus on the impaired cross-arbitrage ahead of price crashes. In particular, it could explain Madhavan's finding that ETFs in strongly fragmented markets suffered most. These are the ETFs that rely most on cross-arbitrage.

The extreme illiquidity in E-mini and its contagious effect on other markets are possibly explained by two recent theoretical studies. Cespa and Foucault (2014) generate such contagion by a feedback loop from a liquidity shock in one market making its price less informative and, therefore, raising uncertainty for suppliers in another market, who, in turn,

reduce their liquidity supply, making the price in that market less informative, thus further hampering liquidity supply in the original market, etc. Goldstein et al. (2013) also generate excess volatility by agents learning from prices in multiple markets. They add an endogenous liquidity demander (hedger) and generate excess volatility through a multiple-equilibria result, which arises through a learning complementarity across potential speculators. Both papers relate their findings to the Flash Crash.

The rest of our paper is organized as follows. Section 2 presents the data. Section 3 provides an overview of the Flash Crash by putting all main events on a timeline, including the events identified in this study. Section 4 studies how broken cross-arbitrage relates to price crashes. Section 5 discusses the large seller's trading. Section 6 concludes. The appendices provide technical details, examine data integrity, and present robustness analysis.

2. Data

The data used in this paper are supplied by Nanex, a firm that specializes in low-latency ("real time") distribution of trade and quote data to its clients. One data set contains an ordered sequence of all May 6, 2010, order book events (trades and book changes) in the June 2010 E-mini futures contract. A second data set contains similar information for an important index tracker: SPY.

All events carry a time stamp with a granularity of 25 milliseconds. They are recorded in Eastern Standard Time (EST). One advantage of this data is that its creator is an information consolidator and distributor, which guarantees consistency in event sequencing and time stamps. In other words, the data captures what low-latency participants saw and when they saw it.[3]

The level of detail on market events differs across the E-mini and SPY data sets. The E-mini contract trades only on the Chicago Mercantile Exchange. The data feed is very detailed as each order book change or trade is recorded in the database. SPY, however, traded on eight different exchanges on May 6, 2010: BATS, BOST, CBOE, CHIC, CINC, ISEX, NQES, and PACF. Its data feed is less detailed as it records only trades and changes in a market's best bid or ask (order book changes away from the best quotes are not recorded).

A separate, proprietary data set was provided by Waddell and Reed, which contains all trades by the large seller who featured prominently in the CFTC-SEC report (2010a). These trades can be matched with E-mini trade records in the public data set. They are fully consistent with the CFTC-SEC description of the large seller's trading: (i) they are all sells, (ii) they span a period of 20 minutes from 14:32 to 14:52, and (iii) they add up to 75,000 E-mini futures contracts.

Appendix B describes our data integrity analysis. We find that (i) there is no evidence of persistent, long-lived delays in quote reporting by exchanges; (ii) incidences of the ask price "touching" or "crossing" the bid are few even in the half hour of the Flash Crash; and (iii) more than 90% of the trades were reported with, at most, a 100-millisecond delay (even though exchanges are allowed to report their trades with a delay of up to 90 seconds).

3. Timeline of Main Events

Figure 1 summarizes the salient features of the Flash Crash by plotting E-mini price, volume, and a timeline reviewing the main events on the day of the crash. These events include both known facts (in regular font) and the new empirical facts (in italic font) and, thus, serve as a brief summary. The figure leads to a couple of noteworthy observations. First, the price–volume graph illustrates that the crash was not trivially a result of lack of volume. In the minute of the steep price drop that triggered a five-second halt, volume spiked up.

Second, the news flow on May 6 emphasized political and economic uncertainty, in particular related to the ongoing European debt crisis. At 14:30, the fear index, VIX, had risen by 22.5% relative to market open. Trading in E-mini echoed the rising anxiety as order flow steadily grew more toxic (Easley et al. 2012, figure 2).

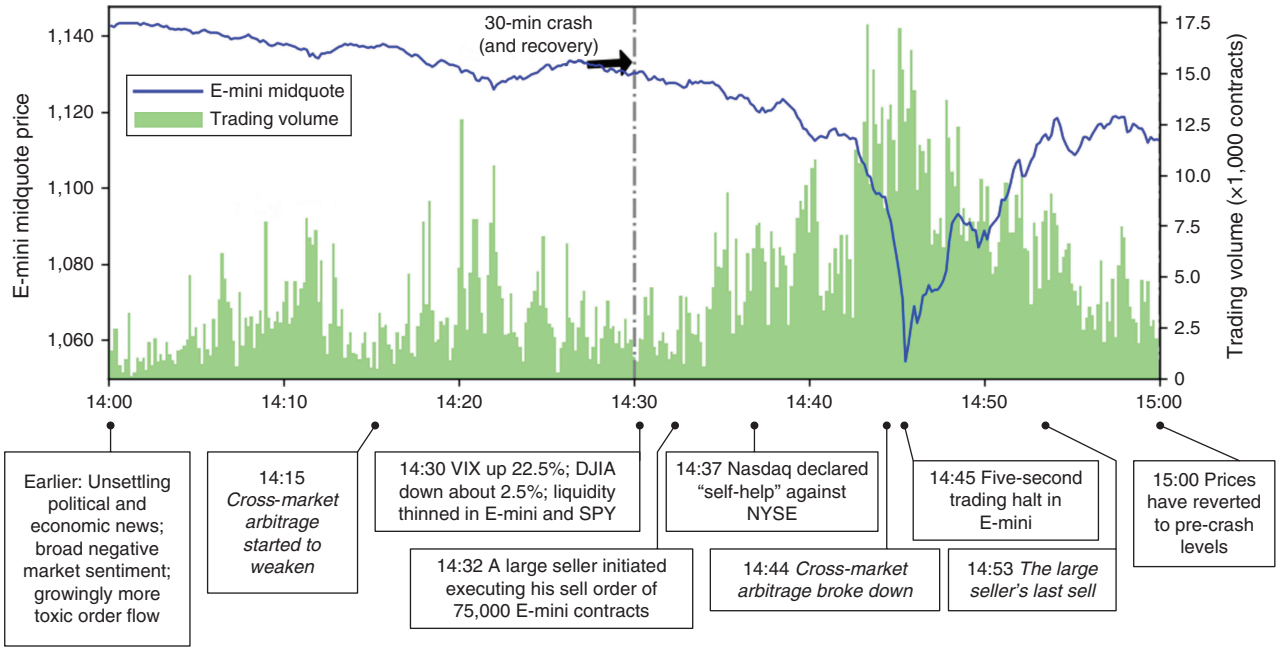
Third, the main contribution of our analysis, cross-arbitrage between the two most active markets trading S&P 500, E-mini and SPY, weakened at 14:15 and broke entirely at 14:44, one minute before the halt. In the midst of this period, at 14:37, Nasdaq declared "self-help" against NYSE Arca, a notification that a glitch had occurred and the exchange should be temporarily bypassed. Such event strained SPY trading as one important venue, NYSE Arca, was no longer available for trade: the best bid and ask prices from NYSE Arca were not "protected" as the declaring exchanges could choose not to route orders to Arca in spite of it showing superior prices (see section III.2 of CFTC and SEC 2010a). This, in turn, reduced the scope for cross-arbitrageurs to offload positions and might have led them to throw in the towel eventually.[4] Cross-arbitrage resumed at 15:00, at a time when the E-mini price had almost fully recovered.

Finally, the large seller started at 14:32. He sold 75,000 E-mini contracts at an average participation rate in volume of 9% and ended at 14:53. He therefore traded his entire order in the crash period.

4. Cross-Arbitrage Activity and Price Crash

This section explores a new channel that might have contributed to the Flash Crash: impaired cross-arbitrage. Fundamental investors trading E-mini effectively relied on cross-arbitrageurs to relay, for example,

Figure 1. (Color online) Timeline of Main Events



Notes. This figure illustrates the sequence of events on the afternoon of May 6, 2010. The graph combines different sources: media ("US Shares Plunge Amid Fears Over Debt," *Financial Times*, May 7, 2010), regulators' reports (CFTC and SEC 2010a, b), and academic studies (Kirilenko et al. 2017, Easley et al. 2012). Earlier in the day, the market experienced "unsettling political and economic news," such as "European debt crisis" and "broad negative market sentiment" (CFTC and SEC 2010a). Easley et al. (2012) show that order flow had grown steadily more toxic (rising VPIN) in the course of the day.

their sells to buyers of the S&P 500 in other markets. The most active alternative market is the exchange-traded fund (ETF) SPY. Under normal market conditions, cross-arbitrage between E-mini and SPY effectively sources liquidity supply from SPY to E-mini and vice versa. If such cross-arbitrage severely weakens, then liquidity supply becomes local. Strong liquidity demand might then command disproportionate price concessions in the form of price crashes. In this section, we show that indeed E-mini/SPY cross-arbitrage severely weakened before the Flash Crash. We then extend the analysis to the cross section of individual stocks to more firmly establish such sequencing.

Cross-Arbitrage Proxies. We develop two proxies for cross-arbitrage activity. The first proxy simply computes the size of a gross profit opportunity resulting from a crossed market. In other words, what is the return on buying one security at the market featuring the lowest ask price and then reselling it at the market with the highest bid? Formally, we compute the first proxy at time t as

$$\max\{0, \max\{B_{1,t}, \ldots, B_{n,t}\} - \min\{A_{1,t}, \ldots, A_{n,t}\}\},$$

where $A_{i,t}$ and $B_{i,t}$ are the (log-transformed) best ask and bid prices at venue i, respectively. This return is strictly positive in a crossed market (where the highest bid is above the lowest ask) and is zero otherwise.[5]

The second proxy is inspired by the broad definition of cross-arbitrage activity, including a market maker's quoting on different venues (van Kervel 2015). If the market maker's bid is hit in one market, the market maker might offload the market maker's position in other markets, potentially by taking the highest bid. The display of quotes in multiple venues, thus, tends to align the best bid and ask prices across venues. It follows that when the best bid in various markets shows substantial dispersion, such quoting behavior is likely to have weakened. This inspired us to construct a second cross-arbitrage proxy based on, say, n venues:

$$\frac{1}{2n}\left(\sum_{i=1}^{n}(A_{i,t} - \bar{A}_t)^2 + \sum_{i=1}^{n}(B_{i,t} - \bar{B}_t)^2\right),$$

where $\bar{A}_t$ and $\bar{B}_t$ are the cross-sectional means of all ask and bid prices, respectively, at time t. Effectively, the proxy computes the standard deviation of quote differentials across venues. Compared with the first measure, this quote dispersion proxy is more sensitive in the sense that it purports to measure cross-arbitrage activity even if markets did not (yet) cross.

