Università della Svizzera italiana

and

Swiss Finance Institute

Doctoral Dissertation

Asset Prices and Demand Shocks

Author:

Andrea
BARBON

Thesis Advisor:

Prof. Francesco
FRANZONI

A dissertation submitted in fulfillment of
the requirements for the degree of
PhD in Economics

SFI PhD in Finance
in the
Institute of Finance
(Faculty of Economics)

Dissertation Committee:

Prof. Francesco FRANZONI, Università della Svizzera italiana
Prof. Laurent FRESARD, Università della Svizzera italiana
Prof. Augustin LANDIER, HEC Paris

Abstract

My dissertation consists of three chapters, each of which focuses on a different area of research in asset pricing.

The first chapter deals with the informational role of brokerage firms during fire sales in the equity market. The second chapter exploits the ETF program by the bank of Japan as a quasi-natural experiment to measure the slope of the equity demand curve. The last chapter presents the application of a novel machine learning model, based on an attention mechanism, to high-frequency data from the Nasdaq.

In the first paper, co-authored with F. Franzoni, M. di Maggio, and A. Landier, we use trade-level data to study whether brokers play a role in spreading order flow information in the stock market. We focus on sizeable portfolio liquidations from large investment funds, which we name "fire sales", resulting in significant temporary price drops for the liquidated stocks. Next, leveraging on our unique dataset, we identify the brokers who intermediate these fire-sale trades. We show that these brokers' clients are more likely to predate on the liquidating funds than to provide liquidity. Predation leads to profits of about 25 basis points over 10 days, on average, and increases the liquidation costs of the distressed fund by 40%. This evidence suggests a role of information leakage in exacerbating fire sales. The highlighted mechanism might be a concern to regulators as well, since it can exacerbate the costs associated with fire sales, especially during times of scarce liquidity.

In the second paper, co-authored with V. Gianinazzi, we focus on the Japanese equity market to study a peculiar intervention by the Japanese central bank. Since the introduction of its quantitative and qualitative easing program in 2013, indeed, the Bank of Japan has been increasing its holdings of Japanese equity through large-scale purchases of index-linked ETFs to lower risk premiums. We exploit the cross-sectional heterogeneity of the supply shock to identify a positive and persistent impact on stock prices, consistent with a portfolio balance channel. The evidence suggests that long-run demand curves for stocks are downward sloping with unitary price elasticity. Moreover, we are able to quantify the effect of the intervention, showing that the purchases of ETFs tracking the price-weighted Nikkei 225 generate pricing distortions relative to a value-weighted benchmark.

The third paper starts from the observation that, even though machine learning methods are able to deliver superior forecasting accuracy, they can hardly be used to make inference. To overcome this limitation, I propose an encoder-decoder neural network augmented with an attention-based mechanism that can autonomously learn to identify the most critical regions of the input data. I first train the model using high-frequency message data from the NASDAQ and show that it outperforms other state-of- the-art models in forecasting future transaction prices. Then, I develop a methodology that uses the attention mechanism to make inference on the relative share of information content of market orders versus limit orders, concluding that the most informative events are executions of market orders while submission and cancellations of limit orders are less relevant. Finally, I test the model's behavior during the execution of real block orders from institutional investors, showing that it favors liquidity provision rather than front-running strategies.

Brokers and Order Flow Leakage: Evidence from Fire Sales

Andrea Barbon, Marco Di Maggio, Francesco Franzoni, and Augustin Landier *

Abstract

Using trade-level data, we study whether brokers play a role in spreading order flow information in the stock market. We focus on large portfolio liquidations which result in temporary price drops, and identify the brokers that intermediate these trades. These brokers' clients are more likely to predate on the liquidating funds than to provide liquidity. Predation leads to profits of about 25 basis points over 10 days and increases the liquidation costs of the distressed fund by 40%. This evidence suggests a role of information leakage in exacerbating fire sales.

*Barbon: USI Lugano and Swiss Finance Institute; Di Maggio: Harvard Business School and NBER; Franzoni: USI Lugano, Swiss Finance Institute, and Center for Economic Policy and Research; Landier: Toulouse School of Economics. Corresponding Author: Francesco Franzoni, (francesco.franzoni@usi.ch). We thank Malcolm Baker, John Campbell, Laurent Fresard, Slava Fos (discussant), Joel Hasbrouck, Terry Hendershott (discussant), Gary Gorton, Owen Lamont, Jongsub Lee (discussant), Andrew Lo, Toby Moskowitz, Abhiroop Mukherjee (discussant), Cameron Peng (discussant), Erik Stafford, and seminar/conference participants at Aalto University, the Becker Friedman Institute CITE conference on New Quantitative Models of Financial Markets, Bocconi University, EIEF, FCA, FINRA Market Structure conference, FIRS, LAEF 2nd OTC Markets and Securities Workshop, INSEAD, Macro-Prudential Conference ECB, New York Federal Reserve Bank, AQR, UC Dublin, and WFA for helpful comments. Augustin Landier directs financial research at AXA-Investment Management Chorus, a quantitative asset management affiliate of Axa-IM. Neither Andrea Barbon, Marco Di Maggio, Francesco Franzoni nor any of their relatives have received significant financial support from any interested party, or have positions in relevant organizations. The authors have signed a nondisclosure agreement for the proprietary data used in this study.

1 Introduction

Large institutional orders are typically split into smaller amounts over time to avoid moving the market (see Garleanu and Pedersen (2013), Di Mascio, Lines and Naik (2016)). One concern when executing an order over time is that other traders might anticipate the intent to trade the stock in the near future and trade in the same direction to benefit from the future price impact. This problem is particularly pronounced in the case of fire sales, during which the seller is forced to bring to the market a large quantity of assets in a limited amount of time (Coval and Stafford, (2007), Ellul, Jotikasthira and Lundblad (2011)). Moreover, if the liquidation occurs during a time of market stress, predatory trading can make the market more illiquid and amplify adverse shocks (Greenwood, Landier, and Thesmar (2015)). Given this possibility, some observers suggest that reducing the frequency of portfolio disclosure can help prevent predatory behavior (Brunnermeier and Pedersen (2005)).

However, the market may possess information about forced liquidations due to broker's close relationship with the liquidating managers. Brokers are uniquely able to observe the daily trades of a fund. In the case of hedge funds, prime brokers also operate as lenders and risk managers, and thus they know when a fund is about to breach some risk limit and deleverage its portfolio. They can also infer their client's trading habits, such as whether a client tends to cut trades into small orders over several days when executing a large order. Thanks to this information, brokers are well placed – indeed they may be best placed among market participants – to predict the future trades of their clients.

In an effort to establish a reputation as a source of valuable information and attract new business, brokers may leak the news that a client's large trade is likely to extend over time, as other investors can use this information to predate on the distressed fund. Alternatively, brokers may be reluctant to foster predatory trading against a client, as doing so may work against their reputation. According to the latter argument, brokers should instead invite other traders to provide liquidity and take the other side of the slow trade. Thus, whether brokers foster predatory trading or whether they support liquidity provision in the case of slow trading by a client is an open empirical question. In this paper, we shed light on such a question using data on forced liquidations of portfolio holdings.[1] Specifically, we exploit proprietary trade-

[1] We focus on large liquidations (which for convenience we label "fire sales"), that is, we do not include large purchases in our analysis, because we aim to cleanly identify liquidity-motivated trades. In our data, the majority of institutional investors are long-only (about 90%) and thus a sale is less likely to be information motivated (as the manager would need to already have the stock in her portfolio) than a buy transaction. In addition, large cash inflows can be allocated slowly over time and are, therefore, less likely to impose a concentrated liquidity demand on the market than large outflows., Fire sales can also

level data to identify asset managers that sell a significant fraction of their portfolio over a relatively short amount of time. We restrict attention to asset managers whose order flow is abnormally negative for at least five days in a row. Moreover, we focus on managers that liquidate multiple stocks (on average about 20 stocks) at a significantly faster pace than usual. We identify approximately 400 events satisfying these criteria over the period 1999 to 2014. We verify that the stock price movements resulting from such sales are only temporary, consistent with the identification of liquidity events – the price impact would have to display a permanent component if sales were motivated by fundamentals.

Not all brokers employed by a liquidating fund will be aware that the fund is in distress. The liquidating fund has little incentive to disclose its intention to liquidate a large fraction of its portfolio. Indeed, it is likely to use multiple brokers (on average 29) to minimize the price impact and information leakage. We therefore label as *aware* only those brokers that intermediate a large enough fraction of volume. We find that the probability of predatory behavior is significantly higher for orders executed through aware brokers. In particular, clients of aware brokers are much more likely to execute sell trades in the same stocks with the same broker over the same period. While clients of aware brokers also engage in liquidity provision, this activity does not appear to be as prevalent as predatory trading.

We next explore the extent of heterogeneity across different clients of aware brokers. If brokers leak information about order flow, they are more likely to leak information to their best clients, that is, to those clients from which they extract the highest rents. As proxies for the strength of the investor-broker relation, we employ both the trading volume and the commissions generated by a client. Using both measures we find that best clients of the aware brokers are significantly more likely than other clients to sell the stocks that the liquidating manager is offloading during the fire sale compared to immediately before the fire sale. [2,3] The magnitude of this result is economically significant: the net probability of predation for the best clients of aware brokers is more than twice that for the small clients of aware brokers. This evidence suggests that predation is more likely than liquidity provision among the best clients of the brokers that intermediate fire sales. As additional evidence of predatory trading, we find that a significant fraction of the positions sold by other managers than the distressed

pose a systemic threat if they cause a propagation of idiosyncratic shocks to the balance sheets of other investors. Hence, studying the effect of information leakage on fire sales is especially relevant, including from a regulatory perspective.

[2] We find that these relations are extremely persistent, consistent with the findings in Goldstein et al. (2009), which suggests that brokers might have an incentive to nurture such relations.

[3] We control for time, manager, event, stock, and broker fixed effects. Hence, differences across stocks, such as their liquidity, or across brokers, such as their ability to execute, cannot explain our results. We also run a specification in which we control for broker-manager relationship fixed effects, which controls for the matching between asset managers and brokers. The results, which are in the Internet Appendix, are qualitatively similar.

fund during the fire-sale period (30% to 42%), are bought back in the 10 days following the fire sale.

We conduct several robustness checks to rule out the possibility that a fire sale's originator and followers are trading in response to a common information signal. We first exclude from our sample all events that occurred during recessions. We next exclude all events that occur around earnings announcements, changes in analyst recommendations, or any other type of negative news as reported by the press and classified by the data provider Ravenpack. We further exclude stocks with negative momentum and high short interest to mitigate the concern that selling managers follow similar trading strategies founded on a negative signal on the stock. In each case, our results continue to hold.

To strengthen our identification of fire-sale events, we examine a natural experiment in which some mutual funds were forced to liquidate their holdings. Specifically, as a consequence of the late-trading scandal of 2003, 27 fund families experienced significant outflows. Anton and Polk (2014) use these outflows to identify an exogenous determinant of mutual funds' selling activity. Kisin (2011) estimates that funds of implicated families lost 14.1% of their capital within one year and 24.3% within two years. Important for our purposes, the brokers of the liquidating funds were aware of both the specific stocks that were being sold and the timing of these liquidations. We show that, after the scandal broke out, the clients of the relevant brokers were significantly more likely to liquidate the same stocks on the same days on which the implicated funds are also selling. Our results thus continue to hold even when we consider plausibly exogenous variation in the source of the liquidation, which increases our confidence in our findings.

One of the contributions of our paper concerns the value of order flow information. We compute the profits that asset managers make during fire sales and show that the best clients of aware brokers, that is, the investors most-likely to benefit from information leakage, are able to earn an additional 25 basis points (bps) in the few days of the fire sale. Given average fund performance, these results suggest that being able to predict fire sales can be quite profitable.

We also provide evidence on the externalities arising the losses incurred by managers exposed to predation. Focusing on the execution shortfall, which we compute as the volume-weighted percentage difference between the execution price and a benchmark price, we find that the price impact is about 40% higher when the trades are executed through brokers that

are aware of liquidations. We interpret this spread as the cost of predation. Our evidence also highlights an important mechanism through which asset price fluctuations are amplified.

We conclude by addressing the question of what brokers gain from leaking order flow information. To do so, we examine brokers' commissions. We find that, relative to other clients of aware brokers and relative to the period before the fire sale, the best clients – those who take advantage of the order flow information by preying on the liquidating funds – pay 16% higher commissions in standard deviation units. We thus document that brokers get rewarded for the information they provide.

Overall, our findings highlight an important trade-off between slow trading execution meant to reduce the price impact, as, for example, in Kyle (1985), and leakage of order flow information. The latter becomes more likely when asset managers trade in the same direction over an extended period of time. This consideration is not confined to fire-sale events. Indeed, we find that the autocorrelation among large trades, that is, those larger than 1% of average daily volume, in our data is about 35%. Hence, as a rule, managers tend to trade in the same direction over multiple days, which opens the possibility for brokers to predict future order flow.

Our paper bridges a vast literature on fire sales[4] and a growing literature on the role of networks among market participants in various domains, including Li and Schürhoff (2018), Di Maggio et al. (2018), Di Maggio, Kermani, and Song (2017), Hollifield, Neklyudov, and Spatt (2016), Afonso, Kovner, and Schoar (2013) and Hendershott et al. (2016). Our novel contribution is to highlight the key role played by brokers during fire sales, which might be amplified due to brokers leaking order flow information.[5]

[4] Theoretically, Shleifer and Vishny (1992, 1997) and Kiyotaki and Moore (1997) suggest that fire sales occur when the natural buyers are unable to purchase assets due, for instance, to agency problems. Brunnermeier and Pedersen (2005) and Di Maggio (2016) show that the market might become illiquid exactly when liquidity is needed most due to unconstrained arbitrageurs taking advantage of the temporary price pressure by selling and then buying the asset back only after the fire sale has ended. See Shleifer and Vishny (2011) for a survey of this literature. Studies on fire sales and price dislocations in financial include, among others, Allen and Gale (1994), Gromb and Vayanos (2002), Brunnermeier and Pedersen (2009), Acharya, Gale, and Yorulmazer (2011), and Garleanu and Pedersen (2011). Recently, Yang and Zhu (2016) provide a two-period Kyle (1985) model of "back-running" where, in addition to informed and noise traders there exists an investor who learns from the order flow generated by the informed speculator after the order is filled.

[5] Our findings also relate to a growing literature examining the way in which information spreads in financial markets, for instance, via to information percolation (Duffie, Malamud, and Manso (2009, 2014)), or network effects (Babus and Kondor, (2016) and Walden, (2016)). We contribute to this literature by providing empirical support to the notion that information can be readily disseminated through interactions between intermediaries and market participants. Furthermore, our results can inform future theoretical developments of this literature as we point out that this information dissemination is *strategic*, a feature that is missing from existing theoretical literature and that drives network formation in financial markets. Also related to our paper, Farboodi and Veldkamp (2017) develop a long-run growth model in which traders have the option to extract information from order flow data-mining and examine the implications for price informativeness and market liquidity.

Our paper also relates to prior studies documenting that brokers leak value-relevant information to selected clients in other contests. For instance, Irvine, Lipson, and Puckett (2006) find evidence of such information leakage with respect to future analyst recommendations, McNally, Shkilko, and Smith (2015) show that brokers share information about firm insiders' order flow, and Di Maggio et al. (2018) document broker dissemination of informed order flow.

In a recent paper closest to our work, van Kervel and Menkveld (2018) study the behavior of high-frequency traders (HFTs) around large orders of institutional investors. The authors find that HFTs provide liquidity if the order is short-lived (below seven hours) but back-run on the order if it lasts several hours within a day, that is, they trade in the same direction of information-motivated orders. The latter behavior increases trading costs for the institution, as predicted by Yang and Zhu (2016). Similar to van Kervel and Menkveld (2019), we study the interplay among institutional investors and we detect trading behavior by other investors that is harmful to the initiator of a large order. Our evidence differs from and complements their results along several dimensions. First, we focus explicitly on liquidity-motivated orders (i.e., fire sales) and show that predation occurs in these circumstances as well, not just around information-motivated trades. Second, we show that predatory behavior also characterizes traditional asset managers, not just HFTs.[6] Third, we identify institutional brokers as instrumental in spreading order flow information and fostering predation. Finally, we highlight a systemic threat caused by predatory trading, showing that it can amplify price dislocations during fire sales.[7]

The remainder of the paper is organized as follows. Section I describes our data sources and presents summary statistics. Section II discusses our main results on the behavior of asset managers and the role of brokers during fire sales. Section III presents results on the value of order flow information. Finally, Section IV concludes.

[6] Our paper does not focus on high-frequency predation because HFTs are not present in our data. However, the question of liquidity provision versus predation, which we address, has received increasing attention in the HFT literature. Moreover, the destabilizing effect of predation during fire sales that we document also finds a counterpart in studies focusing on the impact of HFTs on market efficiency and volatility. Biais and Foucault (2014), O'Hara (2015), and Menkveld (2016) provide surveys of the rapidly growing HFT literature. Several empirical studies find that HFT activity is beneficial in that it reduces transaction costs (Hendershott, Jones, and Menkveld (2011); Hasbrouck and Saar (2013); Menkveld, (2013); Brogaard et al. (2015); van Kervel (2015)) and improves price efficiency (Boehmer, Fong, and Wu (2014); Brogaard, Hendershott, and Riordan (2014)). The evidence on the relation between HFTs and short-term volatility and crashes is mixed.

[7] Our results are also consistent with Chung and Kang (2016), who use monthly hedge fund returns to document comovement in the returns of hedge funds sharing the same prime broker.

2 Data and Summary Statistics

To examine whether and how brokers leak order flow information during fire sales, one needs detailed trade-level data that also contain information on the institutional investors and brokers involved in each trade. We obtain such data from Abel Noser Solutions, formerly Ancerno Ltd. (we use the name "Ancerno" for simplicity). Ancerno performs transaction cost analysis for institutional investors and makes these data available for academic research under an agreement of nondisclosure of institutional identity.

We have access to identifiers for managers that initiate trades and brokers that intermediate those trades from 1999 to 2014.[8] There are several advantages to this data set. First, clients submit this information to obtain objective evaluations of their trading costs, not to advertise their performance, which implies that the data are not likely to suffer from self-reporting bias. Furthermore, while Ancerno collects trade-level information directly from hedge funds and mutual funds when they use Ancerno for transaction cost analysis, another source of information derives from pension funds that instruct the funds they are invested in to release their trading activities to Ancerno for an independent check. Third, Ancerno is free of survivorship bias as it includes information about institutions that report in the past but at some point terminated their relationship with Ancerno.

Previous studies (e.g., Puckett and Yan (2011), Anand et al. (2012, 2013)) show that the characteristics of stocks traded and held by Ancerno institutions and the return performance of these trades are comparable to those in 13F mandatory filings. Furthermore, using an earlier version of our data, Goldstein, et al. (2009) provide a useful description of the institutional brokerage industry and show that institutions value long-term relations with brokers. They further show that, consistent with our results, the best institutional clients are compensated with superior information around changes in analyst recommendations.

Ancerno information is organized into different layers. At the trade level, we have the transaction date and time (at the minute level of precision for a subset of trades), the execution price, the number of shares traded, the side (buy or sell), and the stock CUSIP. We carry our analysis at the ticket level, we aggregate all trades on the same stock on the same side of market (buy or sell) by the same manager executed through the same broker on the same day.

[8] Relative to the standard release of Ancerno that is available to other researchers, we also obtained manager and broker identifiers for the latest years (that is, after 2011), under the agreement that no attempt be made to identify the underlying institutional names.

Next, we define a fire-sale event. Our goal is to identify liquidity-motivated sales that attract brokers' attention and are likely to generate a significant but temporary price impact. We therefore impose two requirements. First, for a given manager, the selling volume needs to exceed the manager's standard trading volume for a protracted period.[9] Second, at the stock level, the liquidation volume needs to make a sufficient fraction of total trading volume.

More specifically, to identify liquidating funds we start by computing the signed volume Z-score for each manager *m* on day *t* as

$$\mathrm{Z}_{\mathrm{t}}^{\mathrm{m}} = \frac{DVol_t^m - E(DVol_t^m)}{\sigma(DVol_t^m)}, \tag{1}$$

where $DVol_t^m$ is the portfolio-level dollar volume traded by manager *m* on day *t*. The mean and standard deviation of $DVol_t^m$ are estimated over a rolling window of 120 trading days ending one week before day *t*. Next, for a given manager we require that during a fire-sale event, $\mathrm{Z}_{\mathrm{t}}^{\mathrm{m}}$ be below -0.25 for at least five consecutive trading days. This requirement ensures that the sale takes place over a sufficiently long period of time for the broker to become aware of the fire sale and for it to represent a significant event in the life of the fund. Given this condition, all of the fire sales that we identify correspond to events in which the order imbalance at the fund level is negative, as can be seen in Table IA.VIII in the Internet Appendix.

In addition, we impose a filter at the stock level to ensure that the sale volume is large enough to generate price pressure. For stock *j* to be part of a fire-sale event, we require that the volume traded by the manager equal at least 1% of the CRSP volume on day *t* for at least four out of the five fire-sale days.

We decide to keep events in which at least 10 stocks are involved in a fire sale. The goal is to reduce the probability that liquidating funds are selling as a consequence of stock-specific

[9] Our analysis is at the manager code level, that is, at the level of the management company. We focus on the management company level for several reasons. First, our definition of fire sales selects events that are particularly large for an asset manager. In this sense, fire sales are likely to arise when multiple clients withdraw their funds from a management company, in which case focusing on a specific client-manager relationship may miss these larger events. Second, if only one fund in the company is in distress, or just a few, other funds can help by providing liquidity. Specifically, the healthier funds can relieve the distressed fund of some of its assets by engaging in cross-trading, a practice that is described in Gaspar, Massa, and Matos (2006) and more recently in Eisele et a. (2017) using Ancerno data. The possibility of intra-family subsidization motivates us to focus on events that involve the entire family of funds. Finally, the choice to focus on the management company, as opposed to specific funds within the family, is also driven by data availability. In the Ancerno data that are available to us, the alphanumeric identifier for the specific fund (manager) is often missing or not meaningful.

information. Focusing on liquidations of a large number of stocks makes it less likely that the sales are driven by information.

Next, we distinguish between *aware* and *unaware* brokers. Broadly speaking, we say that a broker is aware that a stock is subject to fire-sale pressure if it intermediates sufficiently large volume on that stock arising from the originator. More specifically, for broker b, stock j, day t, and event e combination, $Aware_{j,b,t,e}$ equals one if the volume on stock j originated by the liquidating fund that is intermediated by broker b on day t is above 2% of the average daily volume (ADV) for that stock. Note that this definition does not require that the broker have knowledge of the overall size of the liquidation. Rather, it simply requires that the broker realize that the distressed fund is responsible for a significant fraction of daily volume. In Table IA.I of the Internet Appendix we show that our main results are robust to using different ADV-related thresholds to identify aware brokers (for thresholds from 1% to 5%). In addition, we show that our results are robust to also requiring that the broker intermediate a large volume of at least N stocks in the fire-sale basket to become aware of the fire-sale event (for N = 1, 5, 10).

Panels A through C of Table I provide summary statistics for the key variables in our analysis. We identify a total of 385 fire-sale events over the 1999 to 2014 period. Each of these events lasts at least five days, and the liquidating funds sell on average $377 million worth of stock (median $177 million). Figure IA.2 displays the distribution of events over our sample period. As can be seen, these the events are distributed fairly evenly over time. Indeed, even during the recessions marked with red squares, the number of fire-sale events does not spike. This evidence suggests that our methodology identifies funds subject to idiosyncratic shocks rather than market-wide events.[10]

We can compute the fraction of the liquidated portfolio that the liquidation volume represents. To do so, we estimate liquidating funds' portfolios by cumulating their trades over the two years prior to the fire sale. We then divide the total volume of stocks sold by the reconstructed portfolio size. We find this fraction is sizeable, at 9.16% on average. Notably, this methodology is likely to underestimate liquidating managers' actual portfolios, since we do not know their positions at the beginning of the estimation period, in which case the fraction that we compute represents an upper bound. Nevertheless, this evidence further

[10] The lack of clustering of fire-sale events during crisis periods is an intentional feature of our definition of fire sales. In computing the Z-score, in the numerator we subtract from a given day's order flow the average daily order flow over the prior six months. Hence, if the order flow is negative over a protracted period, such as during a crisis, at some point the Z-score will cease to identify fire sales. The advantage of this feature is that we do not generate a sample of fire sales in which crises are overly represented.

suggests that our sample of large sales is unlikely to be inspired by stock-specific information.