Both our proxies can be constructed from standard data sources (e.g., the Trade and Quote data set, TAQ) because they require as input only the best bid and ask quote from each venue. We used Nanex data for all our analysis. We however did redo some key analysis, for example, Table 1, with TAQ data and discovered that the results are largely similar. We decided to use

Table 1. Rank-on-Rank Regression

(a) Crash time based on the first time when the price is below p% of the 14:30:00 price

Crash start at (# detected)	$p = 4.0\%$ (50 stocks)			$p = 5.0\%$ (50 stocks)			$p = 6.0\%$ (50 stocks)			$p = 7.0\%$ (50 stocks)		
Slackness	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Quote dispersion	0.263*	0.329**	0.232*	0.214	0.319**	0.300**	0.227	0.358**	0.324**	0.243	0.316**	0.327**
	(1.89)	(2.41)	(1.70)	(1.51)	(2.33)	(2.21)	(1.56)	(2.59)	(2.31)	(1.68)	(2.26)	(2.34)
	[0.82]	[0.72]	[0.60]	[0.88]	[0.84]	[0.74]	[0.92]	[0.86]	[0.78]	[0.92]	[0.88]	[0.84]
Volatility	−0.056	0.018	0.115	−0.034	0.057	0.123	−0.014	0.074	0.073	−0.088	−0.058	−0.076
	(−0.40)	(0.13)	(0.84)	(−0.24)	(0.41)	(0.90)	(−0.10)	(0.53)	(0.52)	(−0.60)	(−0.41)	(−0.54)
	[0.96]	[0.96]	[0.96]	[1.00]	[1.00]	[0.98]	[1.00]	[1.00]	[0.98]	[1.00]	[1.00]	[1.00]
Order flow	0.260*	0.231	0.255*	0.276*	0.217	0.261*	0.178	0.093	0.155	0.192	0.132	0.180
	(1.78)	(1.62)	(1.87)	(1.86)	(1.52)	(1.92)	(1.17)	(0.65)	(1.10)	(1.26)	(0.90)	(1.28)
	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]	[0.98]
VPIN	0.095	−0.121	−0.341*	0.047	−0.097	−0.216	0.052	0.008	−0.090	0.035	0.052	−0.081
	(0.65)	(−0.79)	(−1.86)	(0.32)	(−0.63)	(−1.19)	(0.34)	(0.05)	(−0.48)	(0.23)	(0.33)	(−0.43)
	[0.44]	[0.26]	[0.12]	[0.48]	[0.30]	[0.14]	[0.50]	[0.32]	[0.16]	[0.52]	[0.32]	[0.16]
R-squared	0.184	0.176	0.186	0.151	0.168	0.197	0.108	0.154	0.140	0.113	0.135	0.146
Total obs.	50	50	50	50	50	50	50	50	50	50	50	50

(b) Crash time based on n consecutive price drops, each more than 2 (normal-time) standard deviations

Crash start at (# detected)	$n = 2$ consecutive drops (50 stocks)			$n = 3$ consecutive drops (48 stocks)			$n = 4$ consecutive drops (40 stocks)			$n = 5$ consecutive drops (29 stocks)		
Slackness	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Quote dispersion	0.190	0.158	0.034	0.248*	0.242*	0.218*	0.154	0.286**	0.294**	0.193	0.234*	0.271**
	(1.35)	(1.18)	(0.26)	(2.01)	(1.95)	(1.76)	(1.22)	(2.29)	(2.19)	(1.53)	(1.85)	(2.03)
	[0.38]	[0.24]	[0.20]	[0.69]	[0.58]	[0.54]	[0.88]	[0.75]	[0.70]	[0.97]	[0.90]	[0.86]
Volatility	0.129	0.338**	0.446***	0.296**	0.332**	0.390***	0.085	0.158	0.199	−0.158	0.058	0.115
	(0.90)	(2.49)	(3.40)	(2.37)	(2.64)	(3.12)	(0.67)	(1.25)	(1.47)	(−1.24)	(0.45)	(0.85)
	[0.98]	[0.98]	[0.94]	[1.00]	[1.00]	[0.98]	[1.00]	[1.00]	[1.00]	[1.00]	[1.00]	[1.00]
Order flow	0.149	0.165	0.187	0.220*	0.258*	0.286**	0.115	0.088	0.144	0.163	0.065	0.143
	(1.01)	(1.18)	(1.44)	(1.69)	(1.99)	(2.31)	(0.87)	(0.68)	(1.07)	(1.23)	(0.49)	(1.07)
	[0.92]	[0.92]	[0.90]	[0.98]	[0.98]	[0.98]	[1.00]	[1.00]	[1.00]	[1.00]	[1.00]	[1.00]
VPIN	0.174	0.048	−0.060	0.174	0.115	0.033	0.454***	0.415***	0.199	0.369***	0.388***	0.163
	(1.17)	(0.32)	(−0.34)	(1.34)	(0.83)	(0.20)	(3.44)	(2.97)	(1.10)	(2.78)	(2.72)	(0.91)
	[0.14]	[0.06]	[0.02]	[0.42]	[0.19]	[0.04]	[0.55]	[0.30]	[0.17]	[0.66]	[0.45]	[0.14]
R-squared	0.166	0.202	0.259	0.355	0.315	0.330	0.331	0.308	0.207	0.276	0.234	0.158
Total obs.	50	50	50	50	50	50	50	50	50	50	50	50

Notes. This table presents regression results of a stock's crash-time rank on four explanatory variables. Crash time is defined in two ways. In Panel (a), it is the first time that a stock's price is p% below its 14:30:00 price, where the threshold p varies from four to seven. In Panel (b), crash time is defined as the first time the NBBO midquote of the stock drops by at least two (normal-time) standard deviations in n consecutive 100-millisecond intervals. The explanatory variables are also rank variables based on a series' break point. Such points are constructed following a standard statistical process control procedure with slackness $k \in \{1, 2, 3\}$ times the (normal-time) standard deviation (see Appendix A). The explanatory variables are quote dispersion, volatility, order flow imbalance, and VPIN (Easley et al. 2012). The t-statistics are in parentheses. The proportion of stocks for which the break preceded the price crash is in square brackets. The superscripts *, **, and *** indicate significance at 10%, 5%, and 1%, respectively.

Nanex data for the paper mostly because we believe time stamps to be more reliable for May 6.

Structural Break. We prefer to only be alerted about weakened arbitrage if the proxy time series shows a "structural break." In other words, one is likely to observe regular peaks in the time series generated by the cross-arbitrage proxy simply because a proxy is subject to noise. If, however, the time series exhibits many right-tail observations in a row, then this suggests the time series mean has shifted upward. This would indicate that weakened arbitrage has become structural and indicative of truly weakened arbitrage. A standard statistical procedure that formalizes this intuition is known as statistical process control (SPC). We will use SPC to identify weakened arbitrage time points. Appendix A describes the procedure in full detail.

The reason for this focus on *structural breaks* is that it identifies "disruptive" market events. We conjecture that (algorithmic) traders only radically change their behavior (e.g., discontinue cross-arbitrage) after

such attention-grabbing events. They suddenly realize that they are in "uncharted territory" or extremely unusual market conditions. We consider SPC appropriate for detecting such conditions as it is engineered to do exactly that: detect the point at which the present becomes disconnected from the past, statistically speaking. The time stamps of these break points then conveniently establish sequences of disruptive events, motivating the rank-on-rank regressions studied in Section 4.2.

Fragility of Cross-Arbitrage and Breakdown. We interpret SPC-detected break points as cross-arbitrage breakdowns. Such breakdowns are unique to May 6. Applying the same procedure to E-mini, SPY, and a large cross section of individual stocks, we only find them on May 6, not on either May 3, 4, or 5 (i.e., the control sample used in Kirilenko et al. 2017). A natural follow-up question is the following: Why did cross-arbitrage break down on May 6?

The drivers could be many, and identifying the exact cause is difficult. While we are unable to pinpoint the exact cause, in the remainder of this section we discuss contributing factors.

To begin with, CFTC and SEC (2010a, p. 1) note that "May 6 started as an unusually turbulent day" as "trading in the U.S. opened to unsettling political and economic news from overseas concerning the European debt crisis." In the afternoon, the market had seen "broadly negative market sentiment." E-mini trading itself grew more toxic as time progressed that day (Easley et al. 2012). We speculate that these conditions might have made cross-arbitrageurs more sensitive to unusual market movements, to the point of deciding to temporarily switch off their engines.

Two events in particular could have added to market participants' nerves that day:

- The market experienced significant delays in quote and trade streams (CFTC and SEC 2010a, III.3). Jones (2013, p. 36) notes that while the delay mainly occurred in NYSE Arca, "it may have even caused some liquidity suppliers at other markets to step back." The delays further rang bells at other exchanges as NASDAQ and BATS sequentially declared self-help against NYSE Arca. Such declaration means that the declaring exchange no longer needs to honor the (possibly stale) quotes displayed at the strained exchange (i.e., the trade-through rule is suspended). For more information, see section III.2 of CFTC and SEC (2010a).
- E-mini trading reportedly experienced substantial spoofing activity that day. A single trader, Navinder Sarao, was charged with purposefully misleading the market by entering quotes without an intention to trade on them.[6]

4.1. Trading S&P 500 Through E-mini and SPY

Figure 2 illustrates how cross-arbitrage weakened on the day of the Flash Crash. Cross-arbitrage in this case pertains to trading the S&P 500 index through E-mini and SPY. Panel (a) shows that gross arbitrage opportunities existed and were large and persistent in particular in the 15 minutes after the price had crashed. The E-mini ask was multiple percentage points below the SPY bid. The other salient pattern is that SPY recovered more quickly than E-mini. The lead–lag relationship between E-mini and SPY during the recovery period echoes the finding in Ben-David et al. (2012), who show that the crash in E-mini slightly led that in SPY.

Panel (b) plots the first cross-arbitrage proxy, the size of gross arbitrage opportunities, both for the 30-minute crash period (first graph) and the whole trading day (second graph). It echoes the finding in Panel (a) that these opportunities were largest after the trading halt. And, more importantly, it shows that these opportunities existed *before* the one-minute steep price drop just ahead of the trading halt (14:44–14:45). They became persistent starting 15 minutes before this steep drop as SPC out-of-control alerts kept coming uninterruptedly as of 14:30. This is illustrated by the gray area in the graph. Note that SPC did not raise much alarm earlier in the trading day as gray bars appeared only sporadically.

Panel (c) repeats Panel (b) for the second cross-arbitrage proxy: quote dispersion. It also finds that cross-arbitrage weakened persistently before the crash. The graph illustrates that this proxy is more sensitive as gray areas started to appear around 14:00, 15 minutes earlier than the first proxy (c.f. the second graphs in Panel (b) and (c)). The reason is that quote dispersion can become elevated without immediately causing markets to become crossed. Consistent with Panel (b), the magnitude of quote dispersion (blue line) was largest after the price crash at around 14:45.

In sum, the structural break analysis suggests a sequencing of events in terms of cross-arbitrage strength. Adding the price co-integration results reported in the online appendix, one gets the following more complete sequencing (time indication in minutes relative to the E-mini trading halt is in brackets):

- (−30) Weakened cross-arbitrage suggested by elevated cross-market quote dispersion.
- (−15) Weakened cross-arbitrage suggested by persistent (gross) arbitrage opportunity.
- (−1) Broken cross-arbitrage suggested by E-mini–SPY price co-integration break.
- (−1) Start of the one-minute steep price drop.
- (0) E-mini trading halt triggered.

4.2. Trading a Stock Through Multiple Venues

Both proxies persistently alerting impaired arbitrage ahead of the crash supports the economic channel we

conjecture but by no means identifies it. We turn to the cross section of the most crashed S&P 500 stocks to generate more evidence on the sequencing of impaired cross-arbitrage and a price crash. These stocks all traded in multiple venues and, thus, also rely on cross-arbitrage. If impaired arbitrage serves as a harbinger for a crash, then the sequencing of impaired cross-venue arbitrage should line up with the sequencing of crashes among the stocks.

The sample of stocks used for this cross-sectional analysis consists of the 50 most crashed stocks in the S&P 500 index. This set includes the six stocks studied thoroughly in CFTC and SEC (2010a).

Figure 3 illustrates trading in these stocks during the Flash Crash by plotting prices by venue as well as the quote-dispersion proxy for cross-arbitrage activity. These graphs are only plotted for the six stocks featured in the CFTC-SEC report; the other stocks show a

Figure 2. (Color online) Price Differential Between E-mini and SPY

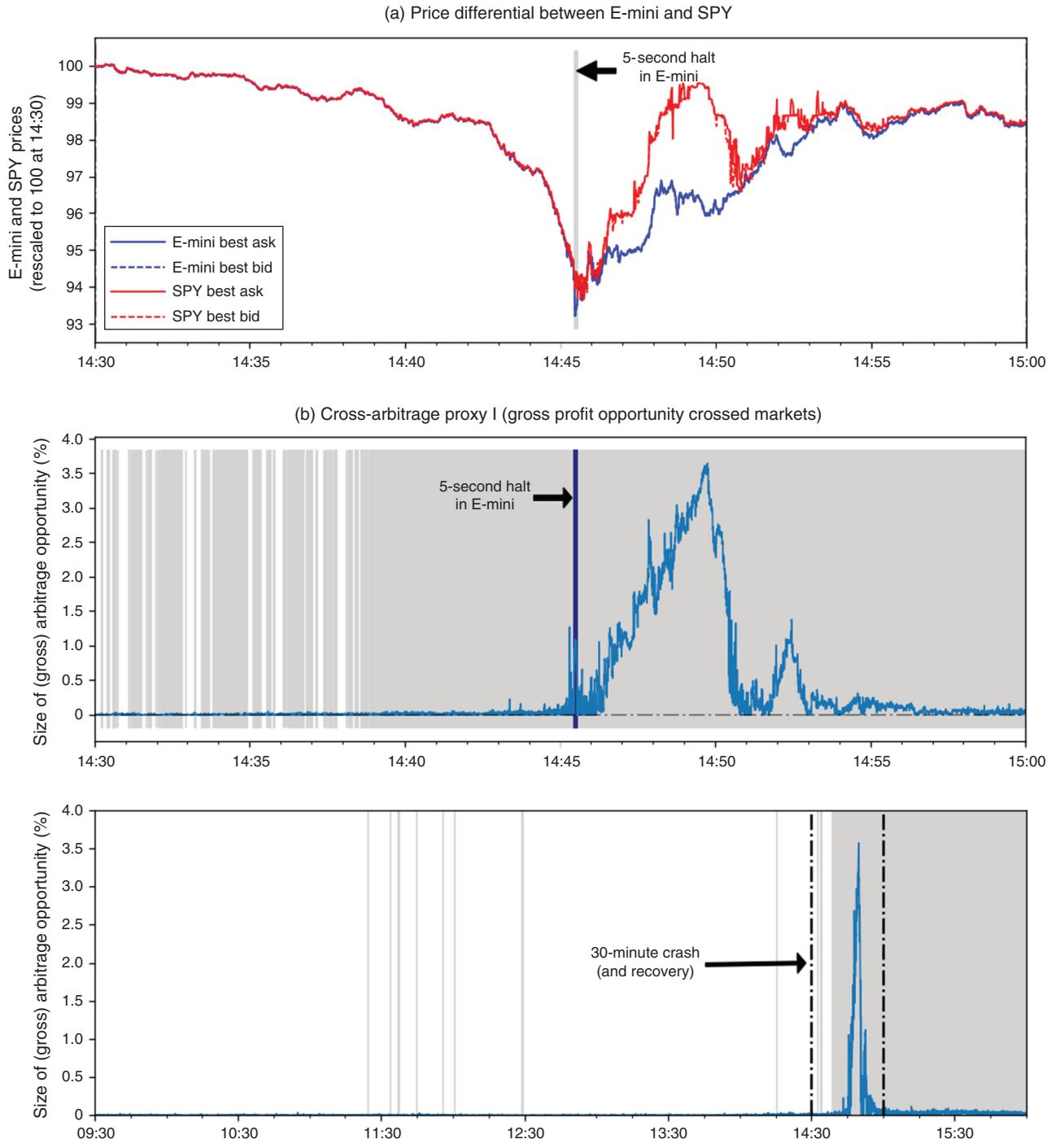
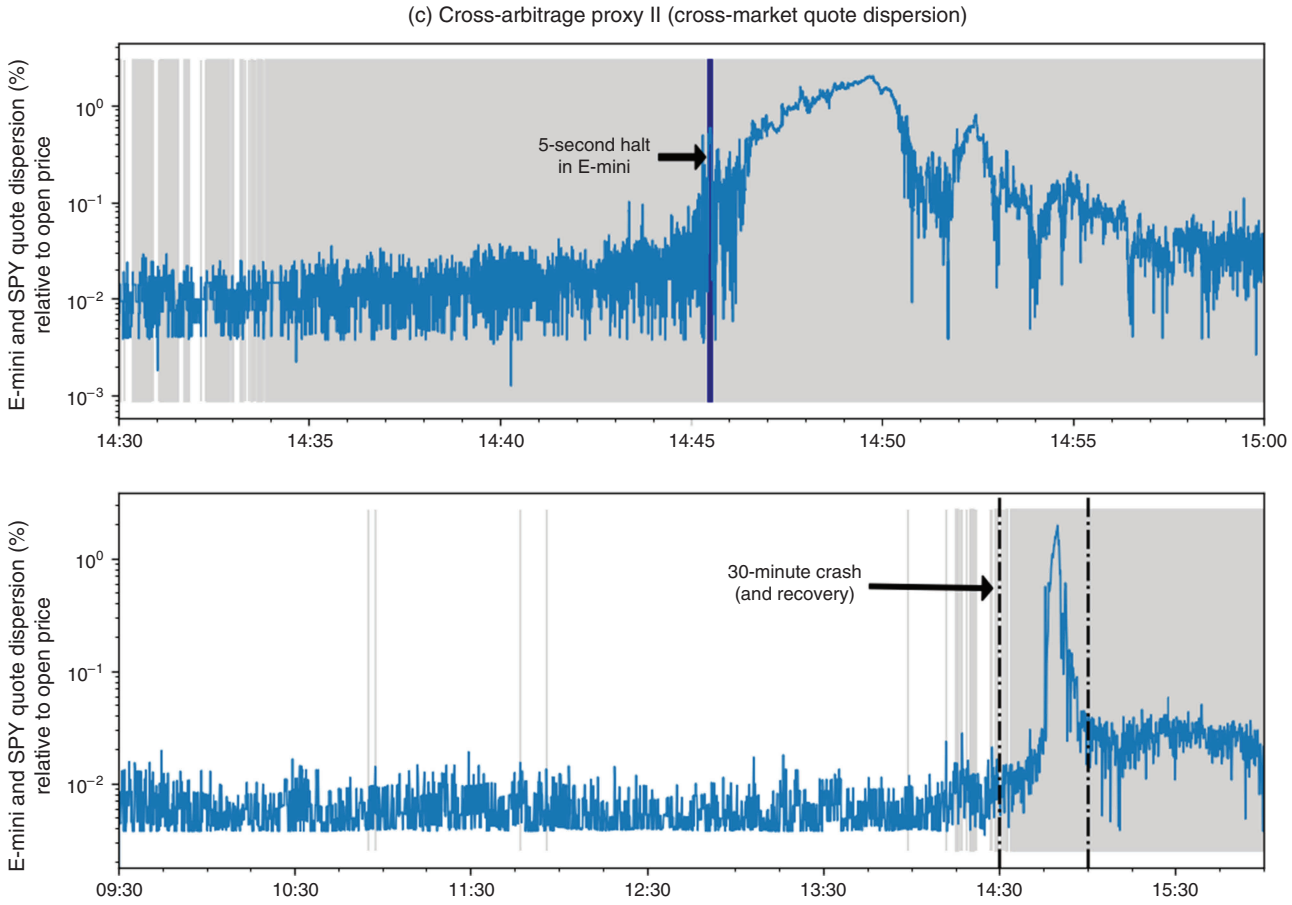


Figure 2. (Color online) (Continued)



Notes. This figure illustrates the price differential between E-mini and SPY. Panel (a) plots the bid and ask quotes for the two markets. Panel (b) plots the first proxy of cross-arbitrage: the size of a gross-profit opportunity resulting from crossed markets (i.e., the ask in one market being strictly above the bid in the other market). Panel (c) plots the second proxy of cross-arbitrage: quote dispersion across markets. To compute this proxy, one first collects the best bids in all markets in a set, does the same for the best asks, demeans both sets, joins them, and then computes the standard deviation and expresses it relative to the open price. The gray bars indicate when the time series turns "out of control" based on standard statistical process control. Both Panel (b) and (c) first focus on the 30-minute Flash Crash period from 14:30 to 15:00 and then zoom out to provide an overview of the entire trading day.

similar pattern. The price graph plots midquotes (i.e., the average of the best bid and ask) for the three most active venues (BATS, PACF, and NQEX). They show that these quotes differ most at the time of the price crash. This suggests higher quote dispersion, an observation that is confirmed by the cross-arbitrage proxy graphs. Note that only the quote-dispersion proxy is useful here as plain vanilla arbitrage opportunities are ruled out by the no trade-through rule that applies to equity trading. If a venue receives a price quote that would create a crossed market, then it is not allowed to post it. It has to be rerouted to the venue that could effectuate a trade (because the quote is a marketable one). The graphs further show that weakened-arbitrage alerts (i.e., gray bars) consistently precede price crashes. This finding is generally true for all 50 stocks in the sample.

To formally establish a statistical link between cross-arbitrage breakdowns and price crashes, we turn to a rank-on-rank regression. The idea is to examine whether for the 50 most crashed stocks the sequencing of cross-arbitrage breakdowns (statistically) explains the sequencing of price crashes. The regressions control for other factors that might have contributed to individual stock crashes. Specifically, we again use SPC to compute break point times for realized return volatility, order flow imbalance, and volume-synchronized probability of informed trading (VPIN, Easley et al. 2012).[7] These break points are then used to sequence the control variables. In the regressions, therefore, both left- and right-hand variables become rank variables, hence the term rank-on-rank regressions. The economic motivation for sequencing these series based on break points is that only a sudden highly unusual pattern will make cross-arbitrageurs worry to the point of potentially pulling out temporarily. (Who wants to operate in market conditions that one is not familiar with?)