We also find that, on average, 22 stocks are heavily sold during a fire-sale event, with about $17.2 million sold in each stock. Thus, our fire-sale events involve more than just isolated stocks. Figure IA.1 shows the distribution of fire-sale events as a function of the number of stocks, for events involving 10 to 50 stocks, as well as the distribution of the volume of trades by the liquidating fund, which can reach more than two billion dollars in some cases.

We further find that fire sales are intermediated by an average of 29 brokers, while the number of aware brokers per event is on average only 1.7. In addition, the price of the stocks sold in a fire sale declines by 83 bps on average during the first five days of the event (Table I, Panel C), but there is significant variation in the magnitude of the change; for the bottom quartile, for instance, the price drops by more than 3%.

Using TAQ data, we find that fire-sale volume is on average 50% of the TAQ order imbalance (median 10%) and 27% of TAQ sell volume (median 19%). These results imply that the liquidating fund imbalances constitute a sizeable fraction of the TAQ imbalance for fire-sale stocks.

In Panel E of Table I, we provide evidence on the type of stocks that liquidating managers are selling. For each stock in a fire sale, we compute the fraction of the total volume of the fire sale that it represents. We then regress the fraction of the fire sale that stock *j* represents on its weight in the selling manager's reconstructed portfolio, market capitalization, volatility, Amihud (2002) ratio, and various measures of past performance at different horizons. We find that, after controlling for the quantity held by the manager (i.e., portfolio weight), funds tend to sell the larger, more liquid, and less volatile stocks in their portfolio. Asset managers also tend to sell stocks with higher past performance. These findings are consistent with the predictions of theoretical models studying the liquidation strategies in the case of distress (Scholes (2000), Brown, Carlin, and Lobo (2010)). Moreover, the highly significant positive coefficient on the reconstructed portfolio weight suggests that the liquidating funds are not building short positions but rather they are selling positions that are already present in their portfolio, which provides further support for our strategy for identifying fire sales.

Finally, in Panel F of Table I, we investigate the sequence of sales by changing the dependent variable to first day in which the stock is sold, which we calculate as the number of business

days from the first day of the fire sale in which a particular stock is sold for the first time. The results show that the most liquid and less volatile stocks are sold earlier.

3 Main Results

In this section, we start by discussing our empirical strategy. We then present our main evidence on the role of brokers in spreading order flow information during fire-sale events.

3.1 Fire Sales

We start our analysis by characterizing the fire-sale events. Figure 1 plots the average (across stocks and events) daily signed volume (i.e., order imbalance) of liquidating funds during the event window, where day zero is the first day of the five-day window over which we identify a fire sale. The large negative volume before day 0 is due to the fact that, while liquidations likely start earlier, we impose stringent criteria for them to be classified as a fire sale. We note that while the daily order imbalance is smaller in magnitude after five days, it is still negative after 15 days. This observation is important, because it indicates that in general liquidating funds do not repurchase the liquidated stocks back (even when we extend the horizon further out) and thus are not simply selling a stock short because they expect the price to decline, after which they buy the stock back.

Figure 2 plots the average DGTW-adjusted (Daniel, Grinblatt, Titman, and Wermers (1997)) cumulative returns for the stocks included in the fire sales across all events. As can be seen, returns are mostly flat pre-event. They then precipitate quite rapidly as the liquidating fund (for simplicity, the *originator*) sells most intensely, that is, during the five-day interval [0,4]. Returns then start to recover slowly over time. Specifically, we find that return are back to pre-event levels after about 20 days. This reversal is faster than what is documented in prior literature on fire sales (Coval and Stafford, (2007)). On average, the price drops by almost 1% during the interval [0, 4], which we refer to as the *liquidation period*. Importantly, the fact that we observe a reversal over such a short horizon helps rule out the possibility that the liquidation and corresponding price decline are due to negative fundamental news on the stock.

On the contrary, the price path is strongly consistent with price pressure following liquidity motivated trades.

3.2 Predation or Liquidity Provision?

The theoretical literature makes mixed predictions as to whether market participants that anticipate a large liquidity order will predate upon it or provide liquidity. Brunnermeier and Pedersen (2005) and Di Maggio (2016) predict that investors that become aware of a liquidation will predate on the distressed fund which will decrease market quality. In contrast, Admati and Pfleiderer (1991) develop a "sunshine trading" model, in which they argue that investors credibly announcing their intention to transact for non-fundamental reasons attract natural liquidity providers to the market. Bessembinder et al. (2016) provide empirical evidence that is consistent with the latter prediction in the context of predictable roll trades of oil futures contracts by a large ETF. The type of behavior that other market participants adopt vis-à-vis a liquidating fund, however, remains an open empirical question.

To determine whether brokers foster predatory trading or liquidity provision, we estimate the following specification

$$Net\ Predation_{m,i,b,t,e} = \beta_1 Aware_{i,b,t,e} + \varepsilon_{m,i,b,t,e}, \tag{2}$$

where $Aware_{i,b,t,e}$ is a dummy equal to one if broker *b* executing the trades is aware of the sale on stock *i* on day *t* as part of fire-sale event *e*. The dependent variable, $Net\ Predation_{m,i,b,t,e}$, is given as the difference between the probability of predation and the probability of liquidity provision. The probability of predation is a dummy equal to one if client *m* of broker *b* trades in the same direction as the originator (i.e., demands liquidity) on stock *i* on day *t* of event *e*, the dummy is equal to zero if the client provides liquidity by trading in the opposite direction of the originator or if the client does not trade on that stock-day.[11] Symmetrically, the probability of liquidity provision is equal to one if the client trades in the opposite direction of the fire sale, and zero otherwise. We also estimate specifications in which the dependent variable

[11] To identify non-trading clients, we consider all managers that traded with the broker in that stock over the previous 20 business days.

is defined as the net predation variable multiplied by the ratio of the dollar volume of the broker's clients to the market capitalization of the stock (this variable is standardized by subtracting the mean and dividing by the standard deviation). The sample includes trades executed by all managers with all brokers in the database on the fire-sale stocks.

These specifications rely on heterogeneity across brokers for identification, that is, they rely on the fact that some brokers are more exposed to order flow information as they intermediate a higher fraction of the liquidating fund's order flow.[12] Standard errors are clustered at the broker level. In the Internet Appendix, we report analogous results with clustering at the broker×stock and the broker×day levels (Table IA.VI).

Table II, Panel A presents the results. Columns (1) and (2) focus on *Net Predation*, that is, the difference between the predation and the liquidity provision dummies, while columns (3) and (4) report results for the volume-weighted version of the dependent variable. Columns (1) and (3) also include manager and broker fixed effects to ensure that our estimates are not driven by unobservable broker or manager characteristics. Columns (2) and (4) also include day×stock fixed effects, which help rule out two alternative explanations. First, asset managers might sell a stock due to stock-specific public news. Day×stock fixed effects would capture these potentially important confounding factors. Second, predation might be driven by information about prices and trades, rather than by information leakage. For instance, the price impact and abnormally high volume resulting from liquidations are publicly observable, and hence information that asset managers can use to spot trading opportunities without relying on brokers.

We find that trades executed by aware brokers have between 11% and 20% higher probability of predation relative to liquidity provision.[13] The results based on volume in columns (3) and (4) confirm this finding.

[12] The subset of brokers classified as aware during our sample period corresponds to roughly 10% of the brokers present in Ancerno.

[13] In Internet Appendix Table IA.IX, we report results from specifications without the fixed effects. In these specifications, the constant, that is, the level for unaware brokers, is virtually zero, while the slope on the aware dummy is 23%. Hence, the economic magnitude is substantial. The Internet Appendix is available in the online version of this article on the Journal of Finance website.

In Panels B and C, we separately study the relation between broker awareness and the probability of predation and liquidity provision, respectively. While we find evidence of a significant positive relation between broker awareness and predation, the effect on liquidity provision is not clear and is positive and significant only in the specifications in which liquidity provision volume is the dependent variable. Even there, the effect is smaller than for predatory volume.

Another way to shed light on predation versus liquidity provision is to compare the cumulative order imbalance from the start of the fire sale, where the cumulative order imbalance is the difference between buy and sell trades divided by the sum of the two. Figure 3 plots this series for aware and unaware brokers with standard error bands during the fire-sale events. In line with the regression evidence reported in Table II, the imbalances through the aware brokers are negative during the first several days of a liquidation event and are significantly lower than those through the unaware brokers.

Overall, the results in this section show that those brokers who are more likely to realize that a fund is engaged in a large liquidation are also more likely to intermediate trades that are consistent with predatory trading. In contrast, evidence that aware brokers facilitate liquidity provision is limited at best.

3.3 Best Clients and Predatory Trading

To sharpen our identification, we test another implication of our information leakage hypothesis. If the aware brokers leak information about order flow from liquidating managers, and if these brokers' information rents can be dissipated by leaking to too many traders, then we should see these brokers selectively disclosing order flow information in an effort to maximize their rents. In particular, we should see aware brokers favoring their best clients.

We construct two proxies for the strength of the manager-broker relationship, that is, best client proxies, based on both the volume and the commissions generated by manager m with broker b over a window of six months ending one month before the fire-sale event. The first variable is defined as the volume generated by the client as a fraction of the total volume intermediated by the broker, expressed in decimal units.

The second measure is computed similarly, but we replace the dollar volume by the dollar trading commissions generated by the manager. Summary statistics reported in Panel D of Table I show that these variables are highly persistent, with an autocorrelation of 90% at the monthly frequency. This finding suggests that brokers are incentivized to nurture these relationships over time and that heterogeneity across clients of the a given broker is a relevant source of variation for identifying the effect of interest.

To provide evidence on the role of broker-client relationships in information diffusion, we estimate

$$Net\ Predation_{m,j,b,t,e} = \beta_1 Best\ Client_{m,b,t} \times Liquidation\ Period_{t,e} + \beta_2 Best\ Client_{m,b,t} + \beta_3 Liquidation\ Period_{t,e} + \varepsilon_{m,j,b,t,e}, \quad (3)$$

where j indexes stocks, b brokers, e events, m managers, and t days, as above. In equation (3) our dependent variable of interest is the difference between the probability of predation and the probability of liquidity provision. As in Table II, we also use a version of this variable that is multiplied by the volume of trade. The dummy *Liquidation period* indicates the first five days of a fire sale, that is, the period of the most intense liquidation activity by a distressed fund. The reference period is the time before the beginning of the fire sale. All specifications include time (at the monthly frequency), manager, event, stock, and broker fixed effects. We conservatively double-cluster standard errors at the stock and the manager level, which allows for arbitrary correlation within trades in the same stock by the same manager.[14]

Table III presents the results. We find that asset managers exhibiting a closer relationship with a fire-sale-aware broker are significantly more likely to sell their holdings of the fire-sale stock with the same broker during the liquidation period. This result is economically significant, with the net probability of predation for the best

[14] In robustness tests reported in Panel B of Table IA.VI, we cluster standard errors along alternative multiple dimensions: event, stock, and manager; event, stock, and day; event, stock, and broker level. In each case the results remain significant.

clients of aware brokers (top decile) more than double that of small clients of aware brokers (bottom decile).[15]

We next employ a more stringent identification strategy that exploits variation across both managers and brokers. More specifically, we compare the difference between the behavior of the best clients of the aware brokers and the behavior of the best clients of the unaware brokers, relative to the behavior of the non-best clients of both types of brokers. Formally, we run the following regression:

$$\begin{aligned} Net\ Predation_{m,j,b,t,e} = {} & \beta_1 Best\ Client_{m,t} \times Aware_{j,b,e} \times LiquidationPeriod_{t,e} \\ & + \beta_2 Best\ Client_{m,t} \times Aware_{j,b,e} \\ & + \beta_3 Best\ Client_{m,t} \times Liquidation\ Period_{t,e} \\ & + \beta_4 Aware_{j,b,e} \times Liquidation\ Period_t \\ & + \beta_5 Best\ Client_{m,t} + \beta_6 Liquidation\ Period_{t,e} + \beta_7 Aware_{j,b,e} + \varepsilon_{m,j,b,t,e} \end{aligned} \tag{4}$$

In this specification, we define $Aware_{j,b,e}$ at the event-broker-stock level by collapsing awareness on the time dimension, that is, for each broker *b* that eventually becomes aware of the fire-sale event *e* on a stock *j,* we assign $Aware_{j,b,e} = 1$. Table III, Panel B reports the results based on both best client proxies. We continue to find that clients with stronger ties to the aware brokers are significantly more likely to sell the stock involved in the liquidation than the best clients of the other brokers involved in the liquidation.

A question that arises at this point is whether there is significant persistence in the set of asset managers that predate and in the set that get predated. We find that more than 60% of predation victims were predated upon only once, that is, the median is one, while on average predation victims are predated upon 3.13 times. These results suggest that the liquidations we are focusing on are unlikely to happen frequently enough for the affected funds to become aware of such and potentially punish the broker. In fact, even among those funds that are predated upon more than two times,

[15] This conclusion is based on the estimates from the regressions without fixed effects in Table IA.X, Panel B. In particular, we find that the net probability of predation increases from 1% in the bottom decile to 2.1% in the top decile of best clients.

the average time between two consecutive events is 2.86 years. It would thus be difficult for a manager to learn about a brokers' leakage, given that, from the perspective of a manager, predation happens rarely and inference is noisy.[16]

We further find that across all of the predators in our sample, 30% predate on more than 10 events during our sample period. Thus, predatory behavior is persistent: conditional on having predated at time *t*, the manager is twice as likely to predate again in *t*+1. Figure IA.3 in the Internet Appendix shows that this result holds for different time horizons. The evidence therefore suggests that, consistent with our hypothesis, brokers leak their information to hose clients that are most likely to take advantage of this information and generate rents for the broker.

An additional dimension that we explore is *when* predators start trading in the same direction. Intuitively, if predation starts on the first day of liquidation, it is potentially much more harmful than if it starts on the last day of the liquidation. We address this question in Internet Appendix Table IA.XII. We find that the best clients of aware brokers begin predating significantly earlier than other predators. In particular, the average predator starts predating on the third day of the liquidation, while the best clients of aware brokers start on the second day, on average. This result suggests that the best clients of aware brokers are rewarded through early access to information.

We also test whether brokers show their best clients preferential treatment when these clients need to liquidate. In Table IA.X, we find that predation is less likely when the fund in distress is one of the broker's best clients. We do not find evidence of significantly more liquidity provision, but the result holds when we look at the difference between predatory and liquidity provision volumes. Overall, clients with closer ties to the brokers appear to enjoy an advantage when they need to liquidate.

Finally, using reconstructed portfolios, we find that predators appear to short fire-sale stocks in 43% of cases. We note, however, that this estimate is subject to large measurement error due to the fact that we approximate the true portfolio. From a theoretical perspective, we do not have a strong prior as to whether predation should occur with stocks already in the predator's portfolio or stocks that the predator needs

[16] We also find that, among the funds involved in a fire sale, 40% also acted as a predator at least once.

to short. Empirically, given that the stocks that are most likely to be predated tend to be the largest and most liquid stocks in the market, it is more likely that these stocks are already in predators' portfolios. This could explain why a slight majority of predatory trades consists of sales of existing positions.[17]

3.4 Robustness to Aggregate and Stock-Specific News

Having established that the best clients of the aware brokers are more likely to sell the same stock as the distressed fund during the liquidation period, we examine whether this result can be explained by factors other than information leakage by the broker. The main alternative hypothesis for this result is that asset managers are responding to the same common shock that occurs during the same event window. This might be the case for two reasons. First, a common disruption in the market that may lead funds to offload their positions. Second, news about the specific stocks might be released, that triggers funds' trading behavior.

As discussed above, some of our prior evidence (e.g., the fact that we control for stock×day fixed effects in Table II) already helps rule out these hypotheses. Nonetheless, we construct several additional tests to rule out these alternative explanations. To ensure that the correlation among traders is not due to general disruption in the market, we first exclude the two recessions from our sample, that is the tech crunch of 2001 and the financial crisis of 2007 to 2009. Panel A of Table IA.III presents the results. We find that our main results are robust to this change in estimation sample, with both the economic and the statistical significance being unaffected.

Next, we examine whether negative stock-specific news might explain our baseline results. To do so, we collect information about earnings announcements and changes in analyst recommendations. Intuitively, earning announcements might serve as a

[17] In Table IA.II in the Internet Appendix we examine the characteristics of the stocks that are subject to greater predation. We split the sample of fire-sale stocks by the median amount of predation. This quantity is the number of manager-days in which a client of an aware broker trades in the same direction as the liquidating fund. Then, for different variables, we compute the average for stocks that are liquidated in events above the median (More Predation) and below the median (Less Predation). Overall, we find that the events with stronger predatory activity involve larger, more liquid, and less volatile stocks.

catalyst, with a negative surprise triggering a series of liquidations. We therefore exclude the 10 trading days around such announcements. Similarly, one might expect multiple liquidations to follow a downgrade, especially if it is unexpected. We therefore also exclude these events from our sample. Of course, earnings announcements and analyst recommendations are not the only news events that might trigger a coordinated response from market participants. We obtain comprehensive information about stock-specific news from Ravenpack, which collects all types of information about a stock, from lawsuit to mergers and acquisitions from newswires. A machine learning algorithm is then employed to classify news as good or bad on a scale from 0 to 100, where 50 is the cutoff below which news is identified as bad. In Panel B of Table IA.III we show that even in the restricted sample excluding bad news, the best clients of aware brokers are more likely to predate on liquidating managers.

Fund managers might also find themselves trading in the same direction when the stocks belong to a strategy, for example momentum, that is adopted by multiple funds. Furthermore, asset managers might be liquidating underperforming stocks. Thus, as an additional robustness check, in Panel C of Table IA.III we exclude from the sample all stocks exhibiting negative momentum. Specifically, we compute the returns of the stocks sold during a fire sale and exclude those with negative returns in the week preceding the fire sale. The results are unaffected.

To check whether our results could be driven by changes in investors' expectations about the stocks, in Panel D of Table IA.III we also consider short-selling data from Markit (formerly DataEx). Intuitively, stocks with high short interest might be subject to correlated sales across funds, which could be triggered by company-specific events or investors' common beliefs about stock performance rather than by a desire to take advantage of a liquidating fund. We find that our results are robust to excluding events for which the liquidated stocks exhibit a significant level of short interest, defined as the utilization ratio (i.e., shares on loan divided by shares available to lend) of the top quartile.

As an additional test to rule out the alternative hypothesis that funds are responding to similar shocks rather than deliberately taking advantage of a fire sale, we examine

the number of stocks affected by the predatory behavior of aware brokers' clients. The idea is that if investors are simply responding to a common shock to a stock, their sales should be concentrated on that stock, whereas if multiple stocks out of those that are involved in a fire sale are sold by the best clients of aware brokers, predation on the liquidating fund is more likely.

To test this conjecture, Table IV reports results where the outcome variable in columns (1) and (2) is the number of fire-sale stocks that the manager sells and in columns (3) and (4) the fraction of fire-sale stocks in which we observe predatory behavior. We find that the top-decile clients of aware brokers tend to sell around eight more stocks than bottom-decile clients do (column (1)) and to predate about 33% more of the stocks involved in the fire sale (column (3)).[18] These results are consistent with aware brokers selectively leaking information to their best clients.

3.5 Evidence of Trade Reversion

To further support the view that our results are driven by predatory behavior by asset managers who acquire order flow information from an aware broker, we examine whether these same asset managers are also likely to cover their positions by repurchasing the stock in the following days.

To do so, we compute the fraction of a manager's negative position that is subsequently reversed. Specifically, we define the percentage of manager m's position in stock j that is reversed during event e as $Rev_{e,m,j} = BoughtBack_{e,m,j} \,/\, Sold_{e,m,j}$, where $Sold_{e,m,j}$ is the dollar sum of all sell orders in the period and $BoughtBack_{e,m,j}$ is the dollar sum of buy orders during the period, where we sum only over the buy orders that are preceded by negative cumulative order flow to avoid counting as reversals those buy orders that occur before sales have taken place. We compute this measure around the 10 days before and after each fire-sale event. We then compare

[18] The bottom decile of the best client proxy based on volume is 0, while that for the top decile is 0.58 (median point between the 90th percentile and the max of the distribution). Hence, we obtain an increase of 14.5 × 0.58 = 8.4 stocks, while the increase in the fraction of predated stocks is 57.8 × 0.58 = 33.5%.

the fraction of the position that is reversed across clients of aware brokers before and after fire-sale events. The liquidating funds are excluded from the sample.

Table V reports the results. We find that a significant fraction of predating managers' positions is covered in the 10 days following the fire sale. We interpret this result as strong evidence that predating managers are motivated by the prospect of short-term gains at the expense of the liquidating fund.[19]

3.6 Late-Trading Scandal as a Natural Experiment

Yet another possible explanation for our results is that the intermediating broker is the original source of the information about the liquidated stocks, that is, the broker triggers both the large sale as well as smaller sales by other managers in the same direction.

To rule out this possibility, we need to identify an exogenous determinant of fire sales. In particular, we need to identify a determinant of liquidations that is manager-specific, that is not induced by the broker, and that does not depend on the identity of the liquidated stocks or the composition of the manager's portfolio.

Anton and Polk (2014) use the liquidations triggered by outflows following the late-trading scandal as a natural experiment to identify exogenous selling activity (see also Kisin (2011)). Following these authors, we exploit the mutual fund scandal that erupted in September 2003. At the time, New York Attorney General Eliot Spitzer announced the discovery of illegal late-trading activities and market-timing practices on the part of several hedge fund and mutual fund companies. The scandal had a significant impact on the 27 fund families involved: they experienced significant

[19] To give a sense of the magnitude of the trade reversal, we compare this unwinding activity to the reversal of sell trades by the same predatory managers over a random sample of five-day intervals that do not include a fire sale (placebo sample). Figure IA.4 in the Internet Appendix compares the unwinding of predators' trades after the fire sale to trades on placebo days, where predators are managers trading in the same direction as the fire sale and placebo days are days in which no fire sale takes place. It is evident that reversals are significantly higher after the fire sales. In particular, only one day after a fire sale, 30% of sell positions are bought back, while this number is closer to 3% in the placebo sample. After one month the reversal plateaus at about 50% of the positions, while it is below 25% in the placebo sample. We conclude that the evidence of trade reversal after fire sales is economically significant. The fact that not all sell trades are reversed suggests that either some investors already intended to sell the stock and took advantage of the price decline to do so, or some investors mistook the price drop for a negative signal on the stock and decided to drop it from their portfolio.

outflows, losing 14.1% of their capital within one year and 24.3% within two years (Kisin, (2011)). This setting is ideal for our purposes because it allows us to identify stocks subject to selling pressure for exogenous reasons. Although market participants were aware that these fund families were experiencing investor outflows, brokers' unique vantage point allowed them to pin down *when* these funds were liquidating and *which* stocks were involved in the liquidation. Both pieces of information which are not publicly available, are crucial to making predation profitable.

To test whether our main result, namely, that brokers leak information about stocks that are liquidating and the timing of these liquidations, extends to this setting, we manually match the identity of the fund families included in Spitzer's complaint with our trade-level data to identify the sell trades of these fund families and the brokers through which they executed them.[20] Corroborating the validity of our matching procedure, we find that the matched managers rank in the top quartile by sales in the two-year period following the scandal.

Next, we focus on the daily transactions of the managers not involved in the scandal for the four-years period centered on the month of the announcement of Spitzer's complaint (September 2003) and construct the dummy $Post\ Scandal_t$, which indicates the two years after the complaint broke out. We then define the broker-stock-day-level dummy variable $Selling_{b,j,t}$ which indicates whether at least one of the charged funds sold stock j on day t through broker b. The dependent variable, *Probability of Predation*, is a dummy variable that equals to one if a non-charged manager sold stock j on day t through broker b, and to zero if a non-charged manager traded on a different day, on a different stock, or with a different broker. In a difference-in-differences setting, we regress *Probability of Predation* on the interaction between $Selling_{b,j,t}$ and the $Post\ Scandal_t$ dummy.