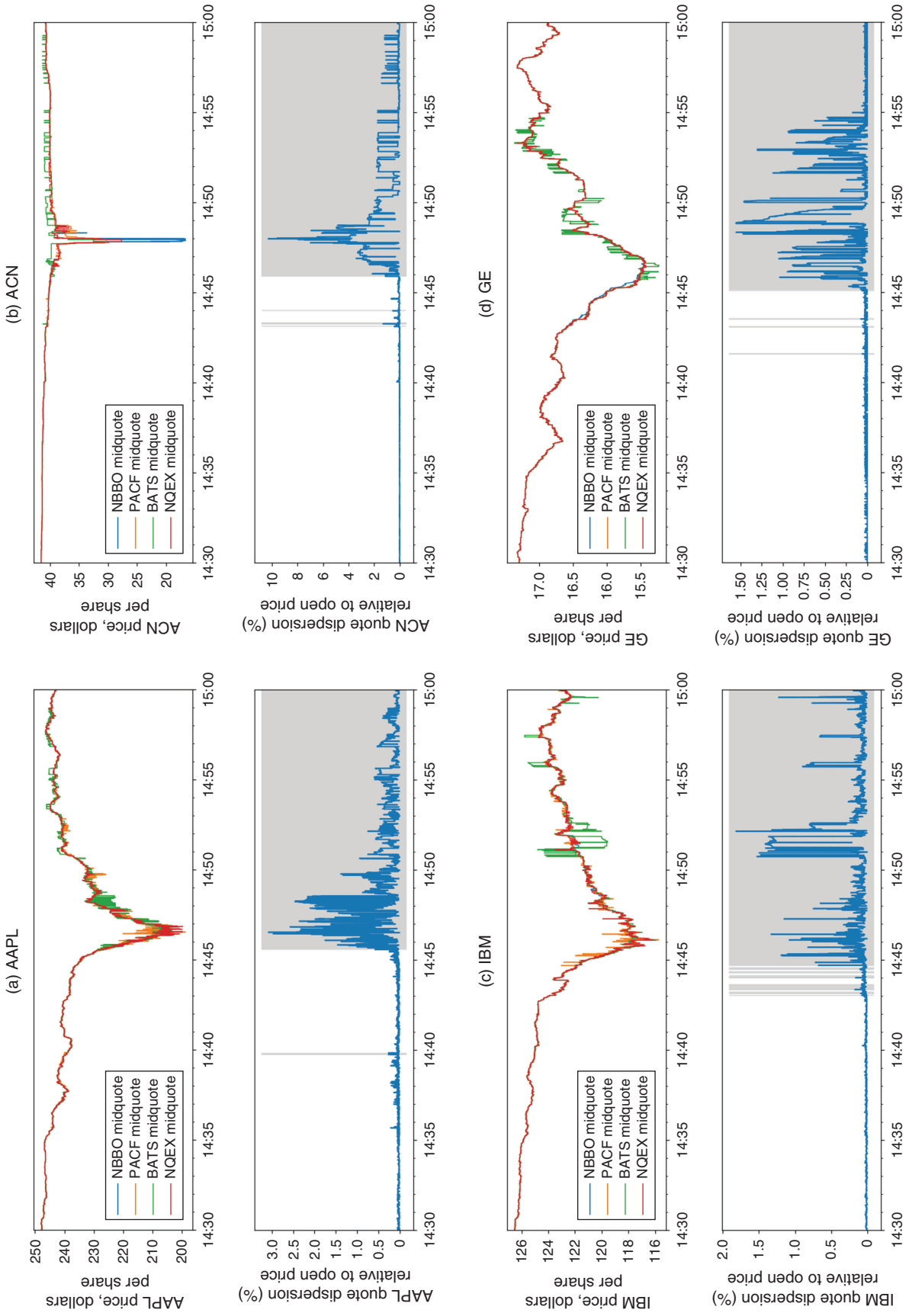
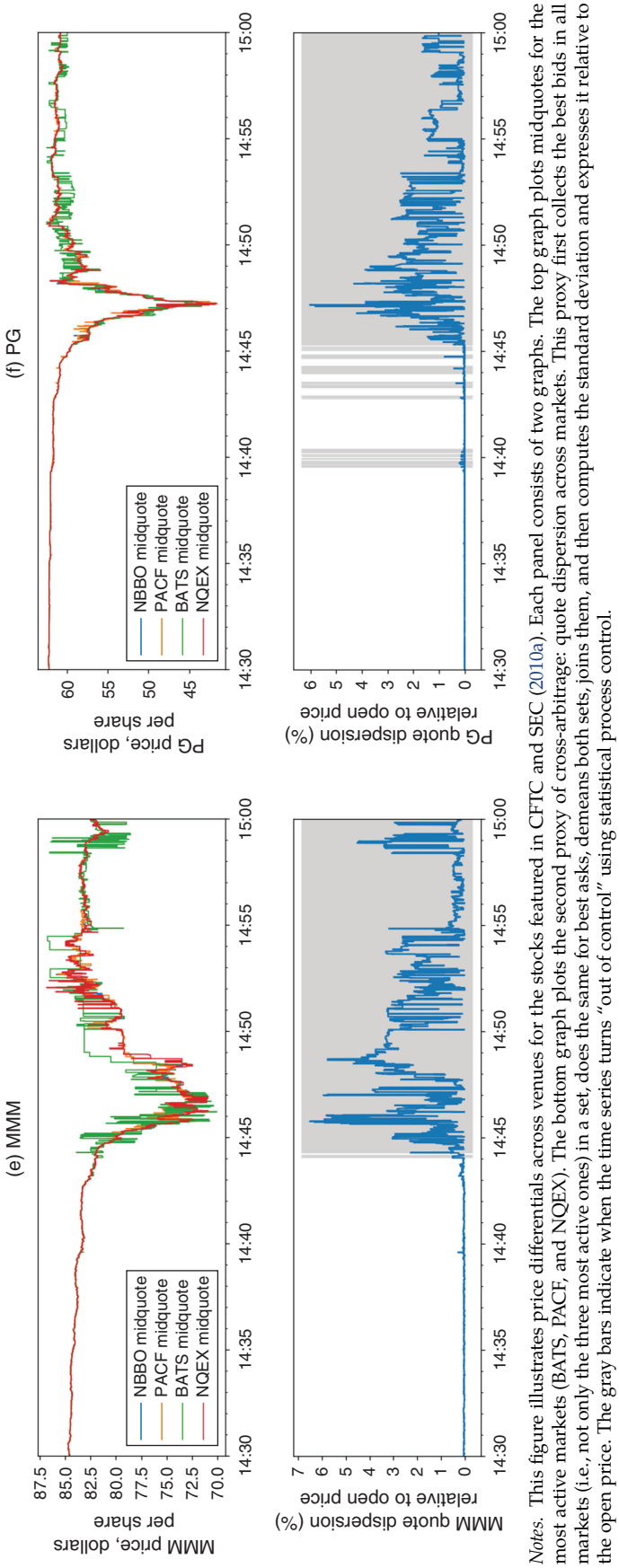


Figure 3. (Color online) Cross-Venue Price Differential for Individual Stocks

Figure 3. (Color online) (Continued)



Notes. This figure illustrates price differentials across venues for the stocks featured in CFTC and SEC (2010a). Each panel consists of two graphs. The top graph plots midquotes for the most active markets (BATS, PACF, and NQEX). The bottom graph plots the second proxy of cross-arbitrage: quote dispersion across markets. This proxy first collects the best bids in all markets (i.e., not only the three most active ones) in a set, does the same for best asks, demeans both sets, joins them, and then computes the standard deviation and expresses it relative to the open price. The gray bars indicate when the time series turns "out of control" using statistical process control.

Table 1 presents the results of the rank-on-rank regressions with the price crash defined in two ways: (i) the first time the price drops by at least several percentage points below its 14:30 level or (ii) the first time the price experienced several consecutive large price declines. The latter measure is inspired by Nanex, a firm whose Flash Crash reports received widespread attention.[8] The empirical results show that, indeed, the timing of cross-arbitrage breaks (proxied by quote dispersion) line up with the timing of stock price crashes. The coefficients are all positive, and many are statistically significant. Such statistical significance is particularly striking given that the sample consists of only 50 stocks. The table further reveals that in the vast majority of cases the cross-arbitrage break happened before the price crash. Finally, note that several control variables also show some significant positive coefficients, but none of them so consistently as quote dispersion. We therefore find that cross-arbitrage breaks have explanatory power for price crashes over and above standard trade variables.

One important worry is that these results might be driven by a third factor. In the half hour of the Flash Crash, 14:30 until 15:00, the data feed from NYSE Arca became unreliable. In response, NASDAQ and BATS declared self-help against Arca, which allowed them to neglect Arca quotes (CFTC and SEC 2010a, section III.2). In other words, it allowed them to "trade through" Arca. Note that in our analysis Arca quotes are included and the self-help events might, therefore, have triggered both extreme quote dispersions and price crashes, thus causing the third-factor issue. To address it, we have redone the rank-on-rank regressions after first removing Arca from the sample. The results are largely unaffected and are reported in Appendix C.

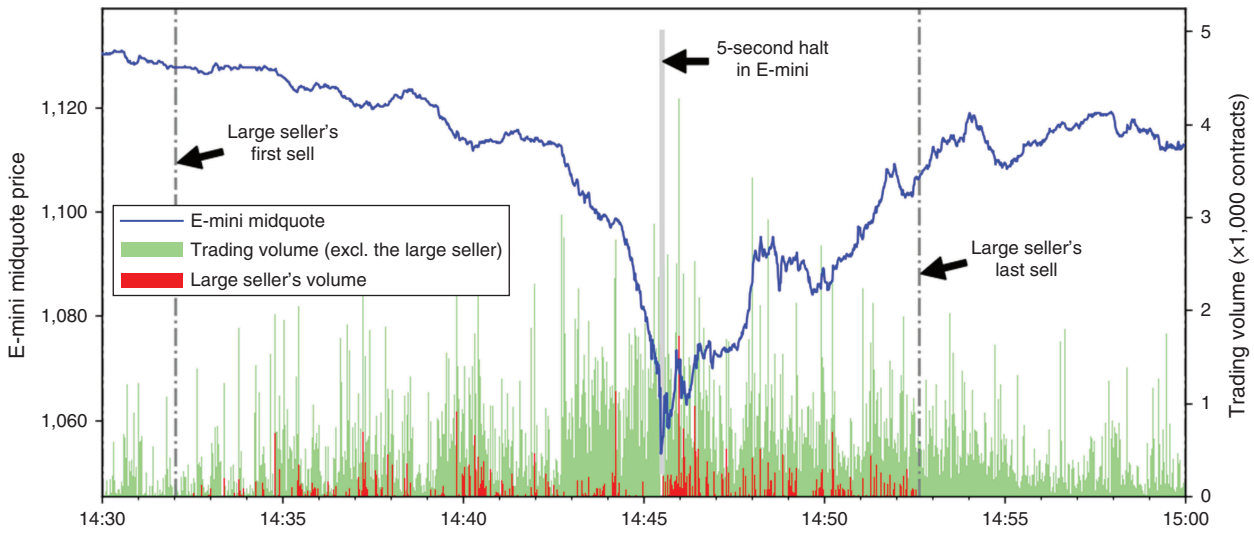
We caution that the evidence presented in Table 1 should be interpreted carefully. We do not claim to have found a causal relationship between cross-arbitrage breakdowns and subsequent price crashes. It is possible that some unknown third factor drove both events. Neither do we claim that the cross-arbitrage breakdown was a primary factor leading to the crash. In fact, the rank-on-rank regressions suggest that other factors—volatility, order flow imbalance, order flow toxicity (VPIN)—all played a role. The analysis only highlights that cross-arbitrage, as proxied by quote dispersion, has *additional* explanatory power for price crashes. We argue that this new angle for looking at the Flash Crash contributes to a better overall understanding of the event.

5. Large Seller's Trading

This section analyzes the large seller's trading in the Flash Crash half hour. It first explores how he implemented selling 75,000 E-mini contracts. It then estimates how much he paid for demanding this liquidity.

Figure 4 illustrates how the large seller traded in the crash period. It does so by plotting E-mini price and volume whereby the dark (red online) parts of the lighter (green online) volume bars denote large-seller volume. The graph shows that the large seller decelerated trading in the minutes before the crash. According to CFTC and SEC (2010a), the large seller set a 9% volume target over the previous minute. Yet, in the minutes leading to the crash, he clearly stayed far below the 9% threshold. This evidence mitigates the concern that

Figure 4. (Color online) The Large Seller's Trading During the Flash Crash



Notes. This figure highlights the large seller's trading during the Flash Crash period. He only sold in the E-mini market. Trading volume bars are stacked: the dark part (red online) is volume that involved the large seller, and the light part (green online) is volume that did not involve him.

the one-minute steep price drop was simply an immediate response to extreme selling by the large seller (in that minute). The other salient feature of the graph is that the large seller sold most of his position in the minutes immediately after the halt when the price started to recover. The price, however, was still far from full recovery, and this selling arguably contributes most of the high price he paid for liquidity (see the calibration in Section 5.2).

The figure is further consistent with the observation in Section 4.1 that SPY recovered more quickly than E-mini. Notice how the large seller sold heavily throughout the entire recovery period, thereby keeping the selling pressure high on E-mini. The lack of cross-arbitrage meant that this pressure was not relayed to SPY, which, therefore, could recover more quickly.

5.1. The Price the Large Seller Paid for Liquidity

The large seller paid for demanding liquidity, not by taking liquidity beyond the best bid (none of his trades walked down the order book), but by driving the price temporarily down in the 20 minutes he took to execute the order. The lion's share of his total order was executed *just after* the E-mini trading halt, a time when the price had just bottomed out (see dark (red online) bars in Figure 4). We first compute the price the large seller paid for demanding such liquidity in adverse conditions (i.e., severely strained cross-arbitrage). We then calibrated such liquidity demand to a standard liquidity supply model. Such calibration allows one to judge whether the price he paid was "excessive."

Table 2 explores what price the large seller paid for immediacy, a calculation that requires us to take a stance on a reasonable counterfactual price. That is, what would the price have been had the large seller not traded? We consider several options: the price at market open, at the time of the first sell, at the market close, and also the volume-weighted average price (VWAP). The cost estimates range from $98.6 to $229.8 million or, in relative terms, from 234.1 to 529.5 basis points (i.e., relative to the value of 75,0000 E-mini contracts). The table further shows that such cost level was extremely high for the large seller. His first-quarter operating income in 2010 was $58.6 million. His liquidity cost experienced in only 30 minutes therefore wiped out at least two times his quarterly operating income.

5.2. Calibrating the Price Paid for Liquidity

We calibrate the price the large seller paid for liquidity to study what market conditions could rationalize the Flash Crash as a price paid for immediacy. In microstructure, there are two canonical motivations for an intermediary to command price impact on incoming orders: information asymmetry or costly inventory. In an information model, the intermediary understands that orders are, with some probability, information motivated. The intermediary therefore protects himself by trading at prices that, on average, at least recover the expected loss from trading with these orders. In other words, if the intermediary trades, the intermediary will sell at a price strictly above fundamental (ask), and the intermediary will buy at a price (bid) strictly below fundamental. The more informed the orders are, the higher price impact the intermediary will charge (e.g., Kyle 1985). An information-based explanation for the Flash Crash would have intermediaries interpret the high selling pressure as signaling extremely informed orders. It could explain the drop, but it is inconsistent with a subsequent price rebound (i.e., price changes reveal information and are, therefore, permanent). A price rebound is naturally generated by an inventory model, which is why we turn to this type of model for the calibration.

In inventory models, a risk-averse intermediary temporarily holds inventory when intermediating between buyers and sellers. The price risk the intermediary runs on such inventory requires the intermediary to charge those who demand immediacy a (temporary)

Table 2. How Much Did the Large Seller Pay for Liquidity?

	Price large seller paid for liquidity		Price paid relative to his total assets[a]	Price paid relative to his operating income[a]
Timing of counterfactual price	($ million)	(bps)	(%)	(%)
9:30 A.M. (market open)	229.8	529.4	24.8	381.7
2:32 P.M. (time of large seller's first sell)	117.3	277.4	13.0	200.0
4:15 P.M. (market close)	8.6	234.1	11.0	168.1
VWAP price for the day	138.78	326.5	15.3	235.4

Notes. This table presents computations on the price the large seller paid for the liquidity he demanded when selling 75,000 E-mini contracts. This price is computed as his average transaction price relative to several candidate values for the counterfactual price—the price that would have been obtained had he not traded.

[a]Taken from the large seller's 2010Q1 quarterly report.

"price pressure." The inventory model used here is Grossman and Miller (1988) with constant *relative* risk aversion instead of the model's constant *absolute* risk aversion—CRRA instead of CARA. One important reason for using CARA in theoretical papers is that models can be solved analytically. CRRA, however, is the de facto choice for calibration (see, for example, Mehra and Prescott 1985).

The calibration model is specified as follows: A seller with an inelastic supply curve of 75,000 E-mini contracts trades with a representative CRRA intermediary who operates on a zero-profit basis (to capture a competitive market). After an interval of length T, the intermediary trades out of the position with buyers who have an infinitely elastic demand curve. Hence, the intermediary does not pay any price pressure when selling in the second stage—the intermediary is able to sell at fundamental value. The price that clears the market in the first stage relative to fundamental value, therefore, captures the price pressure that the seller has to pay. The deep parameters that drive the size of this price pressure are the intermediary's relative risk aversion γ, the intermediary's initial wealth w_0, the length of the holding period T, and the price risk per unit of time σ^2. For simplicity, the intermediary's discount rate is assumed to be one. We consider this to be a relatively innocent assumption as the intermediary's inventory holding period is short (hours or, in the worst case, a couple of days).

Parameter Choice. Various values were picked for all of the model's parameters except for position size (which equals 75,000 E-mini contracts). The risk parameter σ^2 takes three values: one based on realized volatility from 9:30 until 14:30, one based on the value used by Kyle and Obizhaeva (2016), and a forward-looking value based on the highest VIX level observed the Flash Crash half hour (i.e., from 14:30 until 15:00). The latter value assumes intermediaries perfectly foresaw the extreme volatility that was coming.

Three levels are chosen for intermediary wealth w_0: \$4.1 billion, \$78 billion, and \$200 billion. \$4.1 billion is the minimum wealth level in the intermediation sector because CRRA does not admit negative wealth. \$78 billion was approximately the market capitalization of Goldman Sachs on May 6, 2010. \$200 billion is a ballpark figure based on Kirilenko et al. (2017). They find that there were 16 high-frequency traders and more than 100 market intermediaries operating in E-mini at the time of the crash.

Two values are considered for the length of the holding period. A five-hour period assumes normal market conditions. According to CFTC and SEC (2010a, p. 14), the large seller had previously executed a similar size sell program: "On that occasion it took more than 5 hours for this large trader to execute the first 75,000 contracts of a large sell program." The other holding period considered is three days.

Finally, we picked several values for the relative risk-aversion coefficients. Hendershott and Menkveld

Table 3. Calibrating the Price the Large Seller Paid for Liquidity

		Initial wealth and holding period					
		\$4.1 billion[a]		\$78 billion[b]		\$200 billion	
Volatility	Risk aversion	5 hours	3 days	5 hours	3 days	5 hours	3 days
As realized from 9:30 A.M. to 2:30 P.M. on May 6	3.96[c]	6.3	106.5	0.4	5.9	0.2	2.3
	10[d]	15.5	263.1	0.8	14.8	0.3	5.8
	12[e]	18.6	313.2	1.0	17.8	0.4	7.0
	50[f]	76.7	1,136.7	4.2	73.4	1.6	28.9
As implied by highest VIX from 2:30 P.M. to 3:00 P.M. on May 6	3.96[c]	9.9	38.3	0.5	2.1	0.2	0.8
	10[d]	24.8	95.9	1.4	5.3	0.5	2.1
	12[e]	29.8	114.7	1.6	6.3	0.6	2.5
	50[f]	122.4	453.1	6.8	26.3	2.6	10.3
As used in Kyle and Obizhaeva (2016)	3.96[c]	6.1	23.7	0.3	1.3	0.1	0.5
	10[d]	15.3	59.5	0.8	3.3	0.3	1.2
	12[e]	18.1	71.2	1.0	3.9	0.4	1.5
	50[f]	76.0	287.1	4.2	16.2	1.6	6.3

Notes. This table calibrates the price the large seller paid for liquidity (see Table 2) based on Grossman and Miller (1988). It does so by documenting what "price pressure" a zero-profit intermediary with constant relative risk aversion would have charged. Four market environment parameters vary: (i) price risk, (ii) the intermediary's relative risk aversion, (iii) the intermediary's wealth, and (iv) the inventory holding period. Only the highlighted price pressures are at least as large as what the large seller paid.

[a]Just enough wealth for the intermediary to take the full position.
[b]Goldman Sachs market capitalization on May 6, 2010.
[c]Estimate from Hendershott and Menkveld (2014) based on NYSE specialist data.
[d]Mehra and Prescott (1985) use relative risk-aversion coefficients below 10.
[e]Estimate from Barsky et al. (1997) based on experiments.
[f]See chapter 21 of Cochrane (2005).

(2014) estimate that NYSE specialists for large stocks had, on average, a relative risk-aversion of 3.96. Mehra and Prescott (1985) use relative risk-aversion coefficients below 10. Barsky et al. (1997) provide experimental evidence that an average person has an average relative risk-aversion coefficient of about 12. Using a classical asset pricing model, (Cochrane 2005, Ch. 21) finds that a relative risk-aversion of more than 50 is necessary to explain stock returns, the risk-free rate, and consumption growth in 20th century U.S. markets.