[20] A complete list of the fund families involved in the scandal arising from Spitzer's complaint can be found at https://en.wikipedia.org/wiki/2003_mutual_fund_scandal#List_of_implicated_fund_companies.5B4.5D.5B5.5D. Out of the 27 families that were involved, we are able to find a match in our data set for 19 of them. These 19 managers were responsible for seven of the 31 fire-sale events in the two-year period after the scandal broke out, that is, the late-trading scandal families generated 23% of the fire sales. Importantly, the implicated funds represent only about 2.1% of the managers in the database (i.e., 19/900). Hence, the implicated families weigh about 10 times more heavily than the other managers in generating fire sales in those two years.

Table VI reports the results. Consistent with our main results, we find that the clients of the brokers employed by the funds involved in the scandal were significantly more likely to liquidate the same stocks after the scandal broke out. For example, in column (1) we find a 6.2% higher probability of non-charged managers trading in the same direction as a charged manager on the same day through the same broker relative to the pre-scandal period (i.e., 8.7% - 2.5%).

In sum, there results suggest that our main findings are not due to and that these results cannot be explained away by shocks to the market to the single stocks, nor are they a common response to the release of public information, given that the timing of the sales and the identity of the stocks sold is information that only the intermediating brokers would have access to. Moreover, the idea that brokers generate stock-specific trading ideas seems implausible, as there is no reason for such activity to increase after late-trading scandal broke out or for stocks of the implicated funds to be liquidated. These results therefore lend further support to the view that the clients of aware brokers adopt predatory trading strategies to take advantage of temporary price movements due to fire sales.

3.7 Heterogeneity

The most active managers in the sample are likely to be most able to take advantage of liquidating funds' trades. To test this conjecture, we examine whether the results differ for hedge funds and other institutions. Intuitively, compared to mutual funds or pension funds, hedge funds are likely to be able to react more promptly to information released by brokers. We manually identify the hedge funds in Ancerno following the procedure in Franzoni and Plazzi (2015).

Table IA.IV reports the results of regressing equation (3) for hedge funds and other institutions. The results clearly show that hedge funds are the main culprits behind predation, as both statistical significance and economic significance are weaker for non-hedge funds. This evidence supports the view that the behavior we observe is a deliberate attempt by the "smart money" to take advantage of temporary price fluctuations.

To examine whether predation is even more prominent in periods of financial distress, we split our sample based on the value of the VIX during each fire-sale event. The results, reported in Table IA.VII, show that predation through aware brokers is somewhat stronger during periods of financial distress. This finding may be driven by the fact that liquidations are more significant during times of market stress. Additionally, the price impact of liquidations is larger because the market is more illiquid, which creates additional room for predators to profit from liquidation.

4 The Value of Order Flow Information

4.1 Profitability of Predatory Strategies

An important question at this point is whether the asset managers that receive leaked information from aware brokers are able to generate higher abnormal returns.

To address this question, we compute the profits that asset managers generate during fire sales. In particular, starting from the first day of the liquidation (day 0), at the close of each day we compute the marked-to-market value of the net position in a given stock and subtract from this value the net cash expenditure necessary to build that position over the period. To express these profits as a fraction of capital at risk, we divide them by the absolute value maximum dollar outlay over the period in which the profits are computed.[21]

We start by plotting in the left panel of Figure 4 the profits of the best clients (those generating more than 5% of the volume intermediated by the broker in the previous semester) of aware and unaware brokers at the daily frequency after the start of the fire sale. Intuitively, if as shown in Table II the trades executed by unaware brokers are significantly less likely to be predatory, we should find that their clients are less likely to profit from these fire-sale events. In line with this view, the figure shows that the clients of aware brokers are able to capture significant returns after the start of the

[21] To be clear, we subtract stocks that are sold from stocks that are bought to compute the net position, which can end up being negative, as in a short sale. The net cash expenditure to build the position can also be negative if the dollar value of the sell transactions exceeds that of the buy transactions. This observation implies that we need to use the absolute value when we compute the maximum exposure.

liquidation, while the trades of unaware brokers' clients do not generate significant profits. The profits of the best clients peak at about 25 bps on the ninth day after the start of the fire sale. Importantly, most of the profits are generated in the first five days, that is, while the prices of the stocks being liquidated are still decreasing and it makes sense to predate. The profits generated later in the period are instead consistent with liquidity provision.

To provide more systematic regression evidence, we estimate

$$Profits_{m,j,b,t} = \beta_1 Best\ Client_m \times Post[0,9]_t + \beta_2 Best\ Client_m + \beta_3 Post[0,9]_t + \varepsilon_{m,j,b,t}, \quad (5)$$

which indicates whether manager *m*'s profits are significantly higher in the 10 days after the start of a fire sale relative to the 10 days before the fire sale, as a function of clients' proximity to aware brokers. Intuitively, as with the estimation of the predation probability, we are comparing the behavior of managers expected to be aware of the fire sale, given their relationship with the broker, with those not likely to be aware of the fire sale, before and after the beginning of the fire sale. We choose a 10-day window to allow managers time to close the predatory short positions that they likely accumulate during the first five days of the fire sale, the period over which the price of an affected stock declines (see Figure 2).[22]

Table VII reports the results. We find that aware brokers' best clients exhibit significantly higher profits than other managers during the period under consideration. In particular, in the 10 days following the beginning of a fire sale, clients in the top decile of our relationship metric based on trading volume, outperform by more than 50 bps on average relative to managers in the bottom decile trading on the same stocks in the same period (i.e., (136.56 - 48.57) × 0.58 = 51 bps, where 0.58 is the value of the best client proxy in the top decile, while it is zero in the bottom decile). Considering the low average performance of institutional asset managers (see, among others, Busse, Goyal, and Wahal (2010)) these returns are indeed highly economically significant.

[22] Of course, the positions could be closed before day 5 and still be profitable. Our methodology for computing profits is flexible enough to allow for all such possibilities.

One might wonder whether the clients of aware brokers are always able to generate higher profits than the clients of unaware brokers. Although we already control for manager fixed effects, we test for this possibility directly in the right panel of Figure 4, which provides a placebo test for the left panel of the figure. This panel plots the profits of the two groups of managers, but for a random sample of event windows other than those included in our fire-sale analysis. We find that the two groups are indistinguishable in terms of performance during these other times.

We can provide more details regarding the profitability of managers classified as predators, that is, clients of aware brokers that trade in the same direction as the liquidating fund during the five days of a fire sales. Across all predated stocks, the average predatory position of a predator during a fire-sale event is $10 million (median $6 million). The profits arising from these positions total on average $280,449 per event-manager (median $126,420).[23] These predatory profits correspond to 2.1 bps of the portfolio value (median 0.9 bps). This percentage profit is generated over 10 trading days on a predatory trade involving about four stocks on average. Thus, it seems a significant source of returns, given the limited amount of capital required to carry it out.[24]

We can also quantify the commissions generated by predators' trades. On average, aware brokers earn $40,407 per fire-sale event, considering only predatory trades by their best clients on the fire-sale stocks, which corresponds to roughly 50% of the total commission these brokers earn on those stocks during the liquidation period (excluding those generated by the liquidating funds). That said, we expect most of brokers' benefits to originate from the future business that the enhanced closeness with tipped predators brings about. Indeed, in Section III-D we show that in the two

[23] The estimate of 32 bps for additional profits during the predation period (Table VII) is computed by averaging across all best clients of the aware brokers (to avoid hardwiring the result). Additionally, it results from difference-in-difference regressions that include fixed effects. This explains why our estimate here is lower than the estimate of 280 bps (= $280,449/ $10 million) that we obtain for the average profits of predators only.

[24] Because we compute profits using transaction prices, the price impact component of transaction costs is already taken into account. To account for the explicit component of transaction costs, we can use the estimates provided in Ross et al. (2017) based on AQR's internal data. For US equity trades they report that commissions, fees, and taxes erode about 0.3 bps of the notional per trade. Therefore, assuming the trades involve roundtrip transactions, accounting for explicit costs would reduce our estimated average trade-level profit in Table VII (column (1)) from 32 bps to 31.4 bps for the best clients of aware brokers. For the subset of predators, profits as a fraction of the value of the open position would decrease from 280 bps to 279.4 bps.

years following a fire-sale event, predators pay higher commissions per dollar traded to the tipping broker than do other clients.

4.2 Price Impact

Having established that predatory traders are able to capture significant returns, we now investigate the dark side of predation. We conjecture that predatory volume causes stock prices to decline significantly more than they would in the absence of predation, and that this steeper decline in prices leads the liquidating fund to achieve lower returns on its sale trades.

Testing this conjecture requires specification of a counterfactual. Fortunately, we are able to identify 29 fire-sale events (out of the 385, or 7.5%) for which there are no aware brokers. In these situations, no broker observes a large enough fraction of the liquidation to be considered aware according to the criteria specified in Section I. According to our identification strategy, no information leakage occurs around these events. More realistically, information leakage occurs but is expected to be significantly less pronounced.

Based on this strategy, we run regressions of price impact on the broker awareness dummy. In this case, the broker awareness dummy indicates sock-events in which there is at least one aware broker. The price impact is given as the execution shortfall, that is, the volume-weighted percentage difference between the execution price and a benchmark price (e.g., Keim and Madhavan (1997)).

We use three benchmarks to show that our results do not crucially depend on the choice of measure. Specifically, we use the price at the time the first fire-sale trade is placed, the open price on the day of the first fire-sale trade, and the transaction price of the first fire-sale trade. In all specifications, we control for the volume in the fire sale, the volume of the following trades (i.e., the trades in the same direction over the same five-day window), and the liquidity of the stock (Amihud's (2002) illiquidity ratio), as these factors potentially contribute to the price impact. In more detail, for each benchmark price we compute the implementation shortfall at the ticket level for sales by the liquidating funds during the liquidation period as

$$\frac{TransactionPrice - BenchmarkPrice}{BenchmarkPrice}. \quad (6)$$

We average this quantity at the event-stock-broker level, using as weights the volume of each transaction, to obtain an event-stock-broker level measure of price impact. We then further collapse on the broker dimension to study the price impact at the event-stock level.

Table VIII reports the results. In specifications (1) to (3), we regress the event-stock-level price impact measures on the dummy indicating events associated with at least one aware broker. We find that the price impact costs borne by the liquidating funds are higher when at least one broker is aware of the liquidation event, with this result statistically significant in two of the three cases. This finding is also economically significant, as the price impact increases by 14 bps to 36 bps, which amounts to a two-fold increase in the baseline average price impact. In specifications (4) to (6), we exploit the granularity of our data and run a similar specification in an event-stock-broker-level sample. In this case, we can have aware and unaware brokers for the same stock-event. We then include broker fixed effects to control for the possibility that heterogeneity in price impact results from differences in broker execution quality. The results remain significant and the magnitude decreases only slightly.[25]

We also address the concern that our findings are driven by expectations of a large price impact affecting awareness, rather than the opposite. The literature on block trading in equities (e.g., Seppi (1990)) suggests that this alternative interpretation might be due to the fact that brokers generally frown upon clients breaking up a large order if this large order is likely to have a meaningful price impact. The liquidating manager of a large order may therefore choose to intermediate through only a few brokers if he anticipates a price impact. We can examine this question empirically by adding the number of brokers used by the liquidating fund to our main specification.

[25] To further assess magnitudes, we can look at Ross et al. (2017), who employ AQR internal data. These authors document a price-impact for a long-only momentum fund with USD 1.6B under management as of 2016. In particular, they report that during 2009 to 2016 market impact was in the range of 6 to 11 bps per dollar traded. By comparison, in our sample of liquidations, the average execution shortfall is 52 bps per dollar traded by the liquidating funds (median 40 bps). This result is far from the typical price-impact range. One might also be concerned whether the liquidations are too small to attract arbitrage capital. However, Table IV shows that predators are able to trade multiple stocks from a given liquidation event, which through diversification increases the Sharpe ratio of this type of trade.

With the inclusion of this control, identification arises from comparing aware and unaware brokers *conditional* on the number of brokers used in the liquidation. This specification alleviates the concern that broker awareness is an endogenous variable, where the endogeneity emerges because liquidating funds choose to use fewer brokers when they expect a higher price impact. On the contrary, our results show a positive (although not statistically significant) correlation between price impact and the number of brokers used in the liquidation.

Finally, we provide graphical evidence on the difference in price paths between aware brokers and unware brokers. In Figure 5 we plot the cumulative return of fire-sale stocks during fire-sale events. The red line with squares represents the cumulative return averaged across these stock-events for the aware brokers, while the green line with circles is an estimation of the counterfactual cumulative return, based on unaware brokers. The series draw on estimates from a regression similar to that reported in column (3) of Table VIII, but run on daily observations starting on day 0. The vertical distance between the two series thus gives the estimate of the aware broker dummy for a specific day of the fire-sale event.

Figure 5 shows that the liquidating funds' transaction costs increase significantly in the presence of predatory trading. In particular, at the trough of the price impact, that is, the fifth day of the fire sale, the cumulative return is about -105 bps when trading thorugh aware brokers and -75 bps in the case of unaware brokers, that is, in the case in which we conjecture that no leakage occurs, which corresponds to a 40% increase. These results thus speak to the role of information leakage in exacerbating fire sales.[26]

[26] In Internet Appendix Table IA.XI, using the full Ancerno sample, we compute the characteristics of managers that fire-sale eventtrade with aware/unaware brokers during fire-sale events. The results show that managers trading with aware brokers during a fire sale generate more trading volume (row 1). Accordingly, they have more with broker relationships (row 2). However, they use significantly fewer brokers per million dollars that they trade (row 3). This fact suggests that the managers that face predation (i.e., those that turn to aware brokers) are less in the habit of splitting their volume across multiple brokers. We also look at the commissions generated by the managers. Given that they trade more, the managers that deal with aware brokers generally pay more commissions (row 4). In terms of the commissions they pay to different brokers, we find that the first category of managers pay more commissions to aware brokers than to unaware brokers, even outside of fire-sale events, both per dollar traded (row 5) and in total dollar amounts (row 6). Using Ancerno, Goldstein et al. (2009) classify broker-manager relationships into premium and discount. The evidence in the table suggests that managers that are predated are more likely to have a premium relationship with aware brokers. Overall, it appears that managers that end up being predated have more focused brokers relationships. In particular, they appear to be premium clients who entrust the aware brokers with a larger fraction of their traders. These predated clients are possibly trading off the risk of being predated against the advantages in terms of information generation, IPO allocation, etc., which originate from being a premium customer of a broker (Goldstein et al. (2009)).

4.3 Persistence in Broker-Manager Relationships

One may wonder why liquidating funds do not hide their trades better to avoid the higher price impact. Several non-mutually exclusive explanations are possible. First, the evidence suggests that in fact they do try to hide their trades, as on average they employ 29 brokers to intermediate these trades. Second, these funds are likely in a rush to liquidate, in which case they prioritize execution speed over price impact. For the same reason, they are likely to rely on familiar brokers other than to search for other brokers, which can take time. Third, there is a significant amount of stickiness in the trading relationships between brokers and their clients. The autocorrelation of our relationship measures at the monthly frequency is above 90%. Table IA.XIII reports the autocorrelation of various measures of concentration, such as the number of brokers and the Herfindahl index, both on average and during fire-sale events. The relevant finding is that indeed there is significant persistence in the concentration of asset managers' trades among brokers. This result holds both when managers are seeking liquidity, that is during fire sales, and when they are not. It appears, therefore, that managers find it difficult to start interacting with new brokers, that is building new broker relationships, when timely execution of their trades is needed to meet investors' redemption demands.

4.4 Quid Pro Quo

Another natural question that may arise is whether brokers gain from leaking order flow information about their clients. On the other hand, it may be in their best interest to build a reputation as a loyal trading partner by keeping order flow information private. On the other hand, brokers have an incentive to increase the volume they intermediate as they are paid on commissions. We address this question by exploiting the granularity of our data to test whether best clients tend to reward aware brokers by channeling more trades to them. In Table IX we regress the average $Commission\ per\ dollar_{e,m,b,}$ paid by manager m to broker b during month t, calculated as the ratio of the total dollars paid in commissions and the total dollar

volume traded by manager *m* and intermediated by broker *b* in that month, on the interactions between the dummy variable identifying the two years following the fire-sale event and each of our best client proxies. We find that the clients that are more likely to receive order flow information tend to increase their commissions to the brokers, which strongly suggests a quid pro quo between these parties.[27]

5 Conclusion

In this paper we study whether brokers' incentives to attract and retain clients induce sharing of order flow information with other market participants. The evidence suggests that brokers tend to reveal the occurrence of a fire sale to their best clients, which allows them to generate significant profits by predating on the liquidating fund. This information leakage leads to a higher price impact and hence to a more costly liquidation for the fire-sale originator.

These findings have implications for academics, practitioners, and policy makers alike. First, our results highlight an important cost associated with slow execution. Slow execution has been widely advocated since Kyle (1985) as a way to minimize price impact and is routinely implemented by practitioners. However, according to our results, executing large trades over multiple days allows the intermediary brokers to anticipate order flow share this information with other market participants, triggering predatory behavior. This might adversely affect liquidating managers' price impact.

Information leakage might be a concern to regulators as well, since it can exacerbate the costs associated with fire sales, especially during times of scarce liquidity. As Fox, Glosten, and Rauterberg (2015) point out, brokers have a legal duty not to use knowledge of a client's order to their own advantage. A broker that trades on its own behalf before the client's order reaches the market clearly violates such duty.[28] Things

[27] To get a sense of magnitudes, after a fire-sale event a top-decile client (for which the best client proxy based on volume is 0.58) fire-sale eventpays 16% of a standard deviation higher commissions per dollar relative to smaller clients. In Table IA.XIV we show that this estimate increases to 24% when we restrict attention to managers that generated the highest profits during the event window.

[28] This situation describes front running, which is prohibited under common law (e.g., Opper v. Hancock), federal law (Section 10(b) of the Exchange Act and Rule 10b-(5)), and industry self-regulatory standards (FINRA Rule 570(a)).

are less clear, however, if a broker receives an indirect benefit from leaking order flow information to a third party. In theory, the broker is in violation of the aforementioned legal duty.[29] But brokers can argue that information leakage occurs while respecting their fiduciary duty of best execution vis-à-vis their clients. In particular, they can argue that information may leak while searching for counterparties for a large trade. Irrespective of the intentions of the broker, the evidence in this paper shows that liquidity provision is lower among aware brokers than other brokers. A regulatory attempt to stop such information leakage is likely to be challenging, however, as it would have to account for brokers' need to operate as deal-makers, as well as face the reluctance of many asset managers to disclose more information about their trading activities.

Our results shed light on a recent debate over exchanges' and brokers' use of their access to market information to sell data products.[30] Critics maintain that institutional investors, which routinely need to execute large trades anonymously, can be negatively impacted as these products could be used to "reverse-engineer" their strategies and thereby lead to front-running. Our findings show that, even in the absence of such supplemental information, large investors whit strong business relationships with brokers are able to exploit order flow information at the expense of those seeking liquidity provision. Our estimates may serve as a benchmark – probably a lower bound – for the costs associated with releasing such data products.[31]

Future research could build on the insights of this paper towards generating a theory of how the relationship between asset managers and intermediaries (such as brokers) affects trading behavior and asset prices. Specifically, one could structurally estimate how information diffuses among market participants to address questions about the

[29] Moreover, a broker that shares information concerning a customer order to a third party is violating its agency duties of confidentiality, provisions of Regulation ATS (if the broker is an operator of an Alternative Trading System, such as a dark pool), and probably its own marketing material. See RESTATEMENT (THIRD) OF AGENCY (2006) Section 8.05(2) and Rule 301(b)(10) of Regulation ATS.

[30] The most recent dispute involves NASDAQ seeking the SEC's approval for an options data service called the "Intellicator Analytic Tool". This new service would provide subscribers market color by revealing whether a trade was initiated by a small investor or a large money manager. This story was reported in a recent WSJ article: "*Wall Street Fears Nasdaq Proposal Would Expose Trading Secrets*" (available at https://www.wsj.com/articles/could-the-intellicator-spill-the-markets-secrets-1510223403?tesla=y#comments_sector).

[31] Our results also highlight the importance of the fiduciary duty between broker-dealers and their clients. A few states in the U.S. are moving in the direction of tightening such duty for brokers. For instance, Nevada is considering an expanded interpretation of fiduciary duty whereby the brokers would be required to "disclose to a client, at the time advice is given, any gain [the broker] may receive, such as profit or commission, if the advice is followed."

efficiency of such strategic behavior by brokers for price discovery and asset allocation, as well as provide insights into counterfactual results in the presence of new regulations aimed at curbing this practice.

References

Acharya, Viral V., Douglas Gale, and Tanju Yorulmazer, 2011, Rollover risk and market freezes, *Journal of Finance* 66, 1177-1209.

Admati, Anat R., and Paul Pfleiderer, 1991, Sunshine trading and financial market equilibrium, *Review of Financial Studies* 4, 443-481.

Afonso, Gara, Anna Kovner, and Antoniette Schoar, 2013. Trading partners in the interbank lending market, FRB of New York Staff Report, 620.

Allen, Franklin, and Douglas Gale, 1994, Limited market participation and volatility of asset prices, *American Economic Review* 84, 933–55.

Amihud, Yakov, 2002, Illiquidity and stock returns: Cross-section and time-series effects, *Journal of Financial Markets* 5, 31-56.

Anand, Amber, Paul Irvine, Andy Puckett, and Kumar Venkataraman, 2012, Performance of institutional trading desks: An analysis of persistence in trading costs, *Review of Financial Studies* 25, 557-598.

Anand, Amber, Paul Irvine, Andy Puckett, and Kumar Venkataraman, 2013, Institutional trading and stock resiliency: Evidence from the 2007-2009 financial crisis, *Journal of Financial Economics* 108, 773-793.

Anton, Miguel, and Christopher Polk, 2014, Connected Stocks, *Journal of Finance* 69, 1099–1128.

Babus, Ana, and Péter Kondor, 2016, Trading and information diffusion in over-the-counter markets, Working paper, London School of Economics.

Bessembinder, Hendrik, Allen Carrion, Laura Tuttle, and Kumar Venkataraman, 2016, Liquidity, resiliency and market quality around predictable trades: Theory and evidence, Journal of Financial Economics 121, 142-166.

Biais, Bruno, and Thierry Foucault, 2014, HFT and market quality, *Bankers, Markets & Investors* 128, 5-19.

Boehmer, Ekkehart, Kingsley Y. L. Fong, and Julie Wu, 2014, International evidence on algorithmic trading, Singapore Management University, Working paper.

Brogaard, Jonathan, and Andrew Detzel, 2015, The asset-pricing implications of government economic policy uncertainty, *Journal of Management Science* 61, 3-18.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2014, High-frequency trading and price discovery, *Review of Financial Studies* 27, 2267-2306.

Brown, David B., Bruce Ian Carlin, and Miguel Sousa Lobo, 2010, Optimal portfolio liquidation with distress risk, *Journal of Management Science* 56, 1997–2014.

Brunnermeier, Markus K., and Lasse H. Pedersen, 2005, Predatory trading, *Journal of Finance* 60, 1825-1863.

Brunnermeier, Markus, and Lasse H. Pedersen, 2009, market liquidity and funding liquidity, *Review of Financial Studies* 22, 2201–38.

Busse, Jeffrey A., Amit Goyal, and Sunil Wahal, 2010, Performance and persistence in institutional investment management, *Journal of Finance* 65, 765-790.

Chaboud, Alain P., Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega, 2014, Rise of the machines: Algorithmic trading in the foreign exchange market, *Journal of Finance* 69, 2045-2084.

Coval, Joshua and Erik Stafford, 2007, Asset fire sales (and purchases) in equity markets, *Journal of Financial Economics* 86, 479-512.
Chung, Ji-Woong, and Byoung Uk Kang, 2016, Prime broker-level comovement in hedge fund returns: Information or contagion?, *Review of Financial Studies* 29, 3321-3353.
Degryse, Hans, Frank De Jongm and Vincent van Kervel, 2015, The impact of dark trading and visible fragmentation on market quality, *Review of Finance* 19, 1587-1622.
Di Maggio, Marco, 2016, Market turmoil and destabilizing speculation, Columbia Business School Research Paper No. 13-80.
Di Maggio, Marco, Amir Kermani, and Zhaogang Song, 2017, The value of trading relations in turbulent times, *Journal of Financial Economics* 124, 266-284.
Di Maggio, Marco, Francesco Franzoni, Amir Kermani, and Carlo Sommavilla, 2018, The relevance of broker networks for information diffusion in the stock market, *Journal of Financial Economics*, forthcoming.
Di Mascio, Rick, Anton Lines, and Narayan Y. Naik, 2016, Alpha decay and Strategic trading, Working Paper, Columbia Business School.
Duffie, Darrell, Semyon Malamud, and Gustavo Manso, 2009, Information percolation with equilibrium search dynamics. *Econometrica* 77, 1513-1574.
Duffie, Darrell, Semyon Malamud, and Gustavo Manso, 2014, Information percolation in segmented markets, *Journal of Economic Theory* 153, 1-32.
Eisele, Alexander, Tamara Nefedova, Gianpaolo Parise, and Kim Peijnenburg, 2017, Trading out of sight: An analysis of cross-trading in mutual fund families, Working paper, Centre for Economic Policy Research.
Ellul, Andrew, Chotibhak Jotikasthira, and Christian T. Lundblad, 2011, Regulatory pressure and fire sales in the corporate bond market, *Journal of Financial Economics* 101, 596-620.
Farboodi, Maryam, and Laura Veldkamp, 2017, Long-run growth of financial technology, National Bureau of Economic Research, paper No. w23457.
Fox, Merritt B., Lawrence R. Glosten, and Gabriel V. Rauterberg, 2015, The new stock market: Sense and nonsense, *Duke Law Journal* 65, 191.
Franzoni, Francesco, and Alberto Plazzi, 2015, Do hedge funds provide liquidity? Evidence from their trades, Working paper, Swiss Finance Institute.
Garleanu, Nicolae, and Lasse H. Pedersen, 2011, Margin-based asset pricing and deviations from the law of one price, *Review of Financial Studies* 24, 1980-2022.
Garleanu, Nicolae, and Lasse H. Pedersen, 2013, Dynamic trading with predictable returns and transaction costs, *Journal of Finance* 68, 2309-2340.
Gaspar, Jose-Miguel, Massimo Massa, and Pedro Matos, 2006, Favoritism in mutual fund families? Evidence on strategic cross-fund subsidization, *Journal of Finance* 61, 73-104.
Goldstein, Michael A., Paul Irvine, Eugene Kandel, and Zvi Wiener, 2009, Brokerage commissions and institutional trading patterns, *Review of Financial Studies* 22, 5175-5212.
Greenwood, Robin, Augustin Landier, and David Thesmar, 2015, Vulnerable banks, *Journal of Financial Economics* 115, 471-485.
Gromb, Denis, and Dimitri Vayanos, 2002, Equilibrium and welfare in markets with financially constrained arbitrageurs, *Journal of Financial Economics* 66, 361–407.
Hasbrouck, Joel, and Gideon Saar, 2013, Low-latency trading, *Journal of Financial Markets* 16, 646-679.
Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1-33.
Hendershott, Terrence, Dan Li, Dmitry Livdan, and Norman Schürhoff, 2016, Relationship trading in OTC markets, Swiss Finance Institute Working Paper.
Hollifield, Burton, Artem Neklyudov, and Chester S. Spatt, 2016, Bid-Ask spreads, trading networks and the pricing of securitizations, *Review of Financial Studies* 30, 3048-3085.

Irvine, Paul, Marc Lipson, and Andy Puckett, 2006, Tipping, Review of Financial Studies 20, 741-768.
Keim, Donald B., and Ananth Madhavan, 1997, Transactions costs and investment style: An inter-exchange analysis of institutional equity trades, *Journal of Financial Economics* 46, 265-292.
Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun, 2017, The flash crash: High-frequency trading in an electronic market, *Journal of Finance* 72, 967-998.
Kisin, Roni, 2011, The impact of mutual fund ownership on corporate investment: Evidence from a natural experiment, Working paper, Washington University in St. Louis.
Kiyotaki, Nobuhiro, and John Moore, 1997, Credit cycles, Journal of Political Economy 105, 211-248.
Kyle, Albert S., 1985, Continuous auctions and insider trading. *Econometrica* 53, 1315-1335.
Li, Dan, and Norman Schürhoff, 2018, Dealer networks, Journal of Finance 74, 91-144.
McNally, William J., Andriy Shkilko, and Brian F. Smith, 2015, Do brokers of insiders tip other clients? *Journal of Management Science* 63, 317-332.
Menkveld, Albert J., 2013, High frequency trading and the new market makers, *Journal of financial Markets* 16, 712-740.
Menkveld, Albert J., 2016, The economics of high-frequency trading: Taking stock, *Annual Review of Financial Economics* 8, 1-24.
O'Hara, Maureen, 2015, High frequency market microstructure, *Journal of Financial Economics* 116, 257-270.
Puckett, Andy, and Xuemin (Sterling) Yan, 2011, The interim trading skills of institutional investors, *Journal of Finance* 66 601-633.
Moskowitz, Tobias J, Adrienne Ross, Ronen Israel, and Laura Serban, 2017, Implementing momentum: What have we learned?, Working paper.
Scholes, Myron 2000, Crisis and risk management, *American Economic Review* 90, 17–21.
Seppi, Duane J., 1990, Equilibrium block trading and asymmetric information, *Journal of Finance*, 45, 73-94.
Shleifer, Andrei, and Robert W. Vishny, 1992, Liquidation values and debt capacity: A market equilibrium approach, *Journal of Finance* 47, 1343-1366.
Shleifer, Andrei, and Robert W. Vishny, 1997, A survey of corporate governance, Journal of Finance 52, 737-783.
Shleifer, Andrei, and Robert W. Vishny, 2011, Fire sales in finance and macroeconomics, *Journal of Economic Perspectives* 25, 29-48.
Van Kervel, Vincent, 2015, Competition for order flow with fast and slow traders, *Review of Financial Studies* 28, 2094-2127.
van Kervel, Vincent and Menkveld, Albert J., 2016. High-Frequency trading around large institutional orders, *Journal of Finance* 74, 1091-1137.
Walden, Johan, 2016, Trading, profits, and volatility in a dynamic information network model, Working paper, University of California, Berkeley.
Yang, Liyan, and Haoxiang Zhu, 2016, Back-running: Seeking and hiding fundamental information in order flows, Rotman School of Management Working paper 2583915.
Ye, Mao, Chen Yao, and Jiading Gai 2013, The externalities of high frequency trading, Working paper.

Appendix

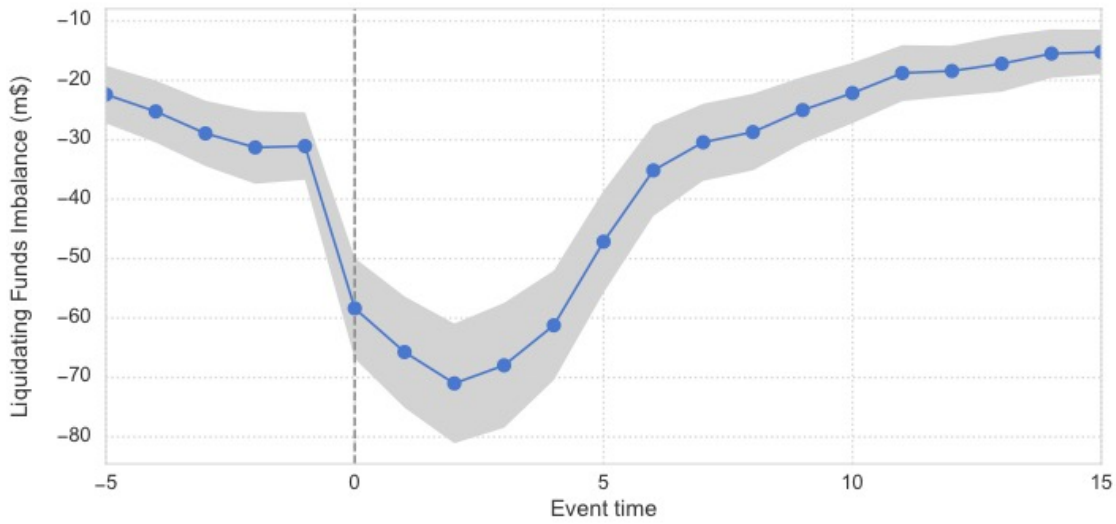


Figure 1. Liquidation volume and price pattern. The figure plots the average daily signed volume (i.e., order imbalance) of the fire-sale originator on the fire-sale stocks, expressed in million dollars.

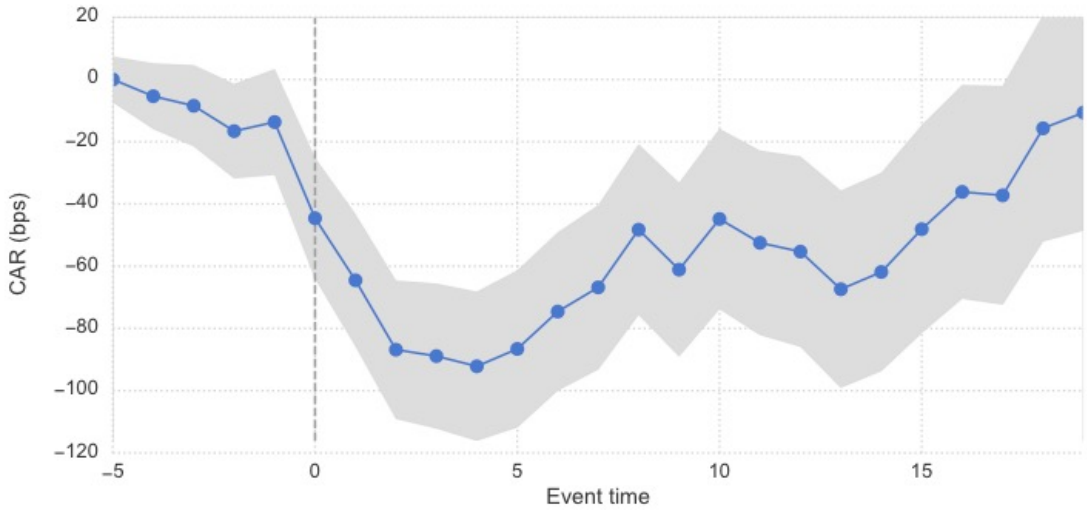


Figure 2. Price pattern. The figure plots the average DGTW-adjusted cumulative returns for the stocks sold during the fire sale along with 95% confidence bands.

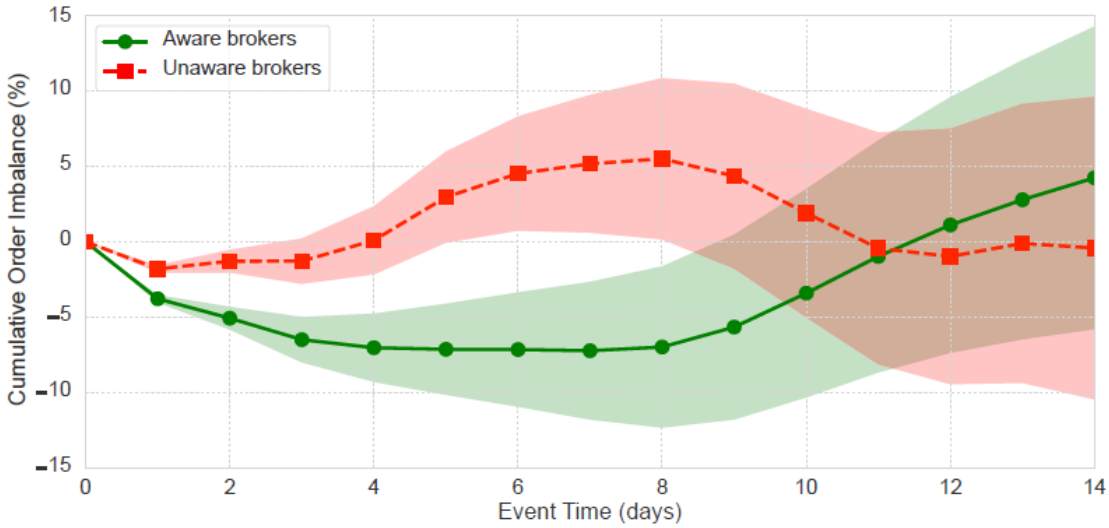


Figure 3. Order flow through aware brokers. The figure plots the cumulative order imbalance of the transactions intermediated by aware brokers (green solid line) and unaware brokers (red dashed line) on the fire-sale stocks, excluding those generated by the liquidating funds. The daily order imbalance is computed as the difference in the volume of buy and sell orders divided by the total absolute volume. The measure is then averaged across fire sales in event time.

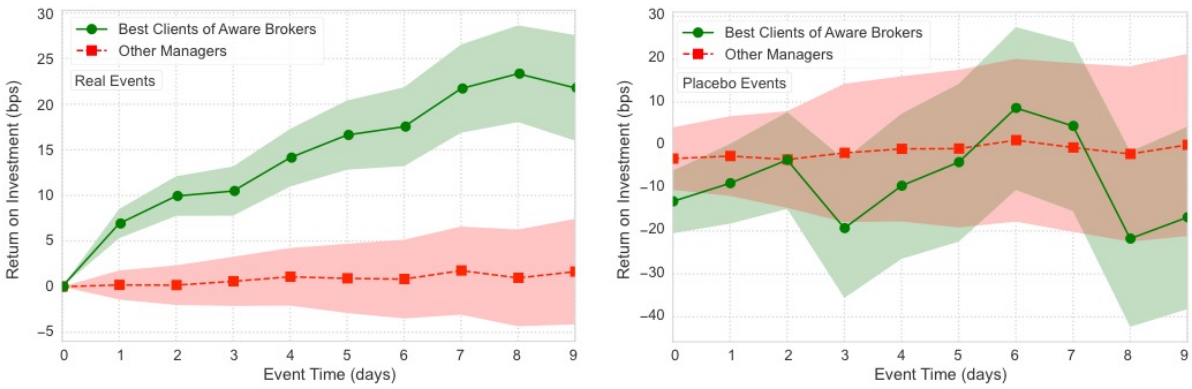


Figure 4. Profitability of predatory trades: liquidation events (left) and placebo sample (right). The left panel of the figure plots the profits of managers that are the best clients of aware broker (green solid line with circles) and unaware brokers (red dashed line with squares) during the fire-sale events. The best clients of a broker are defined as managers generating more than 5% of the volume intermediated by that broker in the previous semester. The right panel repeats this exercise for random event windows other than the actual fire-sale windows employed in the analysis.

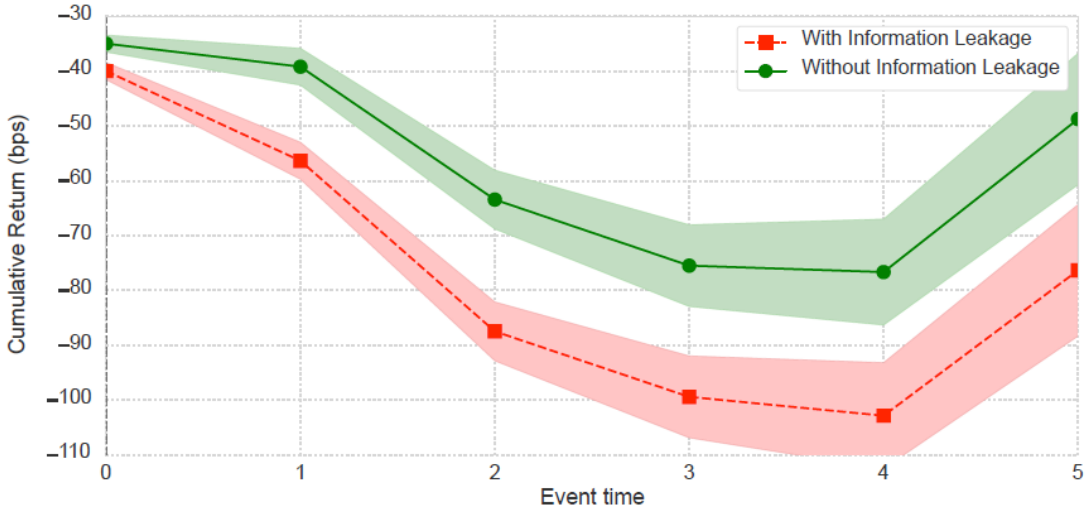


Figure 5. Price paths with and without information leakage. The figure plots the cumulative return of fire-sale stocks during fire-sale events involving at least one aware broker. The red line with squares represents the cumulative return averaged across stock-events in which aware brokers are present. The green line with circles represents the cumulative return averaged across stock-events in which no aware brokers are present. The series are based on estimates of a regression specification similar to that reported in Table VIII, but run on daily observations.

Table I
Summary Statistics

In Panels A, B, and C, we report summary statistics for the volume Z-score and the 385 fire-sale events identified by our methodology. In Panel D we report summary statistics for the manager-broker relationship proxies employed in the paper, expressed in percentage units. To identify fire-sale events, we start by computing the signed volume Z-score Z_t^m for each manager *m* on day *t* as $Z_t^m = (DVol_t^m - E(DVol_t^m))/\sigma(DVol_t^m)$, where $DVol_t^m$ is the portfolio level dollar volume traded by manager *m* on day *t*, and its mean and standard deviation are estimated over a rolling window of 120 trading days ending one week before day *t*. Then, at the portfolio level, we define manager *m* as *liquidating* if Z_t^m is below -0.25 for at least five trading days in a row. Next, we impose a stock level filter: for stock *j* to enter the fire-sale basket, we require that the volume traded by the manager be above 1% of the CRSP daily volume for at least four of the fire-sale days. Finally, we keep events in which at least 10 stocks are sold by the liquidating fund. Standard errors are clustered at the event level. In Panel E we regress the amount of each stock sold as a fraction of the total fire-sale volume on a set of stocks characteristics, while in Panel F we regress the first day in which a stock is sold the first time by the liquidating fund, in event time, on the same set of stock characteristics. *t*-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Panel A: Volume Z-Score

	Obs.	Mean	S.D.	Min	0.25	0.5	0.75	Max
All Managers-Days	941,219	-0.035	3.249	-41.714	-0.369	0.027	0.394	35.889
Fire-Sale Days	2,210	-2.075	4.518	-41.714	-1.768	-1.038	-0.616	-0.251
Fire-Sale Events	385	-2.002	3.410	-37.818	-1.672	-1.172	-0.878	-0.344

Panel B: Fire-sale events

	Units	Obs.	Mean	S.D.	25%	50%	75%	90%
Dollar Volume	Million Dollars	385	-377.062	534.635	-503.571	-177.461	-50.544	-18.244
Fraction of Portfolio	Percentage	385	9.164%	23.921%	1.224%	2.274%	5.879%	15.828%
Number of Stocks		385	21.917	10.090	13	18	29	38
Event Length	Trading Days	385	5.766	1.439	5	5	6	7
Number of Brokers		385	28.803	16.095	18	27	39	52
Number of Aware Brokers		385	1.694	0.968	1	2	2	3

Panel C: Fire-sale stocks

	Units	Obs.	Mean	S.D.	25%	50%	75%	90%
Dollar Volume	Million Dollars	8,438	-17.204	20.305	-23.401	-11.246	-3.542	-1.366
CRSP Volume Ratio	Percentage	8,438	-14.576%	16.000%	-18.749%	-9.922%	-4.585%	-2.409%
Price Decrease in [0,4]	Percentage	8,438	0.831%	4.613%	-1.904%	0.666%	3.388%	7.131%
Number of Brokers		8,438	5.737	5.039	2	4	8	13
Number of Aware Brokers		8,438	0.522	0.603	0	0	1	1

Panel D: Manager-Broker Relathionship Proxies

	Obs.	Mean	S.D.	Min	25%	50%	75%	90%	Max
Ranking Based on Volume	501,568	0.035	0.079	0.000	0.000	0.004	0.031	0.101	0.965
Ranking Based on Commission Paid	501,568	0.032	0.071	0.000	0.000	0.005	0.032	0.088	0.924

Panel E: Fire-Sale Stock Selection						
Dependent Variable:	Amount Sold as a Fraction of Total Fire-Sale Volume					
	(1)	(2)	(3)	(4)	(5)	(6)
Portfolio Weight	1.863***	1.830***	1.319***	1.805***	1.301***	1.318***
	(6.522)	(6.427)	(5.875)	(6.540)	(5.815)	(5.842)
Amihud Ratio		-0.691***			-0.486***	-0.506***
		(-8.419)			(-6.579)	(-6.775)
Market Cap			2.614***		2.427***	2.441***
			(11.580)		(10.926)	(10.977)
Volatility				-6.698***	-3.838***	-3.394***
				(-12.549)	(-7.296)	(-6.438)
One-Month Return						0.112
						(0.981)
Six-Months Return						0.209*
						(1.741)
One-Year Return						0.340***
						(2.783)
Observations	7,948	7,948	7,948	7,948	7,948	7,948
R^2	0.134	0.142	0.237	0.164	0.253	0.257
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Manager FE	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes

Panel F: Fire-Sale stock Timing						
Dependent Variable:	First Day In Which The Stock Is Sold					
	(1)	(2)	(3)	(4)	(5)	(6)
Portfolio Weight	-0.034***	-0.033***	-0.026***	-0.033***	-0.025***	-0.025***
	(-3.862)	(-3.732)	(-3.218)	(-3.831)	(-3.128)	(-3.184)
Amihud Ratio		0.028***			0.025***	0.025***
		-4.008			-3.515	-3.53
Market Cap			-0.040***		-0.036***	-0.037***
			(-5.736)		(-5.233)	(-5.319)
Volatility				0.113***	0.055	0.05
				-2.982	-1.443	-1.311
One-Month Return						0.011
						-1.228
Six-Month Return						0.002
						-0.189
One-Year Return						-0.018*
						(-1.785)
Observations	7948	7948	7948	7948	7948	7948
R^2	0.209	0.211	0.213	0.211	0.215	0.215
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Manager FE	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes

Table II
Predatory Behavior and Broker Awareness

The table reports results on the likelihood of a broker attracting predatory trades. The regressions are run at the ticket level, excluding trades by managers originating the fire sale of interest or another overlapping fire sale. In columns (1) and (2) of Panel A, the dependent variable is the difference between a dummy indicating predation and a dummy indicating liquidity provision, that is, it takes a value of one if the trade is in the same direction as the volume by the liquidating fund for that stock on that day (i.e., it is a sell trade), is negative one if the trade is in the opposite direction (i.e., a buy trade) and zero if the manager is not trading that stock on that particular day. In columns (3) and (4) of Panel A, we multiply this dependent variable by the volume of the trade as a fraction of market capitalization, standardized. The independent variable *Aware* is a dummy defined at the event-broker-stock-day level indicating whether the broker is aware of a fire sale happening on the traded stock on that day. More precisely, a broker is considered aware if stock j on day t of fire-sale event e broker b intermediates transactions on stock j from the liquidating fund originating e that amount at least 2% of the average daily volume of stock j. In Panel B we focus on predation, using the *Predation* dummy as dependent variable in columns (1) and (2) and its volume-weighted counterpart in columns (3) and (4). In Panel C we focus on liquidity provision, using the *Liquidity Provision* dummy as dependent variable in columns (1) and (2) and its volume-weighted counterpart in columns (3) and (4). All specifications include date, manager, and broker fixed effects. We further add broker-stock fixed effects in odd-numbered columns and day-stock fixed effects in even-numbered columns. Standard errors are clustered at the broker level. t-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Panel A: Net Predation

Dependent Variable:	Probability of Predation - Probability of Liquidity Provision		Predatory Volume - Liquidity Provision Volume	
	(1)	(2)	(3)	(4)
Aware Dummy	0.202***	0.113***	0.039***	0.027**
	(7.142)	(5.199)	(3.080)	(2.323)
Time Fixed Effects	Yes	Yes	Yes	Yes
Manager Fixed Effects	Yes	Yes	Yes	Yes
Broker Fixed Effects	Yes	Yes	Yes	Yes
Broker × Stock FEs	Yes		Yes	
Day × Stock FEs		Yes		Yes
Observations	487,605	462,841	487,605	462,841
R^2	0.203	0.229	0.136	0.159

Panel B: Predation				
Dependent Variable:	Probability of Predation		Predatory Volume	
	(1)	(2)	(3)	(4)
Aware Dummy	0.103***	0.124***	0.062***	0.049***
	(7.469)	(11.399)	(4.578)	(4.065)
Time Fixed Effects	Yes	Yes	Yes	Yes
Manager Fixed Effects	Yes	Yes	Yes	Yes
Broker Fixed Effects	Yes	Yes	Yes	Yes
Broker × Stock FEs	Yes		Yes	
Day × Stock FEs		Yes		Yes
Observations	487,605	462,841	487,605	462,841
R^2	0.360	0.274	0.189	0.193