Results of Calibration. Table 3 presents the calibration results. The price pressure for reasonable values of the model's parameters is an order of magnitude lower than the 234.1 basis points price pressure paid by the large seller (the latter is a lower bound; see Table 2). The calibrated price pressure matches the realized pressure *only* if all of the three following conditions hold: (i) the intermediation sector is relatively risk-averse ($\gamma \geq 10$), (ii) the aggregate wealth in the intermediation sector is thin (just enough to buy the \$4.1 billion position[9]), and (iii) the expected holding period is long (three days). This area is indicated by the gray shading in the table. In sum, the price the large seller paid for liquidity seems excessive.

Note that, consistent with our interpretation of broken cross-arbitrage, the calibration suggests that intermediaries' risk aversion should have risen drastically (the first condition). This could be the result of broken cross-arbitrage effectively making intermediaries more cautious of taking inventory and requiring a higher risk premium in return.

6. Conclusion

This paper revisits the Flash Crash with new evidence. The large seller sold \$4.1 billion of S&P500 exposure in about 20 minutes through E-mini only. Cross-arbitrage between E-mini and SPY effectively connected him with buyers in SPY as well as in E-mini. However, in the minutes before the crash, cross-arbitrage severely weakened and eventually broke. Continued selling by the large seller could no longer find local buyers in E-mini and is consistent with the slower price recovery in E-mini as compared to SPY. A cross-sectional analysis based on individual stocks adds further evidence of weakened arbitrage preceding price crashes.

These results suggest that liquidity supply in severely fragmented markets might become vulnerable when lots of liquidity is demanded. Fragmentation might thus limit the "capacity of trading strategies" employed by institutional investors (see Landier et al. 2015, for broader discussion). One way to mitigate price crashes is to alert investors to elevated volatility, for example, by adding a quote dispersion metric to the NBBO stream. The current SEC tick-size experiment might bear on the issue as well. Wang and Ye (2017) theorize that small ticks are more conducive to mini crashes.

Acknowledgments

This paper previously circulated under the title "Anatomy of the Flash Crash." We thank Terrence Hendershott, Charles Jones, Emiliano Pagnotta, Vincent van Kervel, and Shihao Yu and participants in an SEC conference call, conference/seminar participants at Euronext Paris/ Toulouse University, Humboldt University, University College London, and the University of Gothenburg. We are very grateful to Eric Hunsader of Nanex and to Waddell and Reed for providing the data without any strings attached (other than no redistribution). There are no competing financial interests that might be perceived to influence the results and discussion of this article. Menkveld gratefully acknowledges VU University Amsterdam for a VU talent grant and NWO for a VIDI grant. Yueshen gratefully acknowledges Bengt Holmström for sponsoring his MIT visit.

Appendix A. Statistical Process Control

Consider a process x_t, which, under normal conditions, has mean $\bar{x}$ ("under control"). However, the mean of the process might deviate from $\bar{x}$ at unknown times (becoming "out of control") and we wish to detect such structural changes. A standard statistical process control (SPC) method for such purpose is "the cumulative sum control chart" technique (Montgomery 2009). Specifically, construct the upper cumulative sum of $\{x_t\}$ recursively as $s_t := \max\{0, s_{t-1} + x_t - (\bar{x} + k)\}$, where $\{s_t\}$ is the cumulative sum and k is the slack parameter. Intuitively, as x_t is progressively observed, s_t shows the cumulative upward deviation from the mean of $\bar{x}$, subject to a slackness of k units per observation. Along with the slack k, a statistician also chooses the control threshold h, such that the process becomes out of control only if $s_t > h$; that is, the cumulative upward deviation is too much to statistically believe that the structure of x_t has not changed. The smaller k or h is, the more sensitive the detection is.

We use this SPC approach in Section 4 to examine whether and when the two proxies for cross-arbitrage become upwardly out of control (i.e., the times when cross-arbitrage breaks down). While the same method symmetrically applies to detecting downward structural changes, we are primarily interested in upward changes as both our proxies are higher when there is reduced cross-arbitrage activity. We further apply SPC to sequence the various control variables in the rank-on-rank regressions featured in the same section.

Following the practical recommendations of Montgomery (2009, section 9.1.3), we vary the slack $k \in \{\sigma, 2\sigma, 3\sigma\}$ and set the control to be $h = 5\sigma$, where σ is the standard deviation of x_t when under control. We compute the under-control moments of $\bar{x}$ and σ based on 9:30–10:00 on May 6—the "training sample" (normal condition). Trading in this half-hour interval still seems normal as, for example, VPIN had not started to trend up yet (Easley et al. 2012, figure 2).

Appendix B. Data Integrity

B.1. Quote Delays

The CFTC-SEC report mentions substantial price quote delays in the Flash Crash period. Chapter III.3 of CFTC and SEC (2010a) documents that 1,665 NYSE-listed symbols were affected. The list did not include the NYSE-listed SPY. Nanex

reports two instances of SPY quote delays in BATS, each lasting for about one second. Panel (a) of Figure B.1 illustrates one of them. These delays, however, seem short-lived and idiosyncratic relative to the 10-minute-long arbitrage opportunity documented in Section 4. The CFTC–SEC report does not mention quote delays on the CME-traded E-mini contract, nor did we find any other source that reported such delays. We are, therefore, not overly worried about quote delays in our sample and treat time stamps on quotes as accurate.

B.2. Trade Report Delays

Trades might experience a time delay as exchanges are "required to report their trade activity within 90 seconds of execution time to the Consolidated Tape System (CTS)" (see Appendix Q of SEC 2001).

Before going through the data line by line, Panel (c) of Figure B.1 plots the end-of-period bid–ask spread and the volume-weighted average trade price at a one-second frequency. This leads to the following observations. First, for all exchanges, the trade price is within the bid–ask spread up to the 90-second delay that is permitted. The only exception is CINC, where, in the half hour of the Flash Crash, trade delays run up to about three minutes. Second, quote and trade activity at CBOE and CHIC seem to shut down for 5 to 10 minutes in the recovery stage of the Flash Crash. Third, the bid–ask spread is substantially wider after the halt although it does widen in some exchanges even before the halt (see, for example, CBOE).

B.3. Line by Line Check

This subsection performs a basic check on data integrity by going through the data set line by line for each exchange. For quote records, the best quotes are compared to check if the best bid price equals (a "touch") or exceeds (a "cross") the best ask price. For trade records, a trade is counted as an "immediate match" if the trading price is the same as either the immediately preceding best ask or best bid quote. For SPY data, the integrity check also counts "delayed matches" for various delay lengths, a category discussed in Appendix B.4.

Figure B.1. (Color online) Data Feed

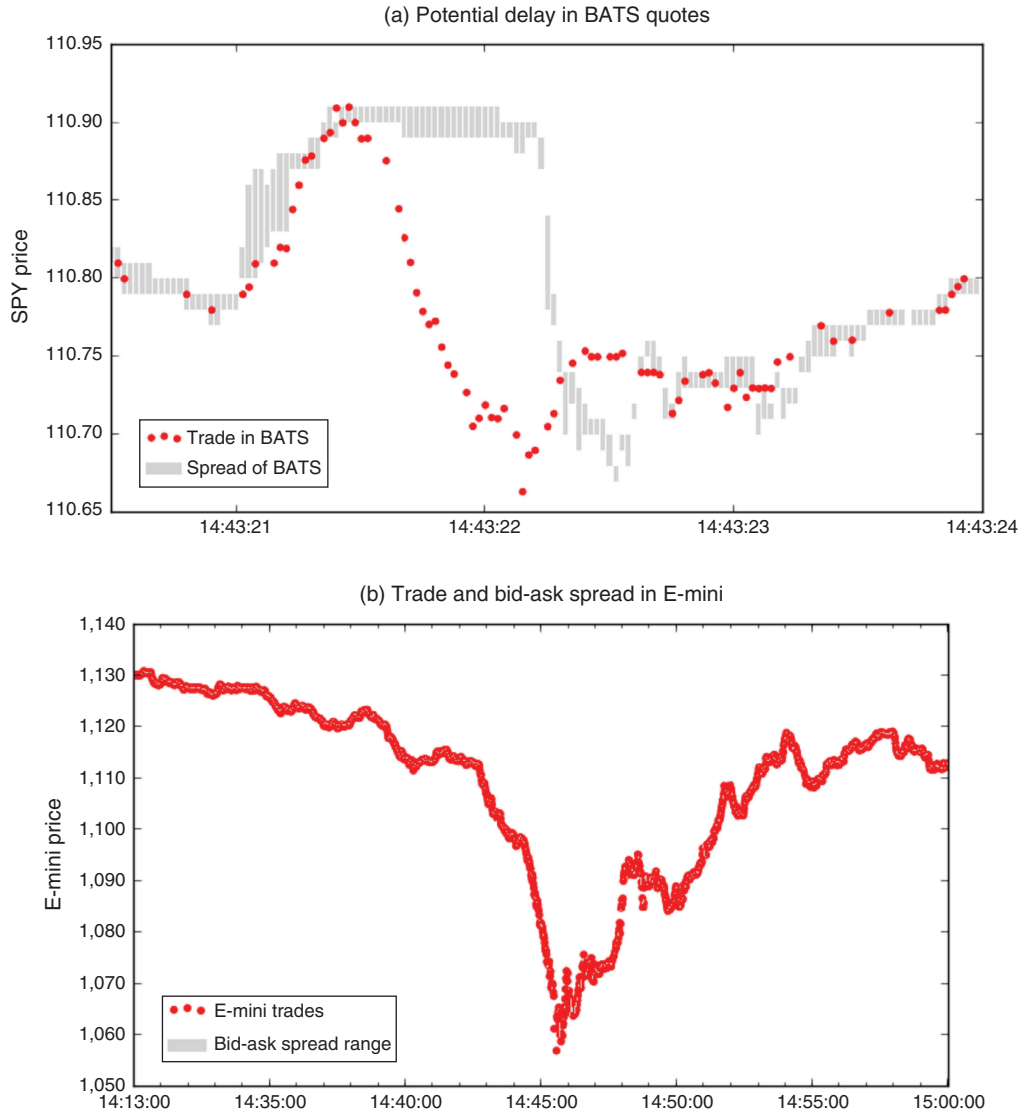
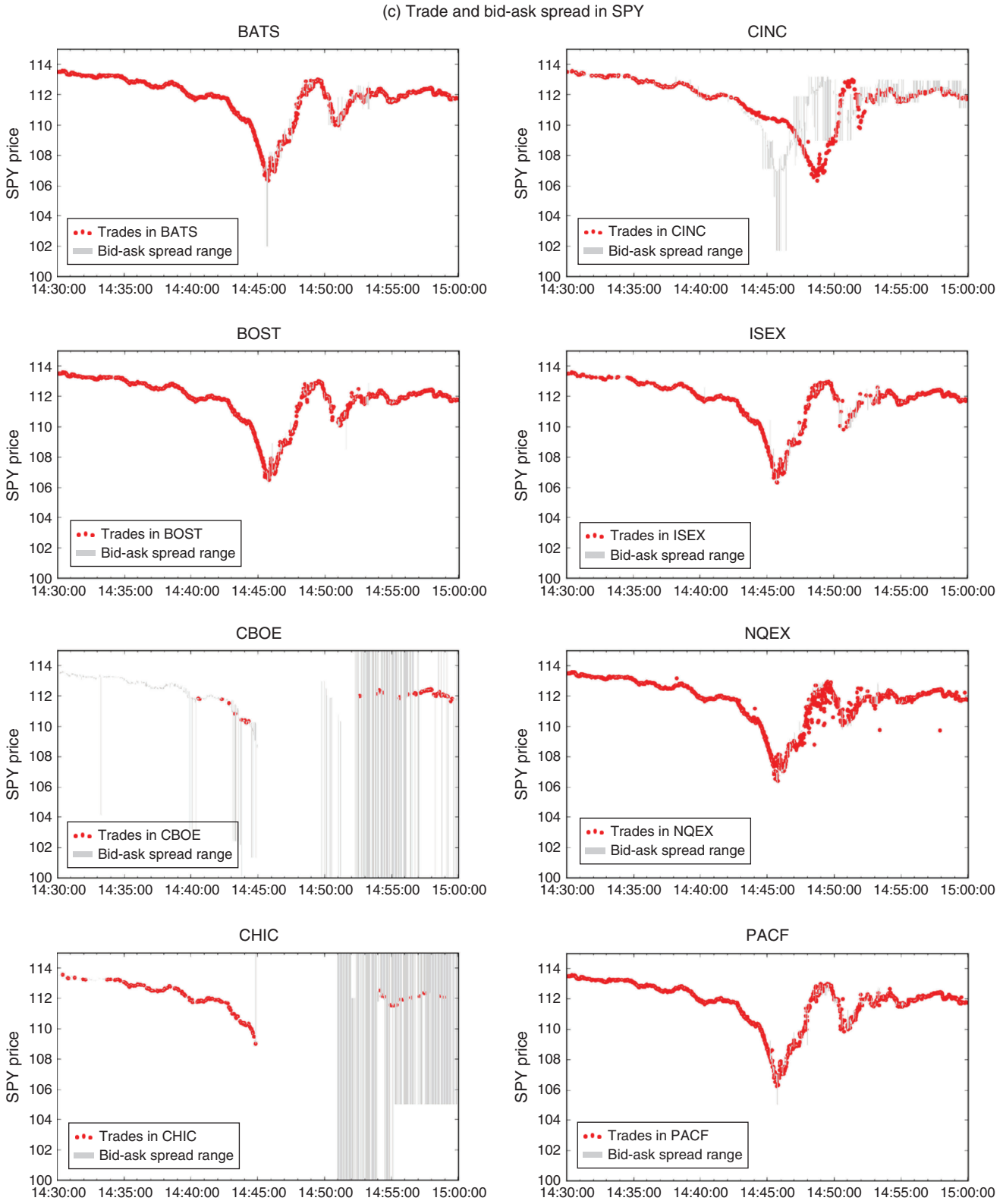