Panel C: Liquidity Provision				
Dependent Variable	Probability of Liquidity Provision		Liquidity Provision Volume	
	(1)	(2)	(3)	(4)
Aware Dummy	-0.099***	0.011	0.022**	0.024**
	(-6.806)	(0.716)	(2.483)	(2.448)
Time Fixed Effects	Yes	Yes	Yes	Yes
Manager Fixed Effects	Yes	Yes	Yes	Yes
Broker Fixed Effects	Yes	Yes	Yes	Yes
Broker × Stock FEs	Yes		Yes	
Day × Stock FEs		Yes		Yes
Observations	487,605	462,841	487,605	462,841
R^2	0.369	0.263	0.155	0.173

Table III
Probability of Predation and Broker-Client Relationship Strength

The table presents evidence on the effect of broker-client relationship strength on the probability of predatory behavior. The regressions are run at the stock-day-manager-broker level, excluding trades by managers originating the fire sale of interest or another overlapping fire sale. In all specifications the dependent variable is the difference between a dummy indicating predation and a dummy indicating liquidity provision, that is, it takes the value of one when the trade is in the same direction as the volume by the liquidating fund for that stock on that day (i.e., it is a sell trade), negative one if the trade is in the opposite direction (i.e., a buy trade) and zero if the manager is not trading that stock on that day. In Panel A we regress the dependent variable on the continuous variable *Best Client*, which measures the strength of the manager-broker relationship, the dummy *Liquidation Period*, indicating the first five days of the fire sale, and the interaction between the two. The relationship strength variables are definined as follows: (i) the ranking based on volume $\mathrm{RVol}_{\mathrm{m,b,t}}$ is the fraction of dollar volume intermediated by broker *b* in the semester preceeding day *t* that is generated by manager *m* and (ii) the ranking based on commissions paid $\mathrm{RCom}_{\mathrm{m,b,t}}$ is the fraction of dollar commissions earned by broker *b* in the semester preceeding day *t* which is generated by manager *m*. Both variables are expressed in decimal units and thus take values in the interval $[0,1]$. The dummy *Liquidation Period* is zero in the five days before the fire sale. We consider all trades on stock *i* intermediated by brokers that eventually become aware that the stock is subject to fire-sale pressure, that is, brokers *b* for which $max_{t\in[0,4]}(Aware_{i,b,t,e}) = 1$. The regression is run on a sample that includes the five days before the fire sale and the five days from the start of the fire sale, defined as the first day in which our liquidation measure crosses the threshold. In Panel B we regress the dependent variable on the triple interaction between: *Aware Broker*, indicating that the broker is aware, *Best Client*, and *Liquidation Period*, indicating the first five days of the fire sale. Time fixed effects are at the monthly frequency. Standard errors are clustered at the event-stock-manager level. *t*-statistics of the differences are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Panel A: Difference in Differences

Dependent Variable:	Probability of Predation - Probability of Liquidity Provision		Predatory Volume - Liquidity Provision Volume	
	(1)	(2)	(3)	(4)
Best Client Proxy:	Ranking Based on Volume	Ranking Based on Commissions Paid	Ranking Based on Volume	Ranking Based on Commissions Paid
Best Client × Liquidation Period	0.055***	0.081***	0.124***	0.127***
	(3.181)	(4.182)	(3.236)	(2.806)
Best Client	0.023	0.048**	-0.001	0.003
	(1.427)	(2.500)	(-0.038)	(0.096)
Liquidation Period	0.006***	0.005***	0.002	0.003
	(5.683)	(4.942)	(0.618)	(0.661)
Time Fixed Effects	Yes	Yes	Yes	Yes
Manager Fixed Effects	Yes	Yes	Yes	Yes
Event Fixed Effects	Yes	Yes	Yes	Yes
Stock Fixed Effects	Yes	Yes	Yes	Yes
Broker Fixed Effects	Yes	Yes	Yes	Yes
Observations	501,567	501,567	501,567	501,567
R^2	0.046	0.046	0.013	0.013

Panel B: Triple Interaction				
Dependent Variable:	Probability of Predation - Probability of Liquidity Provision			
	(1)	(2)	(3)	(4)
Best Client Proxy:	Ranking Based on Volume	Ranking Based on Commissions Paid	Ranking Based on Volume	Ranking Based on Commissions Paid
Aware Broker × Best Client × Liquidation Period	9.814***	11.589***	9.184***	10.794***
	(4.527)	(4.927)	(4.356)	(4.756)
Best Client × Liquidation Period	1.752***	1.287***	1.554***	1.157***
	(5.664)	(5.220)	(4.989)	(4.634)
Aware Broker × Liquidation Period	0.011***	0.011***	0.011***	0.011***
	(10.252)	(9.552)	(10.196)	(9.578)
Best Client × Aware Broker	7.466***	9.082***	7.500***	8.949***
	(3.893)	(4.018)	(3.755)	(3.750)
Best Client	3.747***	2.969***	3.664***	2.931***
	(18.986)	(14.932)	(13.956)	(10.602)
Aware Broker	0.006***	0.005***	0.003***	0.003***
	(7.138)	(5.686)	(4.035)	(3.550)
Liquidation Period	0.012***	0.012***	0.011***	0.011***
	(39.642)	(41.789)	(34.415)	(36.251)
Constant	0.021***	0.022***		
	(99.348)	(100.764)		
Time Fixed Effects			Yes	Yes
Manager Fixed Effects			Yes	Yes
Event Fixed Effects			Yes	Yes
Stock Fixed Effects			Yes	Yes
Broker Fixed Effects			Yes	Yes
Observations	4,226,877	4,226,877	4,128,803	4,128,803
R^2	0.005	0.004	0.022	0.022

Table IV
Evidence of Predation on Multiple Stocks

The table reports results on the number of stocks experiencing predatory pressure. For each fire-sale event, we consider the basket of liquidated stocks, and for each manager actively trading at least one stock in the basket we count the number of stocks traded in the same direction as the fire-sale originator. In the first two specifications, we consider event-manager observations and we regress the number of predated stocks on the best client proxies. These are constructed by interacting the original best client proxies with the broker awareness dummy at the ticket level, and then taking the maximum value at the event-manager level. In other words, the relationship strength assigned to each manager is the value of the best relationship across the aware brokers in the fire-sale event. The number of predated stocks is calculated across all fire-sale stocks predated by the manager across all brokers. In specifications (3) and (4), we repeat this exercise by using as the dependent variable the fraction of predated stocks relative to the stocks in the fire-sale basket. Event, manager, and day fixed effects are included in the regressions, and standard errors are double-clustered at the manager and event level. *t*-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Dependent Variable:	Number of Predated Stocks		Fraction of Predated Stocks	
	(1)	(2)	(3)	(4)
Best Client Proxy:	Ranking Based on Volume	Ranking Based on Commissions Paid	Ranking Based on Volume	Ranking Based on Commissions Paid
Best Client	14.551***	15.066***	57.855***	59.791***
	(4.864)	(4.516)	(5.148)	(4.855)
Time Fixed Effects	Yes	Yes	Yes	Yes
Manager Fixed Effects	Yes	Yes	Yes	Yes
Event Fixed Effects	Yes	Yes	Yes	Yes
Observations	28,168	28,168	28,168	28,168
R^2	0.390	0.386	0.465	0.461

Table V
Predators' Position Reversal

This table reports results on the fraction of sales in a given stock that a given manager subsequently reverses. In a given time period, either before or after the beginning of the fire sale, the percentage of manager m's position in stock j that is reversed during event e is defined as the ratio $Rev_{e,m,j} = BoughtBack_{e,m,j} \ / \ Sold_{e,m,j}$, where $Sold_{e,m,j}$ is the dollar sum of all sell orders in the period and $BoughtBack_{e,m,j}$ is the dollar sum of buy orders during the period, where we sum only the buy orders that are preceded by negative cumulative order flow. We compute this measure around each fire-sale event, for the event time periods $Pre = [-10, -1]$ and $Post = [0,9]$, considering all trades on stock j intermediated by brokers who eventually become aware that the stock is subject to fire-sale pressure. We then compare the percentage of positions reversed across clients with different relationship strength with the aware brokers before (Pre) and during ($Post$) the fire-sale events. Liquidating funds are excluded from the sample. In columns (1) and (2) we present results for the specifications without fixed effects, while in columns (3) and (4) we report results based on time, stock, and manager fixed effects. Standard errors are clustered at the manager level. t-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Dependent Variable:	Percentage of Positions Reversed			
	(1)	(2)	(3)	(4)
Best Client Proxy:	Ranking Based on Volume	Ranking Based on Commissions Paid	Ranking Based on Volume	Ranking Based on Commissions Paid
Best Client × Post[0,9]	25.091***	24.110***	23.352***	20.676***
	(11.788)	(7.413)	(5.404)	(4.151)
Best Client	5.128***	6.448***	-7.022**	-1.939
	(3.441)	(2.791)	(-2.010)	(-0.526)
Post[0,9]	11.427***	12.764***	16.287***	18.320***
	(17.298)	(19.318)	(14.256)	(15.494)
Constant	2.723***	2.878***		
	(5.788)	(6.116)		
Time Fixed Effects			Yes	Yes
Stock Fixed Effects			Yes	Yes
Manager Fixed Effects			Yes	Yes
Observations	37,276	37,276	31,000	31,000
R^2	0.028	0.023	0.258	0.256

Table VI
Evidence from the 2003 Mutual Fund Scandal

This table reports results on the natural experiment based on the late-trading scandal. We first match the list of 27 mutual fund families involved in the 2003 late-trading scandal with managers in our data set and classify them as charged. We focus on daily transactions of the managers not involved in the scandal for a period of four years centered on the month of the announcement of the complaint by Spitzer (September 2003) and define the dummy $Post\ Scandal_t$, indicating the two years after the complaint broke out. Next, we define the broker-stock-day-level dummy variable $Selling_{b,j,t}$ indicating whether at least one of the charged funds is selling stock *j* on day *t* through broker *b*. We then define the dependent variable *Probability of Predation* as a dummy variable equal to one if a non-charged manager sells stock *j* on day *t* through broker *b*, and zero if a non-charged manager trades on a different day, on a different stock, or with a different broker. Finally, we regress the probability of predation minus the probability of liquidity provision (defined as in Table III) on the interaction between $Selling_{b,j,t}$ and the dummy $Post\ Scandal_t$. Standard errors are clustered by manager-stock and *t*-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Dependent Variable:	Probability of Predation - Probability of Liquidity Provision				
	(1)	(2)	(3)	(4)	(5)
Selling × Post Scandal	0.087***	0.097***	0.069***	0.060***	0.046***
	(11.406)	(12.800)	(9.261)	(8.220)	(6.342)
Selling	0.147***	0.141***	0.148***	0.152***	0.179***
	(23.040)	(22.135)	(22.406)	(23.537)	(28.281)
Post Scandal	-0.025***				
	(-9.289)				
Time Fixed Effects		Yes	Yes	Yes	Yes
Manager Fixed Effects			Yes	Yes	Yes
Stock Fixed Effects				Yes	Yes
Broker Fixed Effects					Yes
Observations	12,087,004	12,087,004	12,087,001	12,086,863	12,086,781
R^2	0.001	0.013	0.068	0.076	0.082

Table VII
Profitability of Predatory Trades

The table reports results on the profitability of trades by predators around the fire-sale events. We divide each event into a pre-fire-sale period $[-10,-1]$ and a post-fire-sale period [0,9], where day 0 denotes the day on which the fire sale starts. We then compute the profitability of trades by manager *m* on stock *j* over the window $\pi = [t_0, t_1]$, which denotes either the pre- or post-fire-sale period. The profitability measure that we use as dependent variable in all specifications is calculated using the formula

$$Profitability_{m,j,\pi} = \left(MarkToMarket_{m,j,\pi} - CashFlows_{m,j,\pi} \right) / Exposure_{m,j,\pi}.$$

Here, $MarkToMarket_{m,j,\pi}$ is the marked-to-market dollar value of the position at time t_1, defined as the product of the share position cumulated from t_0 to t_1 with the market price of stock *j* on day t_1, $CashFlows_{m,j,\pi}$ is the dollar amount spent to build the position, that is, the opposite of the dollar volume of each transaction in the stock (based on execution prices) from from t_0 to t_1, and $Exposure_{m,j,\pi}$ is the maximum dollar ependiture over the relevant period, defined as $\max_{t\in\pi} \left| CashFlows_{m,j,[t_0,t]} \right|$. We relate the profitability (expressed in basis points) of trades executed by aware brokers to our relationship strength proxies (i.e. the fraction of the volume intermediated by the broker over the previous semester generated by the manager, expressed in decimal units, as well as the fraction of the commissions) in the pre- and post-fire-sale periods using event-manager-stock-level observations. In rows (1) and(2) we present results for the specifications without fixed effects, while in rows (3)-(4) we report results based on time, stock, and manager fixed effects. Standard errors are clustered at the manager level. *t*-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Dependent Variable:	Return on Capital (basis points)			
	(1)	(2)	(3)	(4)
Best Client Proxy:	Ranking Based on Volume	Ranking Based on Commissions Paid	Ranking Based on Volume	Ranking Based on Commissions Paid
Best Client × Post[0,9]	136.558**	144.821**	147.847**	159.582**
	(2.355)	(2.235)	(2.110)	(1.966)
Best Client	-48.574	-61.826	-78.883**	-108.693***
	(-1.145)	(-1.303)	(-2.414)	(-2.929)
Post[0,9]	-7.160***	-7.102***	-7.719**	-7.665**
	(-2.783)	(-2.761)	(-2.520)	(-2.503)
Constant	8.646***	8.697***		
	(4.651)	(4.679)		
Time Fixed Effects			Yes	Yes
Manager Fixed Effects			Yes	Yes
Observations	263,346	263,346	263,211	263,211
R^2	0.000	0.000	0.034	0.034

Table VIII
Price Impact and Broker Awareness

This table reports results on the price impact experienced by fire-sale originators. Considering all trades by fire-sale originators from the beginning of each fire-sale event (t=0) to the last day of the fire sale (i.e., the last day on which the criteria for being classified a fire sale are satisfied), we construct the following price impact measures: (i) the execution shortfall based on the first placement price, (ii) the execution shortfall based on the first open price, and (iii) the execution shortfall based on the first transaction price. We aggregate the measures taking their volume-weighted average across transactions and express them in basis points. In specifications (1) to (3) we regress the price impact measures on a dummy indicating the presence of at least one aware broker at the event-stock level and the total volume of other managers relative to stock market capitalization. We control for the originator's volume relative to stock market capitalization and the Amihud ratio of the stock, estimated over the previous six months. Time and stock fixed effects are included in the regression. In specifications (4) to (6), we repeat the exercise at the event-stock-broker level and also add broker fixed effects. Continuous explanatory variables are standardized and standard errors are clustered by event. *t*-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Dependent Variable:	Price Impact (basis points)					
Granularity:	Stock Level			Broker-Stock Level		
	(1)	(2)	(3)	(4)	(5)	(6)
Benchmark Price:	First Placement Price	Market Open Price	First Transaction Price	First Placement Price	Market Open Price	First Transaction Price
Aware Broker Dummy	25.176*	36.194**	14.250	11.901***	11.320**	8.970**
	(1.849)	(2.503)	(1.442)	(2.764)	(2.217)	(2.496)
Followers Volume	23.801***	24.286***	8.520*	4.882**	4.898*	2.457
	(2.787)	(2.710)	(1.680)	(2.020)	(1.815)	(1.259)
Generator Volume	6.996	8.560	0.520	21.760***	20.890***	11.607**
	(0.646)	(0.706)	(0.067)	(3.691)	(3.293)	(2.449)
Amihud Ratio	-19.080	-20.435	-18.598	-12.067	-6.532	-8.238
	(-1.070)	(-1.101)	(-1.382)	(-1.262)	(-0.703)	(-1.437)
Number of Brokers	-3.489	1.193	-1.938	6.359	9.710	4.731
	(-0.515)	(0.152)	(-0.373)	(1.137)	(1.509)	(1.334)
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Stock Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Broker Fixed Effects				Yes	Yes	Yes
Observations	6,291	6,291	6,291	28,265	28,265	28,265
R^2	0.430	0.430	0.415	0.323	0.338	0.265

Table IX
Commissions Paid to Aware Brokers

The table presents evidence on the post-event increase in commissions paid by predators to aware brokers. For each month t on a window starting two years before and ending two year after each fire-sale event e, we calculate the average $Commission\_per\_dollar_{e,m,b,t}$ paid by manager m to broker b as the ratio $Comm_{e,m,b,t}/DVol_{e,m,b,t}$, where $Comm_{e,m,b,t}$ is the total dollars paid in commissions by manager m to broker b during month t and $DVol_{e,m,b,t}$ is the total dollar volume traded by manager m and intermediated by broker b in that month. For each event, we consider brokers which are considered *Aware* for at least one of the fire-sale stocks and managers whose trades are intermediated by at least one of these brokers in the 10 trading days around the event. We then regress $Commission\_per\_dollar_{e,m,b,t}$ on the interaction between the dummy variable $Post_{e,t}$, indicating the two years following the fire-sale event, and each of our best client proxies. Standard errors are clustered by event-broker-manager to account for time-series autocorrelation in commissions paid. t-statistics are reported in parentheses. Asterisks denote significance levels (***=1%, **=5%, *=10%).

Dependent Variable:	Commissions per dollar (basis points)			
	(1)	(2)	(3)	(4)
Best Client Proxy:	Ranking Based on Volume	Ranking Based on Commissions Paid	Ranking Based on Volume	Ranking Based on Commissions Paid
Best Client × Post	1.287*	0.858	1.321**	0.857
	(1.808)	(1.308)	(2.171)	(1.391)
Best Client	-12.892***	-10.572***	-4.310***	-2.065**
	(-6.016)	(-5.328)	(-5.617)	(-2.530)
Post	-0.387***	-0.368***	-0.583***	-0.565***
	(-7.874)	(-7.631)	(-11.425)	(-11.144)
Constant	6.402***	6.295***		
	(59.983)	(58.830)		
Time Fixed Effects			Yes	Yes
Manager Fixed Effects			Yes	Yes
Broker Fixed Effects			Yes	Yes
Observations	1,168,535	1,168,535	1,168,521	1,168,521
R^2	0.016	0.009	0.228	0.227

Quantitative Easing and Equity Prices: Evidence from the ETF Program of the Bank of Japan

Andrea Barbon and Virginia Gianinazzi *

Abstract

Since the introduction of its Quantitative and Qualitative Easing program in 2013, the Bank of Japan has been increasing its holdings of Japanese equity through large scale purchases of index-linked ETFs, to lower risk premia. We exploit the cross-sectional heterogeneity of the supply shock to identify a positive and persistent impact on stock prices, consistent with a portfolio balance channel. The evidence suggests that long-run demand curves for stocks are downward sloping with unitary price elasticity. We show that the purchases of ETFs tracking the price-weighted Nikkei 225 generate pricing distortions relative to a value-weighted benchmark.

Keywords: ETFs, quantitative easing, portfolio balance channel, unconventional monetary policy, Bank of Japan

JEL classification: E52, E58, B26

*Andrea Barbon (andrea.barbon@usi.ch, 17 Banks St 02138 Cambridge MA, +393408457736) and Virginia Gianinazzi (virginia.gianinazzi@usi.ch) are with USI Lugano and the Swiss Finance Institute. This work was supported by the Swiss Finance Institute. We would like to thank Francesco Franzoni and Alberto Plazzi for their advice and the stimulating discussions. We are also grateful to Miguel Antón, Harjoat S. Bhamra, Andrea Buraschi, John Campbell, Marco Di Maggio, Ruediger Fahlenbrach, Thierry Foucault, Laurent Frésard, Robin Greenwood, Peter Gruber, Jonathan Kearns, Marco LiCalzi, Loriano Mancini, David Miles, Jouchi Nakajima, Jean-Charles Rochet, Suresh Sundaresan, poster session participants at the AFA meeting 2018 and participants at the OFCE/Science-Po Workshop on Empirical Monetary Economics, SFI Research Days, SFI PhD Workshop, Summer School on Market Microstructure and JEN Conference for helpful comments. All errors are our own.

1 Introduction

With policy interest rates constrained at the zero lower bound, many central banks around the world have resorted to unconventional monetary policy tools. Within the range of unconventional measures, large-scale asset purchase (LSAP) programs have attracted particular attention because of their large size and thus their impact on central banks' balance sheets. The massive expansion of both the assets and liabilities of central banks exposes them to considerable risks and raises questions about the consequences of a potential exit from QE.

There is considerable evidence that central banks' asset purchases can have an economically significant impact on yields in the targeted markets, which has likely motivated central banks to continue these purchases over the past decade (D'Amico and King, 2013, Eser and Schwaab, 2016, Gagnon et al., 2010, Hamilton and Wu, 2012, Krishnamurthy and Vissing-Jorgensen, 2011, 2013, Neely et al., 2010, Swanson, 2011). However, despite the widespread use of LSAP programs, the debate is still ongoing with regard to the mechanisms linking asset purchases to asset prices and the persistence of the impact. Unlike policy rate targeting, asset purchases are explicit decisions on quantities and are designed to have a noticeable impact on market prices. Even though the idea of *easing through quantity* relies on the view that large purchases by the central bank reduce assets' risk premia, there is still no clear theoretical foundation for how and under which conditions this is expected to work. In general, the relationship between the outstanding quantity of an asset and its price is not yet well understood.

Since 2013, the Bank of Japan (BoJ) has been engaging in what they have named *Quantitative and Qualitative Easing* (QQE) program as an attempt to fight against deflation. As part of its broader QQE agenda, the BoJ has been vigorously increasing its domestic equity holdings through purchases of index-linked ETFs. By the end of 2016, the BoJ owned approximately ¥14 trillion worth of TOPIX and Nikkei ETFs, which corresponds to more than 2.5% of the total market capitalization. This unprecedented equity operation has the declared objective of lowering the risk premia of asset prices and reducing the cost of equity capital of Japanese companies (BoJ, 2013).

The BoJ is the first central bank to engage in purchases of domestic equities as part of its QE agenda. This intervention represents a unique laboratory to shed light on the long-standing debate on the elasticity of long-term demand curves for stocks, as well as a unique opportunity to test how QE impacts equity prices and its implications for market efficiency. In this paper we study its impact on the cross-section of stock prices. We propose a novel empirical strategy to identify and quantify the price impact of the change in assets' supply through QE. This provides new evidence on the price elasticity of long-run demand curves.

The literature on the effectiveness of QE has proposed several channels through which central

banks can affect prices. A natural explanation is provided by the so-called "portfolio-balance" channel, first discussed by Brunner and Meltzer (1973), Frankel (1985) and Tobin (1969). According to this channel, when the central bank buys a particular asset, it reduces the amount held by the private sector, effectively changing the risk composition of the aggregate portfolio held by investors. For this to be an equilibrium, prices need to adjust to ensure market clearing.

In this paper, we first propose a simple structural asset pricing model that generalizes the idea of the portfolio balance channel to the case of equities.[1] The key implication of portfolio rebalancing that we derive from the model is that the change in systematic risk of each stock is determined by: (i) the entire vector of central bank purchases and (ii) the covariance matrix of stocks' cash flows.

We then bring the model to the data in a standard event-study framework, exploiting two specific events in which the BoJ announced major expansions of its ETF purchases. On October 31, 2014, the BoJ announced a three-fold increase in the purchase of ETFs and on July 29, 2016, it communicated a further doubling of the budget amount.

We document that both announcements produced a highly heterogeneous response of equity prices at the company level. Figure 1 plots the cumulative returns of two portfolios following the 2014 announcement, formed by ranking stocks on the price impact predicted by the model. The divergence in returns is statistically and economically significant. Results from cross-sectional regressions show that the variation in event returns in the cross-section is consistent with the change in the marginal contribution of each stock to the risk of the aggregate portfolio held by private investors, as predicted by the portfolio-balance channel.