Figure B.1. (Color online) (Continued)



Notes. This figure illustrates the quality of the data feed by plotting the bid–ask spread and (average) trade price. Panel (a) does so for a couple of seconds of BATS data at a 25-millisecond granularity. It illustrates what is, most likely, a delay in the BATS quote feed. Panels (b) and (c) plot the spread and trade price for the Flash Crash half hour period at a one-second granularity. Panel (b) illustrates the E-mini data feed. Panel (c) illustrates the data feed of the eight exchanges that traded SPY.

Table B.1 summarizes the results. It distinguishes between the Flash Crash period (14:30–15:00) and the "normal" noncrash period (9:30–16:15).

E-mini. E-mini contracts traded only on CME Globex. The 2.3 million quote records in the day-trading hours do not contain crossing quotes, and all touching quotes occurred during the five-second trading halt from 14:45:28 to 14:45:33. In the Flash Crash period, more than 99.8% of the trade records can be matched with either the best bid or the best ask quote that immediately preceded them. If instead of a simple count, mismatches are weighted by volume, the matching rate also exceeds 99.8%. In the noncrash period, the matching rate is above 99.9%.

SPY. SPY traded on eight exchanges: BATS, BOST, CBOE, CHIC, CINC, ISEX, NQES, and PACF. Table B.1 shows that there are no best bids touching or crossing the best ask within each exchange. The table further documents that trade records matching is much more problematic. In the Flash Crash period, only 67.4% of the trade records can be matched with immediately preceding quote prices. This low matching rate, however, is a more general characteristic of the market as in the noncrash period too the rate is low: 73.3%. Given the 90-second delay allowed on reporting trades, we should be able to match trade prices with price quotes with a delay of up to 90 seconds. Matching rates improve substantially when allowing for such delay. For example, they become 91.8% and 95.4% for a window length of 100 milliseconds and 98.8% and 99.9% for a window of 90 seconds. This pattern is fairly consistent across exchanges except for CINC. In this case, even the 90-second matching rate during the Flash Crash period remains as low as 83.0% (consistent with Figure B.1).

B.4. Data Used for Analysis

To sum up, the data quality check leads to the following observations. First, price quotes seem largely reliable as there are no intra-market touches or crosses, and time stamps largely line up with the time stamps on trades given the 90-second delay allowed for trade reporting (the only exception being CINC during the Flash Crash period although CINC generates less than 1% of all trades; see Table B.1). All analysis based on price quotes, therefore, uses the available data without applying any filters.

Second, the time stamp on trades is an unreliable indicator of execution time as exchanges are allowed to delay trades by up to 90 seconds. Net flow is obtained by signing trades according to the side of initiation; a trade obtains a positive sign if it was a buyer who hit an ask quote and a negative

Table B.1. Basic Data Integrity Checks

	#Quotes	Touches	Crosses	#Trades	Immediate matches	−25ms′ matches	−100ms′ matches	−500ms′ matches	−10s′ matches	−90s′ matches
				(a) Flash crash period (14:30–15:00)						
E-mini	316,000	1,363[a]	0	191,000	99.8%	—	—	—	—	—
SPY										
BATS	269,000	0	0	50,000	65.4%	81.1%	95.9%	98.2%	99.3%	99.5%
BOST	100,000	0	0	22,000	48.3	66.7	91.3	96.7	98.8	99.2
CBOE	9,900	0	0	404	68.3	69.3	86.1	98.0	99.3	99.3
CHIC	28,000	0	0	547	55.4	72.6	90.3	97.4	99.8	99.8
CINC	49,000	0	0	1,800	28.2	30.3	41.0	57.5	64.8	83.0
ISEX	125,000	0	0	5,000	34.9	53.4	91.8	98.0	99.2	99.4
NQEX	171,000	0	0	144,000	71.7	81.7	90.5	93.8	96.6	98.8
PACF	229,000	0	0	48,000	71.3	83.5	94.9	97.7	99.3	99.6
All SPY	982,000	0	0	272,000	67.4	79.6	91.8	95.2	97.4	98.8
				(b) May 6 trading hours, excluding flash crash period (9:30–14:30, 15:00–16:15)						
E-mini	1,980,000	0	0	839,000	99.9%	—	—	—	—	—
SPY										
BATS	1,963,000	0	0	282,000	72.2%	86.3%	97.6%	99.2%	99.9%	99.9%
BOST	623,000	0	0	75,000	52.4	70.1	91.4	97.0	99.3	99.7
CBOE	58,000	0	0	655	50.4	53.9	76.6	91.3	97.7	98.5
CHIC	180,000	0	0	2,800	64.1	78.9	91.8	94.7	99.7	99.9
CINC	212,000	0	0	3,600	36.7	41.1	65.2	92.0	99.8	99.9
ISEX	735,000	0	0	12,000	48.4	69.0	96.2	99.2	99.7	99.9
NQEX	1,050,000	0	0	570,000	75.9	86.1	94.9	97.7	99.6	99.9
PACF	1,731,000	0	0	282,000	77.2	88.1	96.9	98.7	99.6	99.8
All SPY	6,551,000	0	0	1,232,000	73.3	85.1	95.4	98.0	99.4	99.9

Notes. This table presents some basic data integrity checks. Panel (a) reports statistics for the half-hour Flash Crash period. Panel (b) reports the same statistics for the other trading hours on May 6, 2010. The leftmost columns focus on quotes. They report (i) the number of quote updates, (ii) the number of times the best bid equals the best ask ("touch"), and (iii) the number of times the best bid exceeds the best ask ("cross"). The rightmost columns focus on trades. They report the number of trades as well as how many of the trades can be matched with a best bid or ask quote that preceded them. A match is reported if the trade price matches the preceding bid or ask quote or if the trade price is within the best bid–ask spread in a preceding time interval. The interval lengths considered are 25 milliseconds, 100 milliseconds, 500 milliseconds, 10 seconds, or 90 seconds. This exercise is motivated by the 90-second time delay that is allowed for exchanges when distributing their trade reports. There is no such time delay for distributing quotes.

[a]All these touching quotes occured during the five-second trading halt (14:45:28–14:45:33).

sign if a seller hit a bid quote. For each exchange, the signing is done by comparing the trade price with the bid and ask quotes that precede it. In the Flash Crash period, on average, 99.8% of E-mini trades and 67.4% of SPY trades can be signed this way. The tick rule is used for all other trades (Lee and Ready 1991).

More than 90% of the mismatches are, arguably, the result of a delay less than 100 milliseconds (see Table B.1). It is for this reason that the VAR/VECM analysis aggregates the data over 100-millisecond intervals. This seems to be a fair trade-off between (i) low frequency, to reduce the risk of erroneous sequencing, and (ii) high frequency, to identify the interrelationship between all variables in the system. To further improve data quality, trades from CINC (many delays beyond 90 seconds), CBOE, and CHIC (activity disappears for several minutes) are removed. In total, these trades make up 1% of all trades in the Flash Crash period. The result of all other analysis is not sensitive to trade time stamp inaccuracies in the order of 100 milliseconds.

Appendix C. Rank-on-Rank Regression Without NYSE Arca

This appendix redoes the analysis of signature Table 1 after removing all observations from NYSE Arca. This venue's data feed became extremely strained during the Flash Crash to the point that both NASDAQ and BATS declared self-help. Redoing the rank-on-rank regressions without NYSE Arca, therefore, serves as a robustness check. The results in Table C.1 are largely similar to Table 1 testifying to the robustness of the rank-on-rank regression results.