Looking at longer-horizon returns we find no evidence of reversal over a one-year window after both policy announcements, which supports the main time-series prediction of the model. We estimate the long-term net effect of the portfolio-balance channel at about 22 basis points increase in market value per trillion Yen employed in the program. Given a total equity market capitalization of about ¥500 trillion, this implies an elasticity close to 1, so that each yen invested translates into an increase in total market valuation of roughly one yen. Our estimate is in line with those provded by (Shleifer, 1986a) and (Petajisto, 2011), who find an elasticity of 1 and 0.84, respectivetly, using additions to the S&P 500 index. The analysis based on Dutch auction repurchases of (Bagwell, 1992) also results into a relatively close price elasticity of 1.65 Other authors, instead, find flatter demand curves with price elasticity ranging from the value of 8.24 estimated by (Wurgler and Zhuravskaya, 2002) to that of 10.5 by (Kaul et al., 2000).

[1] It is easy to show that the duration channel discussed in the literature is a special case of our model when all securities in the economy are exposed to a single source of risk.

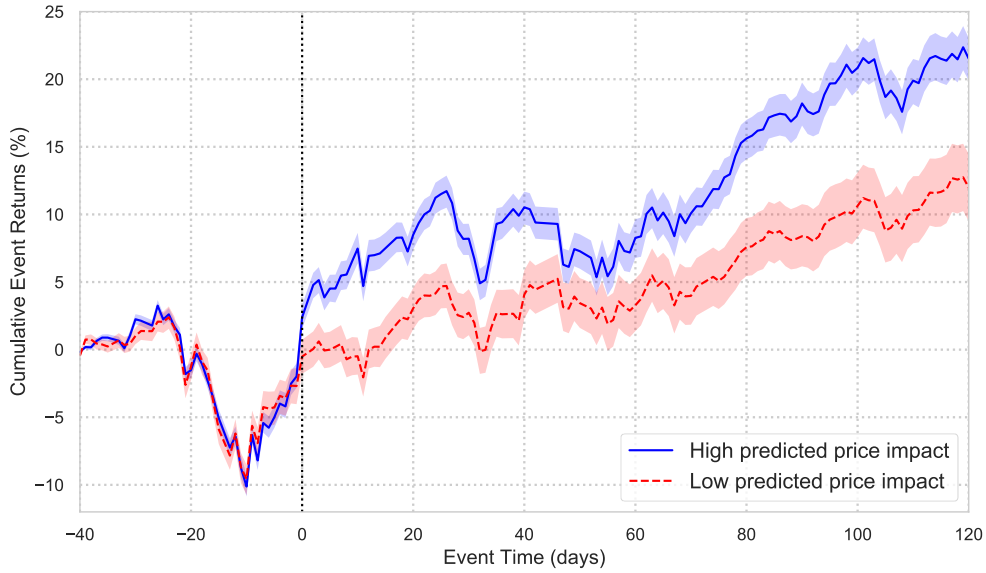


Figure 1: Heterogeneity of Event Returns. This figure shows the time series of the cumulative returns around the BoJ announcement of October 2014, of two portfolios of firms ranked by the model predicted returns. The blue line is the average for first quartile of the distribution (firms with the highest predicted price impact), while the red dashed line corresponds to the average for the last quartile (firms with the lowest predicted price impact). Bands represent bootstrapped 95% confidence intervals.

The two expansions of the policy budget provide us with an ideal natural experiment to examine the net effect of a long-lasting change in supply on prices for three reasons. First, the purchase schedule of the central bank is exogenous to firms' fundamentals in the cross-section. Second, unlike asset purchases by the Federal Reserve, the program of the BoJ affects the supply of each security according to an ex ante well-defined purchase schedule. Third, since roughly half of the capital of the central bank is allocated according to the weights of the price-weighted Nikkei 225 index, the purchases produce variation in the cross-section of supply shocks relative to market capitalization that is as good as random. In general, the identification of the impact of LSAPs on asset prices is a challenging task. The intervention of the BoJ provides us with an empirical framework that mitigates endogeneity concerns and improves the identification of the net (short-run and long-run) effect of a change in supply, which crucially relies on the exogeneity of the shock.

The non-fundamental nature of the Nikkei weighting system has already been exploited in Greenwood (2005, 2007) to establish a causal relationship between uninformed demand shocks and prices in the context of a large redefinition of the Nikkei 225 index. A major difference between the LSAP setting and the one of index redefinitions lies in the nature of the supply shocks. As the central bank buys assets, it is effectively transferring a portion of fundamental

risk from the private sector to its balance sheet, and holds it for an arguably long period of time. This is at least conceptually different from an index redefinition event, in which securities merely change hands from active investors to index funds. The central bank can be thought of as a buy-and-hold long-term investor whose portfolio holdings are not marked-to-market. Its long-term commitment to the policy induces a long-lasting change to supply, making our setting better suited than index redefinitions to identify long-run price effects due to movements along investors' long-term demand curves.

The model that we propose extends the theoretical framework in Greenwood (2005) to account for this difference in setting. As in Greenwood (2005), we consider an economy with multiple assets in finite supply and a CARA-utility representative agent that maximizes her wealth in each period. We introduce quantitative easing in the form of an exogenous shock to the supply of assets, which is first announced and then gradually carried out over a given policy horizon. The agent correctly understands that the QE program will affect the market-clearing portfolio in each future period, which determines the new vector of equilibrium risk premia. Crucially, we extend the model to an infinite horizon to relax the assumption that uncertainty is resolved at a terminal date, which mechanically drives the reversal in Greenwood (2005). In our model, prices adjust to the change in supply to reflect the new risk composition of the aggregate portfolio held by the representative agent. Unless the central bank is expected to unwind its position, this implies that we should not observe a reversal at any horizon. The fact that we observe a persistent effect in the data is consistent with this prediction of the model.

In the data, not only we find no evidence of a reversal of the initial jump in prices, but non-trivial abnormal returns are still observed one year after the announcements. Even though the policy is carried out gradually, the total size of the intervention is revealed to the market in advance. Market efficiency requires that today's prices reflect expectations about future returns, hence they should also reflect expected future changes in the supply of assets. In the model, the observed post-event drift can arise because of two reasons. First, since purchases are scattered over various dates, the model predicts that prices continue to adjust also after the announcement due to the decrease in the residual duration of the program over time. However, for realistic levels of the risk free rate, we show that this effect is expected to be quantitatively small relative to the initial price jumps. A more pronounced drift arises when we allow the representative agent to believe that the central bank will deviate from the announced purchase target. The model produces a sluggish price reaction similar to the one observed in the data when expectations about the size of the purchase program are assumed to increase over time, consistently with investors underreacting to the announcement as well as with learning about additional purchase programs in the future.[2]

[2]Beliefs are exogenous in our model and evolve deterministically over time. Extending the model to a

We address the concern that (part of) the observed price impact and its persistence might be explained by repeated price pressure rather than a portfolio-balance channel. Large trades from the BoJ may give rise to order imbalances, thus pushing prices upwards on purchase days. Such mispricings are expected to be shortly lived in efficient markets. However, arbitrageurs may refrain from trading if the central bank is expected to buy again soon, thus failing to bring prices back to their fundamental value. This would imply a persistent price effect of the program arising from the flow of the purchases rather than the change in the supply of assets, a channel quite distinct from portfolio balance. As in D'Amico and King (2013), we will refer to this effect as the *flow effect* of the program. Even though the difference may seem subtle, disentangling between these two channels has important implications. First, the two channels lead to different conclusions about the elasticity of long-run demand curves for stocks. Second, they imply different consequences of a potential exit from QE. In particular, if QE is mainly effective through repeated price pressure, a slow-down or a suspension of the purchases would cause a sharp drop in prices. On the contrary, in our model of the portfolio-balance channel, it is not the flow into the balance sheet of the central bank that keeps prices up, but its accumulated size. Therefore, suspending the purchases should have a more limited effect on prices. We exploit both the cross-sectional and time series variation in purchase volumes to identify and quantify the flow effect of the policy in the spirit of Eser and Schwaab (2016). We then re-estimate the cross-sectional portfolio-balance channel effect using returns net of the flow-induced component. We find that price pressure effects are positive and persistent. However, this channel might explain at most a minor fraction (between 12% and 23% depending on the specification) of the estimated portfolio balance effect.

Overall, our empirical analysis confirms the concerns raised by the financial press that the intervention of the BoJ might be inducing price distortions due to the deviation of the purchase schedule from market weights. We document a significantly heterogeneous effect of the policy both at company and industry level. A modification of the QQE has the potential to address this problem. Theoretically, the only way to achieve a cross-sectionally homogeneous shift in risk premia is for the BoJ to hold each stock proportionally to the company market capitalization. At the moment, however, still roughly a quarter of the BoJ capital is allocated to the price-weighted Nikkei index.

The rest of the paper is organized as follows. Section 2 describes the ETF purchase program of the BoJ. Section 3 reviews the relevant literature. Section 4 presents the model. Section 5 describes the data, the empirical strategy and estimation procedures. Section 6 presents our main empirical findings. Section 7 considers the flow effect of direct purchases and evaluates its relative importance with respect to the portfolio balance effect. Section 8 discusses policy implications and Section 9 concludes.

setting where beliefs are endogenous is definitely interesting, but beyond the scope of the paper.

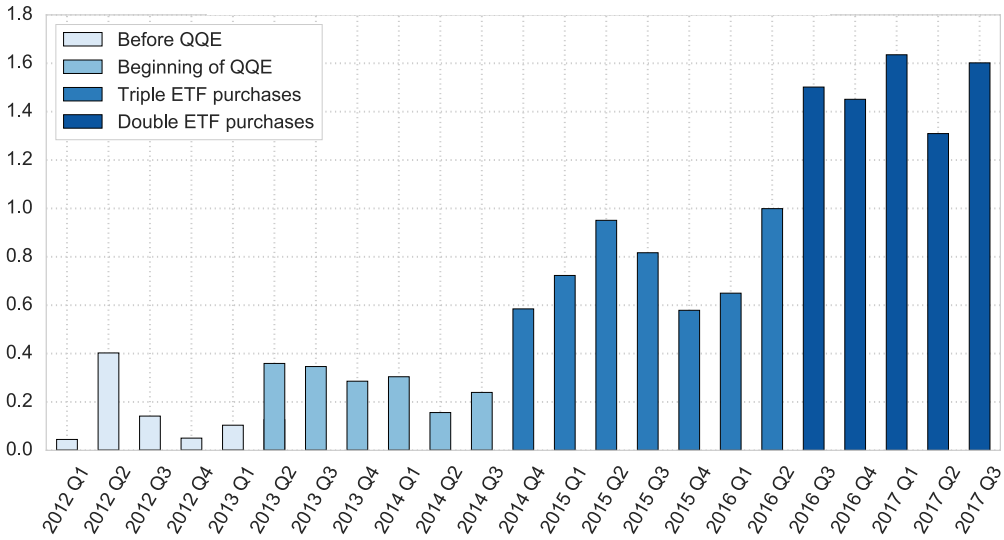


Figure 2: Quarterly ETF Purchases of the Bank of Japan in trillion yen. Changes in the bar color indicate changes in the policy target purchase amounts. In the first phase the target was set to ¥1 trillion per year, in the second phase it was tripled to ¥3 trillion and in the third phase it was additionally doubled to ¥6 trillion. Data is from the BoJ website.

2 The ETF Program of the BoJ

As part of the "Quantitative and Qualitative monetary Easing" (QQE) introduced on April 4, 2013, the BoJ embarked on a large-scale asset purchase (LSAP) program committing itself to buy large quantities of broad market equity ETFs, with the declared view of lowering risk premia (BoJ, 2013). The policy budget was initially set at ¥1 trillion per year (roughly US$ 10 billion). On two occasions, the BoJ announced a sharp expansion of the target amount: on October 31, 2014, the Bank communicated that the annual mark was tripled to ¥3 trillion, and was again doubled on July 29, 2016, to ¥6 trillion. The policy changes are clearly visible along the time series of monthly ETF purchases by the bank, as shown in Figure 2. The time series of aggregate ETF purchases is publicly available at daily frequency on the BoJ website starting from December 2010.

Its holdings accumulated rapidly, and by the end of 2016, the BoJ owned more than ¥14 trillion worth of ETFs. This corresponds to 2.5% of the total capitalization of the First Section of the Tokyo Stock Exchange (TSE), and around 3% of the Japanese GDP. The share of BoJ holdings to aggregate Assets Under Management (AUM) of targeted ETFs has grown from almost zero to more than 70% since the beginning of the program; this is even more remarkable if we consider that the ETF industry in Japan almost tripled in value between 2013 and 2016. In terms of size, the ETF program is comparable to the annual aggregate net flows into or

out of the Japanese equity fund industry and therefore economically relevant.[3]

The purchase program targets two types of ETFs: those tracking the Tokyo Stock Price Index (TOPIX) and those replicating the return of the Nikkei 225 Stock Average.[4] At inception of the program, the money allocated to each ETF was set to be proportional to its assets under management (AUM). The ratio of the aggregate AUM of ETFs tracking the TOPIX Index and those of ETFs tracking the Nikkei 225 Index is roughly 1 to 1.2. This approximately translates into half of the capital flowing into Nikkei ETFs and half into TOPIX ETFs. In turn, this then maps into a demand shock at the stock level that depends on each company's weight in the corresponding index.

The TOPIX is a value-weighted index tracking the roughly 2000 companies listed on the First Section of the TSE, while the Nikkei 225 is a *price-weighted* index of 225 TOPIX companies representative of the Japanese stock market. The constituents of the Nikkei index are typically large blue-chip companies that account for roughly two-thirds of the market capitalization of the TSE First Section on aggregate. The Nikkei 225 is the most widely traded equity benchmark in Japan.

The weighting system of the two indices implies that the BoJ allocates only half of its budget to companies proportionally to their market value. The remaining half of the budget flows instead to the Nikkei constituents proportionally to their price, not accounting for the number of shares outstanding, thus producing mis-allocation relative to market capitalization. Under market efficiency, the market value of a company should reflect all available fundamental information. The dispersion of the ratio between price weights and value weights is therefore expected to be unrelated to firms fundamentals. The relative under-weighting in the BoJ portfolio is clearly more severe for companies not included in the Nikkei index. However, there is a high degree of heterogeneity in the allocation of capital across Nikkei companies as well. This is clear from Figure **??** in Appendix **??**, where we plot the distribution of the log of the ratio between the weight in the Nikkei and the weight in the TOPIX for Nikkei companies to measure the cross-sectional dispersion of the resulting allocation at the stock level.

Given the unusual weights of the BoJ purchase schedule, a sudden expansion of the policy budget produces a natural experiment where stocks are hit by an uninformed demand shock that is highly heterogeneous in the cross-section and orthogonal to firms fundamentals after controlling for market capitalization. In this paper, we exploit the exogenous variation in the

[3] Over the past 10 years, the average net flows into equity funds in Japan was roughly ¥3 trillion in absolute value per year. Data are from the Thomson Reuters Lipper Global Fund Flows database.

[4] On November 19, 2014, the BoJ started buying also ETFs tracking the JPX-Nikkei 400 Index. This approximately corresponds to 43% of the purchases flowing to ETFs tracking the TOPIX, 53% to ETFs tracking the Nikkei 225 and the remaining 4% to ETFs tracking the JPX Nikkei 400. For simplicity, in the empirical analysis we round the share of both TOPIX and Nikkei ETFs to 50%, neglecting the JPX Nikkei 400. This simplification does not affect the results of our analysis.

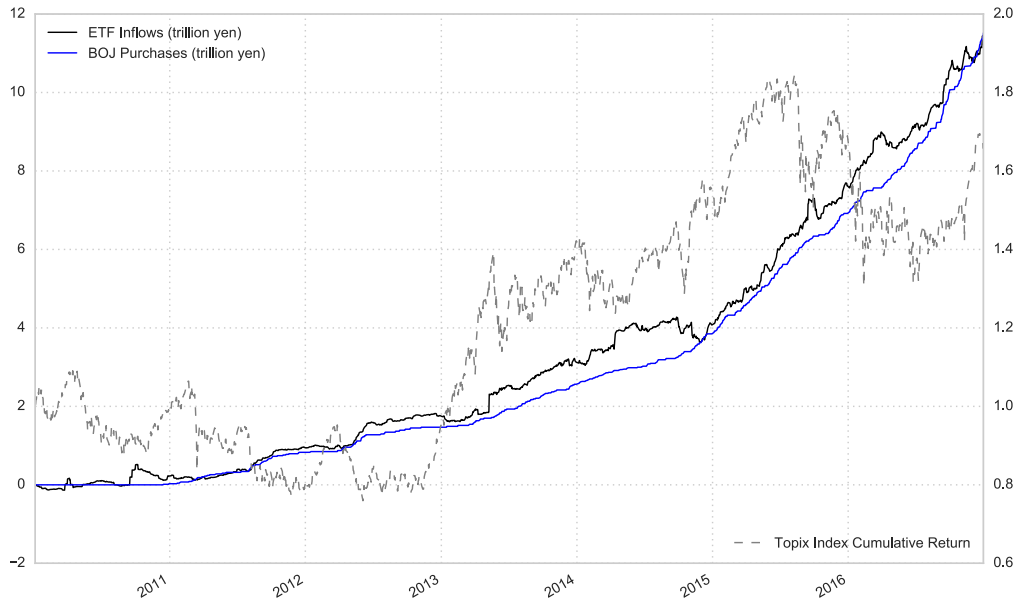


Figure 3: BoJ Purchases and ETF inflows. On the left axis we plot the daily cumulative purchases of ETFs by the BoJ (blue line) and the estimated daily cumulative inflows into ETFs tracking either the Nikkei or the TOPIX index (black line). Both are in trillion yen. On the right axis the figure shows the cumulative return from 2010 of the TOPIX index (gray dashed line).

cross-section of supply shocks to identify the causal impact of the purchase program on equity prices. We rely on a simple asset pricing model to argue that the deviation from a value-weighted allocation allows us to isolate the portfolio-balance channel of the policy impact. Section 5.2 discusses the identification strategy in detail.

Overall, the portfolio of the BoJ ends up deviating significantly from the allocation that market capitalization would dictate. To illustrate the extent of this distortion, take three companies with fairly similar market capitalization and therefore similar TOPIX weights (between 0.45% and 1% in 2014): Canon, Fast Retailing and Nintendo. Canon and Fast Retailing are both among the Nikkei constituents, though with very different weights, namely around 1.2% versus 9.5%, respectively. Nintendo, on the contrary, is not included in the Nikkei index. It follows that the BoJ allocates to Fast Retailing 4 times more capital than to Canon, and 19 times more than to Nintendo. The effects of the departure from a value-weighted allocation are reflected in the indirect ownership that the BoJ accumulated over time. According to estimates by the Financial Times, through its purchases the central bank has indirectly become the largest shareholder in a quarter of TOPIX stocks. In Table 1 we report the ten stocks with the highest estimated indirect ownership share by the BoJ.

We argue that purchases of ETFs by the BoJ translate into supply shocks at the individual stock level. This is a consequence of the creation-redemption mechanism in the ETF primary

market and the physical replication of the underlying basket. When demand exceeds supply in the ETF secondary market, new shares of ETF are issued to keep the ETF price close to its NAV. In the case of physical ETFs, creation requires the physical purchase of the basket of securities that composes the tracked index, for a value equal to the creation unit. Securities are then held by the ETF sponsor on behalf of the owner of the ETF shares, who now bears the associated risk. ETF creation thus reduces the quantity of assets available for trading in the underlying market. This mechanism is visualized in Figure **??** in Appendix **??**. Given the direct correspondence, in the rest of the paper we will consider ETF purchases by the central bank equivalent to an intervention in the underlying equity market.

We can infer whether central bank purchases triggered creation of new ETF shares from data on ETFs AUM. We first obtain the list of the ETFs listed on the Tokyo Stock Exchange (TSE) that track either the Nikkei or the TOPIX index from the website of the Japan Exchange Group (JPX). We then get daily data on AUM for each ETF from Bloomberg. We estimate inflows simply as the difference between the actual increase in AUM and the increase in AUM due to the return on the index that the ETF is tracking. Figure 3 plots the time-series of ETF inflows versus the amount purchased by the BoJ. It is apparent that the flows into these ETFs are almost completely due to the asset purchase program. In turn, this implies that the purchases by the BoJ have consistently triggered creation of new ETF shares.

It must be noted that the bias towards Nikkei companies did not go unnoticed among practitioners and the BoJ was frequently accused by the financial press of distorting the market. In response to the criticism, on September 21, 2016, the BoJ amended the terms and conditions of the program and announced it will change the maximum amount of each ETF to be purchased. Since October 2016, the BoJ allocates ¥2.7 trillion a year (US$ 26.4 billion) to TOPIX ETFs, while the remaining ¥3 trillion are spread out between the TOPIX, the Nikkei 225 and the JPX-Nikkei Index 400. For the Nikkei-ETFs this means a drop from 55% to about 25% of the annual purchases by the BoJ, which brings the allocation of the flows closer to what market capitalization would justify. Yet, the accumulated balance sheet of the BoJ remains tilted away from a value-weighted allocation.

3 Related Literature

"Extraordinary times call for extraordinary measures", stated the Chairman of the Federal Reserve Ben Bernanke in 2009 (Bernanke, 2009). Since then, a number of central banks around the world have adopted unconventional monetary policy tools and most of them have been trying to support asset prices through LSAPs in order to boost economic activity in the face of severe dislocations in financial markets. With actual data on the implementation of LSAPs becoming available, a large body of academic research has investigated their impact

Company Name	BOJ Share (%)	BOJ Flow (bn JPY)	Market Cap (bn JPY)	Nikkei weight (%)
Mitsumi Electric Co Ltd	10.3	5.8	56.1	0.17
Advantest	8.9	27.5	309.1	0.63
Fast Retailing	8.7	336.4	3854.7	9.17
Taiyo Yuden	7.8	10.0	129.0	0.32
Toho Zinc	7.7	3.3	43.2	0.09
Tdk Corporation	7.4	71.0	959.0	1.41
Konami Holding	7.2	37.6	524.5	0.65
Trend Micro	7.0	36.2	514.9	0.90
Comsys Holding	6.6	18.3	275.7	0.39
Nissan Chem In	6.2	30.6	489.7	0.53
Average	6.1	3.8	255.6	0.44
Median	5.9	0.2	43.8	0.20

Table 1: BoJ indirect shareholdings. Summary statistics on indirect ownership by the BoJ for the ten companies with the highest BoJ share. BoJ Flow are the cumulative compounded BoJ purchases at company level since the beginning of QQE and Market Cap is the company's market capitalization. BoJ Share is the ratio of BoJ Flow and Market Cap. Average and median values are calculated over the universe of TOPIX firms. The values in the first three columns are as of August 31, 2016. The last column reports the average company weight in the Nikkei 225 index over the study period. Notice that the ten companies with the highest BoJ share have all positive weights in the Nikkei 225 index.

on financial markets and the real economy.

Most of the work on the impact of QE on market prices relies on evidence from purchases of government bonds by the Fed, the ECB or the BOE, and usually shows a significant impact on yields (Buraschi and Whelan, 2015, D'Amico and King, 2013, Eser and Schwaab, 2016, Gagnon et al., 2010, Hamilton and Wu, 2012, Joyce et al., 2012, Krishnamurthy and Vissing-Jorgensen, 2011, 2013, Neely et al., 2010, Swanson, 2011). There is however little empirical evidence on the large-scale purchases of the BoJ. Perhaps closest to our paper is Ueda (2013), who looks at the time series of LSAP announcements by the BoJ and finds a positive correlation with the TOPIX index and the yen-dollar exchange rate. The BoJ is the first central bank to purchase domestic equities as part of its QQE agenda, and, to the best of our knowledge, this paper is the first to study this program in depth and to analyze its impact on the cross-section of stock prices.

Although there is general agreement that LSAPs do indeed affect prices, there is less consensus regarding the channels through which these policies work. A standard explanation in the literature is the so-called portfolio balance channel (Brunner and Meltzer, 1973, Frankel, 1985,

Tobin, 1969). According to this channel, when the central bank buys a particular asset, it reduces the amount held by private investors, effectively forcing them into a different portfolio. For this to be an equilibrium, prices need to adjust to ensure market clearing. In particular, through this channel asset purchases are expected to push up the price of the target asset and of its substitutes, implying that demand curves are downward sloping. Some papers find that the observed price impact is consistent or partially consistent with portfolio balance explanations (e.g. D'Amico and King (2013), Gagnon et al. (2010), Joyce et al. (2011))[5]. However, the portfolio balance channel of monetary policy is subject of debate, in part because standard asset pricing models do not generally allow exogenous changes in the supply of a security to affect its price. For instance, Miles and Schanz (2014) argue that LSAPs by central banks since 2008 had significant effects because markets were dysfunctional and that in normal times portfolio-balance effects would be weak.