Table C.1. Rank-on-Rank Regression Without NYSE Arca

(a) Crash time based on the first time when the price is below p% of the 14:30:00 price

Crash start at (# detected)	$p = 4.0\%$ (50 stocks)			$p = 5.0\%$ (50 stocks)			$p = 6.0\%$ (50 stocks)			$p = 7.0\%$ (50 stocks)		
Slackness	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Quote dispersion	0.284**	0.249*	0.252*	0.247*	0.276*	0.321**	0.294**	0.304**	0.352**	0.220	0.288**	0.345**
	(2.07)	(1.73)	(1.80)	(1.78)	(1.93)	(2.35)	(2.12)	(2.16)	(2.57)	(1.56)	(2.05)	(2.48)
	[0.64]	[0.56]	[0.52]	[0.74]	[0.64]	[0.62]	[0.76]	[0.70]	[0.66]	[0.80]	[0.78]	[0.76]
Volatility	0.028	0.046	0.161	0.090	0.100	0.159	0.073	0.088	0.135	−0.016	−0.049	−0.017
	(0.21)	(0.32)	(1.15)	(0.66)	(0.71)	(1.15)	(0.53)	(0.64)	(0.97)	(−0.11)	(−0.35)	(−0.12)
	[0.96]	[0.96]	[0.96]	[0.98]	[0.98]	[0.96]	[0.98]	[0.98]	[0.96]	[0.98]	[0.98]	[0.96]
Order flow	0.374***	0.307**	0.253*	0.345**	0.254*	0.241*	0.307**	0.164	0.151	0.335**	0.217	0.200
	(2.69)	(2.09)	(1.81)	(2.44)	(1.75)	(1.75)	(2.17)	(1.15)	(1.10)	(2.33)	(1.53)	(1.43)
	[0.96]	[0.96]	[0.94]	[0.98]	[0.96]	[0.94]	[0.98]	[0.96]	[0.94]	[0.98]	[0.96]	[0.94]
VPIN	0.005	−0.097	−0.111	0.005	0.005	−0.031	0.024	0.189	0.136	0.030	0.207	0.156
	(0.03)	(−0.60)	(−0.65)	(0.03)	(0.03)	(−0.19)	(0.16)	(1.20)	(0.81)	(0.20)	(1.32)	(0.92)
	[0.44]	[0.24]	[0.14]	[0.46]	[0.26]	[0.16]	[0.50]	[0.26]	[0.20]	[0.52]	[0.28]	[0.20]
R-squared	0.215	0.134	0.173	0.191	0.147	0.207	0.193	0.183	0.205	0.160	0.185	0.177
Total obs.	50	50	50	50	50	50	50	50	50	50	50	50

(b) Crash time based on n consecutive price drops, each more than 2 (normal-time) standard deviations

Crash start at (# detected)	$n = 2$ consecutive drops (50 stocks)			$n = 3$ consecutive drops (48 stocks)			$n = 4$ consecutive drops (40 stocks)			$n = 5$ consecutive drops (29 stocks)		
Slackness	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Quote dispersion	0.104	0.055	−0.034	0.199	0.161	0.154	0.196	0.229*	0.256**	0.109	0.073	0.096
	(0.70)	(0.37)	(−0.25)	(1.45)	(1.18)	(1.15)	(1.47)	(1.85)	(2.09)	(0.86)	(0.57)	(0.73)
	[0.26]	[0.22]	[0.20]	[0.62]	[0.58]	[0.56]	[0.83]	[0.75]	[0.69]	[0.79]	[0.71]	[0.71]
Volatility	0.166	0.233	0.423***	0.059	0.123	0.221	0.013	0.176	0.281**	−0.043	0.081	0.144
	(1.14)	(1.59)	(3.04)	(0.44)	(0.92)	(1.64)	(0.10)	(1.45)	(2.27)	(−0.34)	(0.65)	(1.08)
	[1.00]	[0.94]	[0.92]	[1.00]	[0.96]	[0.94]	[1.00]	[1.00]	[1.00]	[1.00]	[1.00]	[1.00]
Order flow	−0.019	0.046	0.059	0.201	0.175	0.162	0.113	0.081	0.120	0.116	0.082	0.161
	(−0.13)	(0.30)	(0.43)	(1.44)	(1.26)	(1.21)	(0.84)	(0.64)	(0.98)	(0.90)	(0.63)	(1.21)
	[0.88]	[0.86]	[0.82]	[0.98]	[0.96]	[0.92]	[1.00]	[1.00]	[0.97]	[1.00]	[1.00]	[1.00]
VPIN	0.139	0.082	0.117	0.273*	0.356**	0.401**	0.369**	0.493***	0.440***	0.351**	0.421***	0.240
	(0.87)	(0.49)	(0.70)	(1.86)	(2.32)	(2.46)	(2.59)	(3.55)	(2.94)	(2.57)	(2.96)	(1.48)
	[0.16]	[0.08]	[0.02]	[0.44]	[0.19]	[0.06]	[0.61]	[0.39]	[0.22]	[0.71]	[0.50]	[0.25]
R-squared	0.071	0.071	0.194	0.213	0.219	0.246	0.240	0.348	0.349	0.205	0.217	0.134
Total obs.	50	50	50	50	50	50	50	50	50	50	50	50

Notes. This table redoes the signature Table 1 after removing all observations from NYSE Arca. This venue's data feed became extremely strained during the Flash Crash to the point that both NASDAQ and BATS declared self-help. Redoing the rank-on-rank regressions without NYSE Arca therefore serves as a robustness check. The superscripts *, **, and *** indicate significance at 10%, 5%, and 1%, respectively.

Endnotes

[1] The CFTC-SEC Advisory Committee on Emerging Regulatory Issues included academics as well as industry professionals. The academic members are Robert Engle, Maureen O'Hara, David Ruder, and Joseph Stiglitz.

[2] http://www.sec.gov/news/otherwebcasts/2012/ttr100212.shtml, accessed August 6, 2018.

[3] The Nanex system itself did not seem to suffer any delay in the Flash Crash period. One check that has been performed is to compare the time stamps on SPY quotes from NASDAQ ITCH with Nanex time stamps. The distance between the two time stamps was the same in the half hour Flash Crash period compared with earlier trading hours that day.

[4] HFTs often engage in cross-arbitrage. Scott Patterson claimed that at least two of them, Tradebot and Tradeworx, closed down their computer systems during the Flash Crash. "Did Shutdowns Make Plunge Worse?," *Wall Street Journal*, May 7, 2010.

[5] We do not use a net-return proxy because trading fees are nontrivial to determine. First, fee structures differ across traders (e.g., because of volume discounts). Second, information on fees is often not publicly available, further complicating calculation of the proxy.

[6] "Flash Crash" Charges Filed. *Wall Street Journal*, April 21, 2015.

[7] Realized volatility is computed as the standard deviation of log-price changes at a 100-millisecond frequency, exponentially weighted with decay 0.9. Order flow imbalance is the net difference between buy and sell volume in 100-millisecond intervals. VPIN is constructed following Easley et al. (2012) but with smaller volume buckets so as to have sufficient observations in the SPC training sample (see Appendix A).

[8] More specifically, Nanex defines the start of a crash as "the stock had to tick down at least 10 times before ticking up—all within 1.5 seconds and the price change had to exceed 0.8%" (http://www.nanex.net/FlashCrashEquities/FlashCrashAnalysis_Equities.html, accessed August 6, 2018).

[9] Coincidentally, this wealth level roughly corresponds to the value of the total number of resting orders in the E-mini book at the start of that day (100,000 contracts worth about $5.5 billion) and the amount available just before the large seller began selling (90,000 contracts) (CFTC and SEC 2010a, p. 11).

References

Barsky RB, Juster FT, Kimball MS, Shapiro MD (1997) Preference parameters and behavioral heterogeneity: An experimental approach in the heath and retirement survey. *Quart. J. Econom.* 112(2):537–579.

Ben-David I, Franzoni F, Moussawi R (2012) ETFs, arbitrage, and shock propagation. Working paper, The Ohio State University, Columbus, OH.

Borkovec M, Domowitz I, Serbin V, Yegerman H (2010) Liquidity and price discovery in exchange-traded funds: One of several possible lessons from the flash crash. Report, Investment Technology Group, New York.

Born BE, Brennan JJ, Engle RF, Ketchum RG, O'Hara M, Philips SM, Ruder DS, Stiglitz JE (2011) Recommendations regarding regulatory responses to the market events of May 6, 2010. Report, U.S. Commodity Futures Trading Commission, Washington, D.C.

Cespa G, Foucault T (2014) Illiquidity contagion and liquidity crashes. *Rev. Financial Stud.* 27(6):1615–1660.

CFTC and SEC (2010a) Findings regarding the market events of May 6, 2010. Report, U.S. Commodity Future Trading Commission and U.S. Securities and Exchange Commission, Washington, D.C.

CFTC and SEC (2010b) Preliminary findings regarding the market events of May 6, 2010. Report, U.S. Commodity Future Trading Commission and U.S. Securities and Exchange Commission, Washington, D.C.

Cochrane JH (2005) *Asset Pricing* (*Revised Edition*) (Princeton University Press, Princeton, NJ).

Easley D, López de Prado MM, O'Hara M (2012) Flow toxicity and liquidity in a high frequency world. *Rev. Financial Stud.* 25(5):1457–1493.

Goldstein I, Li Y, Yang L (2013) Speculation and hedging in segmented markets. *Rev. Financial Stud.* 27(3):881–922.

Grossman SJ, Miller MH (1988) Liquidity and market structure. *J. Finance* 43(3):617–633.

Hendershott T, Menkveld AJ (2014) Price pressures. *J. Financial Econom.* 114(3):405–423.

IIROC (2010) Review of the market events of May 6, 2010. Report, Investment Industry Regulatory Organization of Canada, Toronto.

Jones CM (2013) What do we know about high-frequency trading? Working paper, Columbia Business School, New York.

Kirilenko A, Kyle AS, Samadi M, Tuzun T (2017) The flash crash: High-frequency trading in an electronic market. *J. Finance* 72(3):967–998.

Kyle AS (1985) Continuous auctions and insider trading. *Econometrica* 53(6):1315–1336.

Kyle AS, Obizhaeva AA (2016) Large bets and stock market crashes. Working paper, University of Maryland, College Park, MD.

Landier A, Simon G, Thesmar D (2015) The capacity of trading strategies. Working paper, MIT Sloan School of Management, Cambridge, MA.

Lee CM, Ready MJ (1991) Inferring trade direction from intraday data. *J. Finance* 46(2):733–746.

Madhavan AN (2012) Exchange-traded funds, market structure and the flash crash. *Financial Analysts J.* 68(4):20–35.

Mehra R, Prescott EC (1985) The equity premium: A puzzle. *J. Monetary Econom.* 15(2):145–161.

Montgomery DC (2009) *Introduction to Statistical Quality Control*, 6th ed. (John Wiley & Sons, Hoboken, NJ).

SEC (2001) Report of the advisory committee on market information: A blueprint for responsible change. Report, U.S. Securities and Exchange Commission, Washington, D.C.

van Kervel V (2015) Competition for order flow with fast and slow traders. *Rev. Financial Stud.* 28(7):2094–2127.

Wang X, Ye M (2017) Who supplies liquidity, and when? Working paper, Nanyang Technology University, Singapore.

Zhang F, Powell SB (2011) The impact of high-frequency trading on markets. *CFA Magazine* 22(2):10–11.