The question whether demand curves for stocks slope down has a long tradition in the asset pricing literature. The empirical evidence so far mostly comes from event studies around index redefinitions and fire sales by institutional investors (Coval and Stafford, 2007, Greenwood, 2005, Harris and Gurel, 1986, Hau et al., 2009, Mitchell et al., 2004, Petajisto, 2009, Scholes, 1972, Shleifer, 1986b, Schnitzler, 2016). The general finding is that large non-fundamental trades have a significant but temporary price impact, even though there is considerably heterogeneous evidence on the speed and the extent of reversal. The standard interpretation is that limits to arbitrage can justify temporary deviations from fundamental value: under market efficiency, uninformed shocks cannot have a long-lasting impact on prices.[6]

Quantitative easing provides an ideal laboratory in which to test asset pricing theories such as the long-held belief of flat demand curves for stocks. However, from the success of QE in pushing up prices alone, one cannot conclude much about the elasticity of demand curves. A large number of papers show that LSAPs by central banks have effects beyond those due to portfolio balance, and provide evidence of alternative transmission channels that are consistent with flat demand curves. For the case of purchases of long-term bonds, Krishnamurthy and Vissing-Jorgensen (2011) provide compelling empirical evidence that the so-called signalling channel explains a significant fraction of the drop in bond yields observed after the Federal Reserve's QE announcements. The idea behind this channel is discussed in Eggertsson and Woodford (2004), who claim that financial markets may interpret LSAPs as signals about the

[5] Vayanos and Vila (2009) try to reconcile the predictions of the portfolio balance channel with the observed lack of spillovers across maturities, building on market segmentation and preferred-habitat theories as proposed by Culbertson (1957) and Modigliani and Sutch (1966).

[6] The traditional view in finance is that, in a frictionless world, a simple expansion of the balance sheet of the central bank should have no effect. This neutrality result is formalized in Eggertsson and Woodford (2004) and crucially relies on the assumption of a rational infinitely lived agent with no credit restrictions, who sees no difference between its own assets and those held by the central bank.

central bank's intention to keep interest rates low, thus influencing long-term yields through investors' expectations about the future path of interest rates. Other papers attribute the beneficial effect of the Fed's MBS purchases on risk premia during the financial crises to a capital constraints channel motivated by the distress in the financial intermediary sector (Curdia and Woodford, 2011, He and Krishnamurthy, 2013).

In general, the identification of the impact of market interventions through a specific channel is a challenging task. Our paper contributes to this literature proposing a new identification strategy for the transmission channel of monetary policy and providing new insights on the elasticity of demand curves. Moreover, the results of the empirical literature suggest that the specific workings of LSAPs depend on the asset purchased and the economic conditions under which these purchases take place. We complement the existing evidence by documenting the effects of the ETF program by the BoJ, a unique case in which a central bank is targeting the equity market.

4 The Model

In this section we develop a theoretical framework to describe the portfolio balance channel as the transmission mechanism from LSAP to asset prices. The idea is that asset purchases shift part of the fundamental risk from the market to the balance sheet of the central bank. Because the premium demanded for a given security is proportional to its marginal risk contribution to the aggregate portfolio held by the representative agent, the price effect of the monetary intervention is proportional to the implied change in this quantity. Therefore, the net effect on asset prices through this channel is not simply proportional to the purchased amounts, but it crucially depends on the correlation structure of firms fundamentals.

Our model features the central bank only in reduced form, in the sense that the policy rule is exogenous. We also assume that asset purchases are deterministic. This assumption holds also when we allow investors to believe that the central bank will deviate from the announced purchase target. With no policy uncertainty, asset purchases do not represent a source of risk that has to be priced in equilibrium. Moreover, we assume firms fundamentals to be neutral with respect to monetary policy, excluding the possibility that asset purchases affect market prices through the change in future investment opportunities. We make these choices to keep the model simple and to focus on the direct effect of supply on prices. These assumptions also allow us to restrict our attention to the covariance-stationary equilibrium of the model, which immediately follows once we assume covariance-stationary dividends. The limitation is that the model abstracts from potential additional channels related to uncertainty about future supply and endogenous responses of firms.

4.1 Model Setup

Consider an economy with n risky assets in fixed supply $Q = (Q^1, \ldots, Q^n)$, paying dividends in every time period. The dividend $D_{i,t}$ paid at time t is

$$D_{i,t} = D_{i,0} + \sum_{s=1}^{t} \varepsilon_{i,s}, \qquad \forall i \in 1, \ldots, n \tag{1}$$

where each $\varepsilon_{i,t}$ is revealed at time t. The fundamental innovations $\varepsilon_{i,t}$ are modelled as zero-mean jointly normal random variables, iid over time.

The representative agent optimally chooses her time-t demand N_t to maximize her next period expected utility, subject to a standard budget constraint

$$\max_{N} \ \mathrm{E}_t \left(-\exp(-\gamma W_{t+1})\right) \tag{2}$$

$$\text{s.t.} \qquad W_{t+1} = W_t(1+r) + N_t'(p_{t+1} + D_{t+1} - p_t(1+r)) \tag{3}$$

where W_t is the total wealth, N_t' denotes the transpose of the vector N_t and γ the aggregate risk-aversion. At date $t = 1$ the central bank announces share purchases described by the vector $q = (q^1, \ldots, q^n)$, distributed over M periods after the announcement. We refer to M as the policy horizon. Let q_t denote the vector of cumulative purchases by the central bank up to date t. One can think of q_t as the active side of the balance sheet of the central bank at any time t.

We assume, first, that $q_t = tq$ for $t = 1, \ldots, M$ and, second, that $q_t = Mq$ for $t > M$. The first assumption implies that in our model the central bank's balance sheet evolves deterministically and grows linearly over time. Assuming non-stochastic asset purchases allows us to abstract from policy uncertainty as a priced risk factor and to focus on how QE affects prices through the change in supply [7]. The second assumption implies that the central bank never unwinds its position nor engages in further purchases beyond horizon M. This assumption might be restrictive once we go to the data since the BoJ never announced such a stringent commitment. Still, given that the BoJ position have not been unwound (and neither announced to be so) over the window of our empirical analysis, we believe it to be a reasonable benchmark.

The realized demand shocks negatively affect the net supply of assets in each period. Setting $Q_0 = Q$ yields

$$Q_t = Q - q_t \tag{4}$$

[7] The assumption that the central bank spreads its purchases equally over the policy horizon is instead innocuous. Relaxing this assumption does not improve the economic intuition and only adds technical complexity to the model.

Asset purchases by the central bank affect the quantity at which the equity market clears given the equilibrium condition $N_t = Q_t$. Notice that equation (4) also implies that the quantity of assets available to the market can only change through purchases of the central bank. This excludes the possibility for companies to respond endogenously to changes in prices by issuing new stocks or buying back those outstanding.

The central bank buys the vector q of securities in exchange for cash. We assume that the representative agent invests the proceeds in the risk-free asset and, since risk-free returns are uncorrelated with those of Japanese equities, omitting the risk-free asset from the model does not change the predicted policy impact on stock prices. This assumption may be interpreted as a form of market segmentation, in that the representative agent cannot re-invest the proceeds in assets outside the Japanese equity universe. We discuss the implications of this assumption in Section 6.4. The fact that we do not model other asset classes that equity investors might hold in their portfolios does not affect the model predictions even in case of non-zero correlation with equities. It is easy to show that including securities that are not targeted by the asset purchase program has no effect on the predicted price impact on stock prices. Stock purchases will spillover to correlated asset classes, but in this paper we are not interested in these effects.

Appendix **??** shows that the pricing equation of the covariance-stationary equilibrium, in matrix notation, is given by

$$p_t = \frac{1}{r}\left(D_t - \gamma V \Omega_t\right) \tag{5}$$

where $V \equiv \mathrm{Var}_t(p_{t+1} + D_{t+1})$ is the stationary covariance matrix of asset returns and Ω_t is the vector of time-t expected future asset supply, properly discounted by time, defined as

$$\Omega_t \equiv \frac{r}{1+r} \sum_{i=0}^{\infty} \frac{\mathrm{E}_t[Q_{t+i}]}{(1+r)^i} \tag{6}$$

Notice that the covariance matrix V is not time-varying because the supplies of each asset are fixed and the schedule of purchases by the central bank is deterministic, so there is no uncertainty on future shocks to the asset supply. Equation (6) shows that at any point in time prices reflect the path of future asset supply. Given the time-discounting, today's prices are less sensitive to quantities further into the future.

Staring at the vector of risk premia $\gamma V \Omega_t$ in the pricing equation (5) one can see that, for each stock, priced risk is an increasing function of the stock's covariance with the market portfolio and the risk aversion parameter γ. The vector $V\Omega_t$ admits an interpretation very similar to the CAPM beta and should be thought of as a measure of systematic risk.[8] This is easier to see in the absence of monetary policy shocks, in which case $V\Omega_t$ reduces to VQ.

[8] While the model is written in price changes, market betas are usually defined in terms of returns. In Appendix **??** we derive an expression of systematic risk that determines expected returns in the model. While the notation becomes messier, the intuition carries through.

By plugging equation (4) into equation (5) it is easy to see how the portfolio balance mechanism works in the model. Asset purchases change the amount of each security in the market clearing portfolio. This affects systematic risk and in turn prices. Notice that this change in systematic risk is fully consistent with our assumption of a constant covariance matrix V, since what determines systematic risk is the product $V\Omega_t$, and central bank purchases affect only the latter term in the model.

Let's now turn to how the representative agent builds expectations about future asset supply, before and after the purchase program is announced. These expectations enter the pricing equation (5) and so determine the impact of the policy. At time $t > 0$, the representative agent's expectation about the quantity in period $h = t+1, \ldots, M$ is given by

$$E_t[Q_h] = E_t[Q_0 - q_h] = Q - E_t[q_h] \tag{7}$$

We assume that the central bank intervention is fully unexpected at $t = 0$ before the announcement, i.e. $E_t[q_h] = 0$ for every $t \leq 0$ and $h \geq t$. At each period $t \geq 1$ after the announcement date, we allow the investor to believe that the central bank will deviate from its purchase target. We restrict to a family of investors beliefs parametrized by a time-varying scalar λ_t. Formally, let $\lambda_t \geq 0$ be real numbers such that

$$\mathrm{E}_t\left[Q_h\right] = Q - \lambda_t q_h, \qquad t \geq 1 \tag{8}$$

The parameter λ_t is assumed to change over time in a deterministic fashion. The path of λ_t determines how the representative agent updates her beliefs regarding the size of the purchase program. We first solve the model for a general mapping $t \mapsto \lambda_t$ and then we present, in the next section, results for the special case $\lambda_t \equiv 1$, in which the pricing equation takes a simpler form that better conveys the intuition for the portfolio balance channel. Imposing $\lambda_t \equiv 1$ is equivalent to assuming that the representative agent expects the central bank to commit to the announced target exactly. This implies that she also expects the central bank to never unwind its positions and to never engage in additional purchase programs in the future. Changing the parameter λ_t allows us to study how deviations from this benchmark case impact the effect of the policy.

As shown in Appendix **??**, the model's pricing equation predicts price changes given by

$$p_t - p_{t-1} = \frac{1}{r}\left(\varepsilon_t + \gamma\xi(t)Vq\right) \tag{9}$$

where the function $\xi(t)$ is defined piece-wise as follows

$$\xi(t) = \begin{cases} 0 & \text{if} \quad t \leq 0 \ \text{ or } \ t > M \\ \lambda_1(M - \varphi(1)) & \text{if} \quad t = 1 \\ \Delta\lambda_t M - (\lambda_t\varphi(t) - \lambda_{t-1}\varphi(t-1)) & \text{if} \quad 1 < t \leq M \end{cases} \tag{10}$$

and $\varphi(t) < M$, defined in Appendix **??**, is a deterministic function of time representing the residual duration of the purchase program.

In the first part of equation (10), $\xi(t) = 0$ implies that both before the announcement ($t \leq 0$) and after the purchase program has been completely carried out ($t > M$), price changes only reflect shocks to dividends and are therefore unpredictable. The functional form of $\xi(t)$ in the second and third pieces of the domain determine event ($t = 1$) and post-event price changes ($1 < t \leq M$), respectively. Even when future supply changes are fully predictable and the average path of future prices can be perfectly anticipated ($\lambda_t \equiv 1$), the shock to supply is impounded into stock prices immediately after the policy announcement only up to the term $\varphi(1)$. Prices will then continue to adjust in the following days. The reason why prices do not fully adjust on the event day is that until purchases are actually realized at future dates, the representative agent bears dividend risk and requires a compensation for it. Consistent with this intuition, $\varphi(t)$ is decreasing in t and increasing in M. So, even though $\xi(t)$ drives predictable post-event price changes, these are fully consistent with market efficiency and do not represent an arbitrage opportunity.

The relative magnitude of the initial price reaction and the subsequent adjustments depend on the level and the dynamics of λ_t. More specifically, the price jump at $t = 1$ is increasing in the initial expectation of future supply λ_1 since prices are effectively responding to a purchase program of size $\lambda_1 Mq$. Post-event price changes are then linked to the time series evolution of λ_t. An increasing λ_t over time means that the agent is revising upward her expectations about the size of the program. One can think of different reasons for why this might happen. For example, the agent may not immediately believe that the central bank will commit to the full size of the program and thus update her expectations only once she observes the purchases actually being carried out. Or, she may start believing over time that the central bank will engage in additional purchases beyond the announced policy horizon M. Similarly, a decreasing λ_t means that the agent revises downward the expected size of the program, either because she starts to believe that the central bank will not complete the announced program or that it will unwind the portfolio soon after. In the Internet Appendix we show simulations of the price dynamics implied by different functional forms for λ_t.

Even though we mainly think of λ_t as controlling the agent's beliefs on the central bank actions conditional on time-t information, this reduced-form suits a number of non-mutually exclusive interpretations. For instance, as in Barberis and Thaler (2003), the slow reaction may be due to the bounded rationality of agents who fail to correctly process the consequences of the BoJ announced program.

4.1.1 Benchmark Case

In this section we focus on the special case where $\lambda_t \equiv 1$, which implies that expected and realized purchases are the same at any point in time

$$E_t[q_h] = q_h = h\,q \tag{11}$$

It follows directly from equation (9) that the price adjustment at $t = 1$ is given by

$$p_1 - p_0 = \frac{1}{r}\left(\varepsilon_1 + \gamma V(Mq - \varphi(1)q)\right) \tag{12}$$

Ignoring fundamental innovations, equation (12) predicts a positive price jump of magnitude $\gamma V(Mq - \varphi(1)q)$. This swing in prices is due to the fact that the policy is unexpected at $t = 0$, but it is impounded into prices as soon as it is revealed.

In the following periods ($t \geq 1$), price changes are instead given by

$$p_{t+1} - p_t = \frac{1}{r}\left(\varepsilon_{t+1} - \gamma V(\varphi(t+1) - \varphi(t))q\right), \qquad t = 1, \ldots M \tag{13}$$

Equation (13) shows that price changes in the post-announcement period include a non-stochastic component $\gamma V(\varphi(t+1) - \varphi(t))q$ which accounts for the time delay between the announcement of the supply shocks and their realizations. Since the cross-sectional distribution of these predictable price adjustments is always parallel to and of the same sign as the initial price impact, they add up to create a propagation (drift) of the initial cross-sectional effect.

4.2 Testable Predictions

In this section we derive testable predictions from the model. To make these predictions more suitable to be tested in the data, we state them in terms of returns. In order to go from the expressions in price changes derived in Section 4.1 to predictions about returns, we first need to introduce some new notation. We define u as the vector of yen amount purchased by the BoJ of each security, so that

$$u_i \equiv p_{i,t}\, q_i, \qquad \forall i \in 1 \ldots, n \tag{14}$$

where q_i is the number of shares purchased of stock i and $p_{i,t}$ the stock price at time t. Then, we define Σ to be the stationary covariance matrix of stock returns, i.e.

$$\Sigma_{i,j} \equiv \mathrm{Cov}(R_i, R_j), \qquad \forall i, j \in 1 \ldots, n \tag{15}$$

where R_i is the daily percentage return of stock i.

Dividing equation (5) by p_{t-1} leads to the following two propositions about event returns and post-event returns. Proofs are in Appendix **??**.

Proposition 1 (Event returns)**.** *The vector of returns $R_1 = (p_1 - p_0)/p_0$ on the announcement day is positively related to the vector $\pi \equiv \Sigma u$ in the cross-section.*

Proposition 2 (Post-event returns)**.** *Assume $\Delta\lambda_{t+1} = \lambda_{t+1} - \lambda_t \geq 0$. Then the vector of post-event returns R_{t+1} is positively related to $\pi = \Sigma u$ in the cross-section for every $t = 1, \ldots, M$. Moreover, the vector of expected cumulative returns is given by*

$$\sum_{s=1}^{t} \mathrm{E}[R_s] = \theta_t \pi \tag{16}$$

where $\theta_t = \frac{\gamma}{r} \sum_{s=1}^{t} \xi(s)$ is a positive and increasing function of t, which follows from the definition of $\xi(t)$ in equation (10).

Proposition 1 states that through the portfolio balance channel, the policy announcement leads to abnormal event returns proportional to the change in systematic risk captured by the vector $\pi = \Sigma u$. Notice that if the central bank were to buy stocks proportionally to their market weight, abnormal event returns would be proportional to the product of Σ and the vector of market capitalizations, i.e. the vector of each stock's covariance with the market portfolio. Proposition 1 therefore implies that an exogenous shock to supply parallel to the market portfolio would cause price adjustments proportional to market betas. At the same time, it also implies that shocks to supply that are orthogonal to market capitalization produce abnormal returns orthogonal to market betas. This prediction is key to identify the effect of the policy shock from the cross-section of realized event returns in the empirical part of the paper.

As summarized in Proposition 2, the model predicts post-event returns in the same direction of event returns, i.e. proportional to π, until the purchase target is met at $t = M$. This generates a post-event drift, whose magnitude depends both on the value of the risk free rate and the beliefs dynamics parametrized by λ_t. In the Internet Appendix, we show analytically and from model simulations that for realistic value of the risk free rate and λ_t constant, this drift is small. The model produces a more pronounced drift under the assumption that the representative agent revises her expectations on the size of the program over time ($\Delta\lambda_{t+1} > 0$).

From Proposition 2 it follows that a permanent change in the supply of assets generates a permanent change of risk premia, and hence of prices. Unless the central bank unwinds its positions, prices will not revert to the pre-event level.[9] By stating that changes in supply can

[9]Notice that a reversal would be observed as soon as investors update the expected path of future supply

have long-lasting impacts on prices, the proposition implies downward sloping demand curves for stocks through the portfolio balance mechanism.

5 Data and Empirical Methodology

5.1 Data Sources

From Compustat Global we collect stock-level data on daily returns, volumes and shares outstanding for the roughly 2000 stocks of the TOPIX universe for the period 1990-2016. Daily returns and volume data for the TOPIX index as well as the monthly time-series of TOPIX and Nikkei 225 index weights for every stock in our sample are obtained from Thomson Reuters Datastream. The USD/JPY exchange rate is from Japan Macro Advisors Inc. The time-series of ETF purchases by the BoJ is publicly available at daily frequency on its website.

5.2 Identification Strategy

To test the model predictions from Proposition 1 and Proposition 2 we estimate the following cross-sectional regression at different horizons H around the two policy announcements made by the BoJ

$$R_{i,e}^{H} = \alpha_e + \beta_e^H \pi_{i,e} + \delta_e' W_{i,e} + \eta_{i,e} \tag{17}$$

where R_i^H is the cumulative return of stock i computed over H days from the event day and W is a matrix of stock-level observed covariates. The estimation of the vector π is described in the next section. All variables are event specific and therefore indexed by the subscript $e \in (2014, 2016)$. Regression coefficients are also indexed by the event because we estimate the model separately for the two announcements.

The coefficient of interest β^H measures the portfolio balance effect of the policy and is identified from the cross-sectional heterogeneity of the model-implied change in systematic risk π. Notice that β^H has a similar interpretation as the coefficient on the interaction term in a diff-in-diff estimation where π measures the intensity of the treatment.

Following Proposition 1, if stock returns respond to the exogenous shock to supply through the mechanism described in the model, we expect $\hat{\beta}^H$ to be positive and significant at short horizons. As our baseline specification of the short-run effect, we choose $H = 10$ days.

to include a sale of the portfolio of the central bank ($\Delta\lambda_t < 0$). Also, we would observe a reversal if the central bank was to surprise the market by ceasing the purchases before reaching the expected target. Since we do not provide any empirical evidence of an exit from LSAP, we do not formalize this scenario into a proposition.

Proposition 2 implies a positive and significant coefficient at any horizon H. We therefore look at $\hat{\beta}^H$ estimated from a regression of cumulative returns over longer horizons (one month, three months, six months and one year) on π. Estimating a positive $\hat{\beta}^H$ at short horizons followed by a lower $\hat{\beta}^H$ at longer horizons would indicate that the initial event return is, at least partially, reversed after some time. Such evidence would be inconsistent with the portfolio balance channel described by our model and would rather suggest a temporary price pressure story, where arbitrageurs with limited capital need some time to absorb the demand shock coming from the central bank.

Proposition 2 also implies that $\hat{\beta}^H$ should be found to be weakly increasing in H. An increasing $\hat{\beta}^H$ indicates that the divergence in the cross-section of returns in the direction of the vector π not only does not vanish, but it becomes larger with time.

In the recent literature on the effect of monetary policy, the common approach to address endogeneity concerns is to employ high-frequency data and to focus on very short windows around policy announcements (e.g. Andrade and Ferroni, 2018, Cochrane and Piazzesi, 2002, Hanson and Stein, 2015, Nakamura and Steinsson, 2018). Estimating long-run effects is challenging in this type of studies that rely on time-series variation for identification due to increasing exposure to confounding factors when moving away from the time of the announcement. Since our identification relies on the cross-section of returns, we can expand the post-event window to estimate the long-run effect of the policy and say something about the elasticity of long-run demand curves for stocks.

Our empirical framework relies on a set of assumptions to conclude that the observed returns are due to the supply shock induced by the policy.

First, to allow for a causal interpretation of the coefficient on π we need the OLS exogeneity assumption to be satisfied. The cross-section of asset purchases would be endogenous if, for instance, the central bank were to overweight underpriced stocks and underweight overpriced ones. Clearly, in this case one could not claim a causal link from the supply shock to the observed subsequent returns. Endogeneity is a concern in papers that try to pin down portfolio balance effects in the case of interventions by the Federal Reserve in Treasury markets where the purchases were likely targeting particularly illiquid or underpriced securities. Since the BoJ is buying ETFs tracking market indices, concerns about the endogeneity of the purchases at stock level are mitigated.

Second, in order for the expected future change in supply to be correctly and immediately impounded into price, we need to assume that the purchase schedule is known by the market. If, instead, there was uncertainty about what securities the central bank would purchase, event returns would react to these expectations rather than to the actual change in supply. Unlike other asset purchase programs by the Federal Reserve or the ECB, the BoJ announced

it would buy according to a well defined and predetermined rule. Since index weights are public information, this ensures that the vector q of supply shocks at stock level is known at the announcement.

Proposition 1 from the model shows that a supply shock parallel to market capitalization would produce returns proportional to stock market betas in the cross-section. Since in the model monetary policy affects prices *only* through the portfolio balance channel, those returns are by definition caused by the change in assets' supply. In reality, asset purchases can affect stock prices through channels other than portfolio balance, possibly in proportion to their market beta. In such case, using the market model as our benchmark, we would not be able to identify the portfolio balance effect of the policy in the cross-section of returns. Our empirical strategy relies on the model prediction that a purchase schedule orthogonal to market valuations would leave a characteristic footprint in the cross-section of abnormal returns computed against a market model. We are able to perform this test empirically since the BoJ is tilting its purchases away from market capitalization. Our identification strategy therefore relies on a third assumption, namely that there is no transmission mechanism of monetary policy other than portfolio balance that would affect prices proportionally to π.

Finally, our specification assumes that the theory-implied measure π of predicted price impact reflects the change in systematic risk correctly. Recall that π is derived in the model under the assumption of market segmentation, i.e. that agents keep the proceeds from the sale of the securities in cash. We discuss violations of this assumption in the Internet Appendix.

5.3 Variable Construction and Summary Statistics

This section presents the data and defines the empirical proxies for the vector u of expected purchases and of the covariance matrix Σ of asset returns, defined in Section 4.2.

To be conservative, in Section 6 we test the model predictions separately for the two policy announcements of the BoJ. All variables are therefore calculated or estimated twice, in order to have two sets of variables, one for each event. For all estimated variables in our analysis we use an estimation window of one year, ending two trading weeks before each BoJ announcement.

5.3.1 Expected Purchases

In the guidelines to the LSAP program, the BoJ states that it would spread its purchases among index-tracking ETFs proportionally to the aggregate AUM of each ETF. In practice, this roughly corresponds to a 50-50 allocation of capital between TOPIX and Nikkei ETFs.

When we go to the data, we assume that this allocation rule not only holds on aggregate over the policy horizon, but also each time the central bank makes a purchase. Under this assumption, the vector u of purchases by the BoJ (in yen) can then be expressed as

$$u_i = Tw_{i,T} + Nw_{i,N} \tag{18}$$

where T and N indicate the amount of BoJ capital allocated to TOPIX and Nikkei ETFs, respectively, and $w_{i,T}$, $w_{i,N}$ are the weight of stock i in the TOPIX and Nikkei indices.

Since $T \cong N$ in the current purchase program, for the empirical analysis we compute the vector u simply as $w_{i,T} + w_{i,N}$. The vector u proxied in this way still encodes the cross-sectional variation in purchases at the heart of our identification strategy. Given that index weights are time varying, for each event we take $w_{i,T}$ and $w_{i,N}$ to be the index weights as of the end of the month preceding the announcement.

Figure **??** in Appendix **??** shows in the top row of each panel, the cross-sectional distribution of stock weights in the TOPIX and the Nikkei 225 index in the month before the event. The top-right panel plots the cross-sectional distribution of the resulting stock-level weights in the BoJ purchase vector $(w_{i,T} + w_{i,N})$. The percentile plots in logarithmic scale clearly show that variation in weights across stocks is substantial.

5.3.2 Covariance Matrix

We estimate the variance-covariance matrix Σ of stock returns using daily returns data from Compustat Global. Because the cross-sectional dimension of our data is larger than the sample size, the sample covariance matrix of returns is a poor estimator of Σ. We therefore use the shrinkage method proposed by Ledoit and Wolf (2004) to obtain a well-conditioned and more accurate estimator, which also ensures that the resulting matrix is always positive definite.

In the model described in Section 4 returns are driven only by fundamentals innovations and changes in supply. However, when we go to the data, this assumption may not hold. We are especially concerned about the impact on the returns moments of other monetary policy announcements during the estimation window. To address this concern, we look at stock returns net of market returns and we estimate Σ as the cross-sectional covariance of the fitted residuals $\hat{e}_{i,t}$ from a simple market model specified as

$$R_{i,t} = \alpha_i + \beta_i^{mkt} R_{mkt,t} + e_{i,t} \tag{19}$$

where $R_{i,t}$ are daily returns of stock i and $R_{mkt,t}$ is the return on the TOPIX Index used as proxy for the market portfolio. As reported in Table **??** of the Appendix, our results are robust to estimating Σ on raw returns rather than abnormal returns.

5.3.3 Control Variables

We estimate stocks' sensitivities to changes in the exchange rate by running the following regression separately for each stock i

$$R_{i,t} = \alpha_i + \beta_i^{mkt} R_{mkt,t} + \beta_i^F F_t + e_{i,t} \tag{20}$$

Here F_t is the daily percentage change in the exchange rate from US Dollar to Japanese Yen. Estimation results for market and Forex betas are reported in Table **??** and Figure **??** in Appendix **??**. In the bottom rows we plot the cross-sectional distributions of pre-event market betas and of Forex betas together with companies' market capitalization. Table **??** in Appendix **??** presents a break down of the summary statistics by Nikkei and non-Nikkei companies.

6 Empirical Results

In this section we test the empirical predictions of the model described in Section 4. Our results show that the ETF program of the BoJ had a significant impact on stock prices and that both the cross-sectional and time-series patterns of the price effect are consistent with a portfolio balance channel. We first perform event studies around the two BoJ announcements and show that the observed price impact is positively related at the stock-level with our ex-ante measure of systematic risk change. We then look at the effect over different horizons and conclude that the impact of the policy is persistent and increasing in time. We propose a simple back-of-the-envelope calculation to quantify the net aggregate portfolio balance effect of the implemented program. We estimate a 22 basis points increase in aggregate market valuation per trillion Yen invested, which corresponds to a unitary price elasticity.

6.1 Event Study

Proposition 1 states that we should observe a positive relationship between each security abnormal event return and the change in its marginal contribution to the risk of the aggregate portfolio.

As a preliminary test of this relationship we rank stocks in the TOPIX universe by the predicted abnormal event return $\pi_i = (\Sigma u)_i$ into four equally-weighted portfolios. Figure 4 presents cumulative returns of the low and high π portfolios. Plots on the left show the event returns around the first policy change in 2014 (when the target purchase amount of ETFs was tripled), while those on the right present the effect of the second change in 2016 (when the

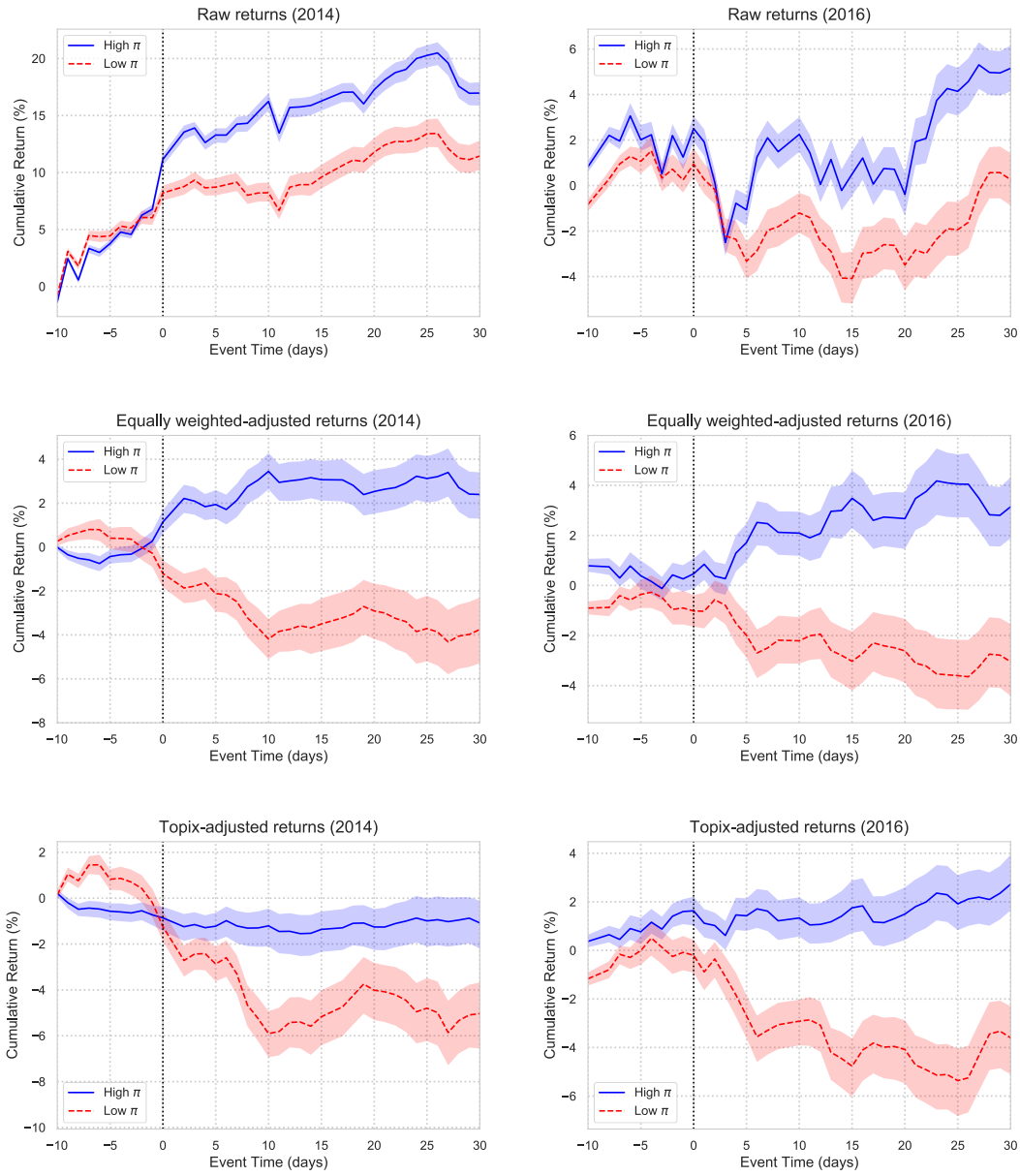


Figure 4: Cumulative returns of high versus low π stocks (in percentage). This figure shows the time series of the mean cumulative returns around the BoJ announcements of stocks with high predicted price impact π against that of low π stocks. The plots on the left refer to the announcement on October 31st, 2014, while those on the right show the reaction to the announcement on July 29th, 2016. The two top panels plot the unadjusted returns. In the four remaining panels returns are adjusted using a market model estimated in a window of one year, as described in Section 5.3.3. An equally-weighted portfolio of stocks in the TOPIX universe is used a proxy for the market portfolio in the middle panels, while the return of the TOPIX index is used in the bottom panels. The blue line is the average for the first quartile of the distribution (firms with the highest predicted price impact), while the red dashed line corresponds to the average for the last quartile (firms with the lowest predicted price impact). Bands represent bootstrapped 95% confidence intervals.

target was doubled further). We consider raw returns and abnormal returns based on two versions of the market model with different proxies for the market portfolio, the TOPIX index and an equally-weighted index, respectively. The reported bands represent bootstrapped 95% confidence intervals.

Each plot shows a sizeable and highly significant spread between the returns of high and low π firms opening after the two announcements. While for the 2014 event the reaction seems to be slightly anticipated, in 2016 the effect is delayed by a couple of days. Overall, the pattern of abnormal returns is similar for the two events, with the performance of the high π portfolio being significantly higher than that of the low π portfolio. There is no sign of reversal over 30 days after the announcement, and rather the gap between the two groups appears to increase over time. This preliminary evidence is consistent with both predictions of the model.

6.2 Cross-Sectional Regressions

One might be concerned that, by sorting on π, we are implicitly ranking stocks based on firms' characteristics such as size, export share or market beta, which might explain the heterogeneous response to the announcements and thus the divergence in returns. We therefore run security-level cross-sectional regressions of event returns on the predicted price impact π_i and a set of control variables

$$R_i^H = a_0 + a_1\,\pi_i + a_2\,u_i + a_3\,\log(cap_i) + a_4\,\beta_i^{mkt} + a_5\,\beta_i^F + a_6\,\text{Amihud}_i + \eta_i \tag{21}$$

For the purpose of these regressions, event returns are defined as the cumulative returns computed over the 10 trading days following the announcement ($H = 10$). We control for each security's weight in the purchase schedule of the BoJ (u), the natural logarithm of its market capitalization, its market beta, its Forex beta and its Amihud ratio as a proxy for illiquidity.

The Internet Appendix **??** reports summary statistics of the control variables by quartile of π. Consistent with the fact that the policy is heavily skewed towards Nikkei companies, which are on average larger than non-Nikkei ones, we find a positive correlation between π and market capitalization. Market capitalization is therefore an omitted variable in a regression of stock returns on π and we need to control for it. Moreover, the policy announcement could affect equity prices through its impact on the foreign exchange market. Since π is weakly negatively correlated with the Forex exposure β^F, if the yen depreciated as a consequence of the announcement, we would spuriously observe returns proportional to π. We therefore control for the exposure to the exchange-rate by adding β^F to the regressions. We also control for market betas. Notice though that there is no obvious relationship between π and market betas. We include the weights of the BoJ purchase schedule u_i to control for

Panel A: October 31st, 2014

	Raw Returns				Abnormal Returns			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
π	57.86***	59.15***	31.92***	23.27***	36.06***	37.75***	39.95***	30.60***
	(8.59)	(8.00)	(4.40)	(3.49)	(4.94)	(4.81)	(5.54)	(4.62)
u		-0.00	-0.02***	-0.02***		-0.00	-0.02***	-0.02***
		(-0.95)	(-3.83)	(-3.43)		(-1.34)	(-3.62)	(-3.20)
Market Beta			0.040	0.025			-0.05	-0.07*
			(0.75)	(0.51)			(-1.48)	(-1.97)
Forex Beta			0.040*	0.043**			0.040*	0.041**
			(1.87)	(2.32)			(1.96)	(2.32)
log(Market Cap)			0.007	0.007			0.005	0.006
			(1.17)	(1.36)			(0.97)	(1.16)
Amihud			0.000	0.000			0.000	4.618
			(0.53)	(0.49)			(0.22)	(0.03)
Observations	1,851	1,851	1,807	1,701	1,851	1,851	1,807	1,701
R-squared	0.106	0.106	0.162	0.203	0.046	0.047	0.108	0.160
Industry FE	NO	NO	NO	YES	NO	NO	NO	YES

Panel B: July 29th, 2016

	Raw Returns				Abnormal Returns			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
π	14.07*	14.30*	12.17*	11.88	15.33*	16.33*	17.49**	16.69**
	(2.09)	(1.93)	(1.68)	(1.78)	(2.10)	(2.08)	(2.43)	(2.52)
u		-0.00	0.001	0.002		-0.00	0.004	0.004
		(-0.29)	(0.19)	(0.38)		(-1.33)	(0.56)	(0.67)
Market Beta			0.006	-0.00			0.002	-0.01
			(0.12)	(-0.06)			(0.06)	(-0.26)
Forex Beta			0.016	0.012			0.019	0.014
			(0.78)	(0.66)			(0.94)	(0.80)
log(Market Cap)			-0.00	-0.00			-0.00	-0.00
			(-0.10)	(-0.05)			(-0.58)	(-0.49)
Amihud			0.000	0.000			0.000	0.000
			(0.43)	(0.18)			(0.41)	(0.11)
Observations	1,905	1,905	1,839	1,734	1,905	1,905	1,839	1,734
R-squared	0.017	0.017	0.021	0.043	0.019	0.019	0.028	0.050
Industry FE	NO	NO	NO	YES	NO	NO	NO	YES

Table 2: Cross-sectional regressions. The tables report the regression coefficients of the cross-sectional regression of returns (in percentage points) on the predicted price impact π and a set of control variables (standardized). Regressions are run separately for the two events. The dependent variable in columns 1-3 is the cumulative raw return, while in columns 4-6 is the cumulative abnormal return with respect to the market model estimated in the pre-event window. Cumulative returns are computed over a 10 days horizon after the announcement date. t-statistics from placebo regressions are in parenthesis; asterisks denote conventional significance levels (***=1%, **=5%, *=10%) based on empirical p-values.

alternative explanations based on the direct effect of purchases in which LSAPs affect asset prices proportional to the amount purchased. Finally, we include industry fixed effects to some specifications, to make sure that our results hold within industries.

We run these regressions on the entire universe of TOPIX firms. Panel A of Table 2 investigates the cross-sectional effect of the BoJ announcement on October 31, 2014 (when the target purchase amount of ETFs was tripled), while Panel B analyzes event returns following the announcement on July 29, 2016 (when the target was further doubled). In either cases, no change was made to the weighting scheme of the purchases. In the first four columns the dependent variable is the cumulative raw return of the stock, while in the last columns the left-hand side variable is the cumulative abnormal return from a market model calculated using pre-event market betas.

On a given day, stock returns are expected to be correlated in the cross-section and therefore the OLS assumption of iid residuals is likely to be violated. We therefore run placebo regressions on the period from January 2009 to March 2013 to get the empirical distribution of the coefficients in the absence of policy shocks, which we use to compute robust standard errors. The placebo event days are chosen randomly on non-overlapping periods to ensure that the empirical distribution is constructed from independent draws. For regressions involving short-horizon returns (up to 3 months) we impose that placebo event periods do not include BoJ meetings on which important monetary policy announcements were made. Namely, we exclude the meetings of February 1st 2013, March 25th 2013, June 18th 2012 and the announcement of the post-tsunami intervention in March 14th 2011. On all regression tables of this paper we report the empirical p-values computed using this methodology.

Consistent with Proposition 1, the coefficient on the predicted price impact π is positive and significant across specifications and events. For the 2014 announcement, the baseline specification with raw returns reported in columns (1) shows a remarkable R^2 above 10%, suggesting that our expected price impact π is crucial to explain the heterogeneity of event returns. As it was already visible from the plots in the previous section, the results are weaker for the 2016 event. In particular, the portion of explained variance for the 2016 policy announcement is significantly lower, consistent with the smaller change in the target purchase amount. Still, the coefficient on π is positive and significant at the 10% confidence level in most specifications. The coefficient turns however insignificant when we include industry fixed effects in the specification with raw returns.

Results show that the effect of π is robust to the inclusion of the vector u of purchased amounts. This horse race provides additional support for the portfolio-balance channel against a local channel where spillovers are negligible. The model predicts that the effect of u should be insignificant once we control for π. This is indeed what we find in the second specification of each panel. The coefficient on u turns however negative and significant in columns (3) and (4)

of panel A. We find that this is due to the inclusion of the control for market capitalization since u and market cap are highly correlated, as it is natural to expect. Still, this does not affect the size and the significance of the coefficient on π.

Results also show that controlling for the exposure to the exchange rate does not impair the significance of the coefficient on π. The coefficient on β_F is positive and significant in 2014, when the BoJ announcement was followed by a rise in the Forex. In 2016, on the other hand, the coefficient on β_F is not significant, consistent with the fact that the Forex did not move significantly (see Figure **??** in the Appendix).

In column (2) of the regression using raw returns as dependent variable, the coefficient on π drops significantly. This is, as expected, due to the fact that control variables play an important role in explaining cross-sectional returns variation, as documented by a significantly larger coefficient of determination. In particular, market beta, Forex beta and market capitalization incrementally increase the regression's R^2 and dampen the coefficient on π. In columns (3) the number of observations drops slightly because of missing data on trading volume needed to estimate the Amihud ratio. In columns (4) the sample is further reduced because of missing information on industry classification.

6.3 Time-Series Pattern

In this section, we test the long-run predictions of the model summarized in Proposition 2. To this end, we estimate the cross-sectional model specified in equation (21) at different horizons H over which cumulative returns are calculated. Results are reported in Table 3.

	Abnormal Returns 2014						Abnormal Returns 2016					
	5	10	21	63	126	252	5	10	21	63	126	252
π	17.78**	39.95***	28.31***	72.21***	171.9***	305.5***	17.43**	17.49**	33.70***	40.86**	93.52***	120.2***
	(3.09)	(5.54)	(2.89)	(4.52)	(8.97)	(11.91)	(3.03)	(2.43)	(3.44)	(2.56)	(4.88)	(4.69)
u	-0.00	-0.02***	-0.01	-0.03***	-0.03**	-0.07***	0.011**	0.004	0.026**	0.016	0.052***	0.118***
	(-0.30)	(-3.62)	(-1.21)	(-3.20)	(-2.46)	(-3.43)	(2.22)	(0.56)	(2.76)	(1.48)	(3.31)	(5.13)
Market Beta	-0.02	-0.05	-0.07	-0.16*	-0.29**	-0.40***	0.026	0.002	0.025	0.025	0.041	0.068
	(-0.84)	(-1.48)	(-1.36)	(-2.02)	(-2.04)	(-2.51)	(0.86)	(0.06)	(0.47)	(0.32)	(0.28)	(0.43)
Forex Beta	0.036**	0.040*	0.101***	0.097**	0.043	-0.13	-0.00	0.019	0.061*	0.070*	0.220***	0.216
	(2.17)	(1.96)	(3.83)	(2.29)	(0.61)	(-0.02)	(-0.23)	(0.94)	(2.30)	(1.67)	(3.10)	(0.04)
log(Market Cap)	0.001	0.005	0.001	0.001	0.005	-0.00	-0.00*	-0.00	-0.01*	-0.02	-0.07***	-0.11***
	(0.32)	(0.97)	(0.17)	(0.06)	(0.25)	(-0.27)	(-1.60)	(-0.58)	(-1.81)	(-1.80)	(-3.52)	(-6.82)
Amihud	0.001	0.000	-1.02	0.001	0.016***	0.018***	0.000	0.000	-0.00	-0.00	-0.00**	-0.00
	(0.78)	(0.22)	(-0.01)	(0.70)	(6.27)	(7.17)	(0.47)	(0.41)	(-0.47)	(-0.92)	(-3.12)	(-2.16)
Observations	1,807	1,807	1,807	1,807	1,807	1,807	1,839	1,839	1,839	1,839	1,839	1,839
R-squared	0.055	0.108	0.073	0.098	0.153	0.119	0.051	0.028	0.079	0.077	0.178	0.140

Table 3: Cross-sectional regressions over different horizons. The table report the coefficients of cross-sectional regressions of cumulative returns (in percentage points) computed at different horizons on the predicted price impact π and a set of control variables (standardized). Regressions are run separately for the two events. The dependent variable is the cumulative abnormal return with respect to the market model estimated in the pre-event window. t-statistics from placebo regressions are in parenthesis; asterisks denote conventional significance levels (***=1%, **=5%, *=10%) based on empirical p-values.

At portfolio level, Figure 4 suggests that the cross-sectional effect of the BoJ announcements is long-lasting and weakly increasing over time. The regression analysis confirms that the evidence holds at stock-level and after controlling for security specific characteristics. The vector π is positively and significantly related to cross-sectional stock returns at every horizon H after the announcement. In other words, the model implied changes in systematic risk estimated ex-ante are a significant predictor of post-event abnormal returns across stocks.

We find no evidence of reversal of the initial price impact even one year after the event. The positive coefficients on π at longer horizons provide evidence of a persistent effect of the purchase program on prices. This result is consistent with the portfolio-balance channel, whereby a permanent reduction in the stock of assets held by private investors leads to a permanent decrease in risk premia.

Claims about the persistence of the effect of QE are hard to make in event studies where identification relies on time-series evidence over a short window around the announcement. The main concern in those type of studies is that long-lasting effects might be due to changes in expectations about the future path of policy rates (signalling channel), and more generally that results might be confounded by the release of macro news in the subsequent days. Since the shock to supply induced by the BoJ has a unique cross-sectional shape, it is unlikely that the observed effect is due to shocks other than the purchase program, suggesting a causal effect of the policy.

The absence of reversal is a key prediction of the portfolio-balance channel. Still, we cannot completely rule out other explanations. While price pressure and limits to arbitrage would generally predict a temporary effect, the fact that the purchase program is ongoing over the entire sample period might prevent us to observe a reversal. In Section 7, we address the possibility that (part of) the effect might be due to a continuous price pressure that prevents prices from reverting to the pre-announcement level.

A second finding from Table 3 is that the estimated coefficients on π are generally increasing in H.[10] This suggests that the effect of the BoJ policy is not immediately reflected on prices, but it is increasingly impounded over time. Post-event returns in the same direction of the announcement effects are predicted by the model through the decrease in residual duration of the program. However, as we discuss in Section 4, this effect is expected to be small for realistic values of the interest rate. The significant post-event abnormal returns observed in the data are consistent with investors expectations about the size of the program increasing over time. In terms of the model, the results are consistent with a λ_t increasing in t. This suggests that investors might be extrapolating current purchases above and beyond the policy horizon or that they might not believe to a full commitment of the central bank to the announced

[10]Notice that the cumulative returns on the left-hand side of the regression are computed as cumulative sums rather than cumulative products in order to avoid a mechanical effect when increasing the horizon.

purchase target at first, but slowly update their beliefs. In the current setting we cannot disentangle between these explanations, nor convincingly claim that post-event returns are in fact driven by updating in beliefs. The question is therefore open for future research.

In Appendix **??** we present some robustness evidence. We show that the observed price impact cannot be explained by industry effects in Table **??**, which presents regression results including industry fixed effects. To show that the difference in returns is not simply driven by an over-performance of Nikkei stocks versus non-Nikkei stocks, in Table **??** we re-run the analysis including a dummy variable for Nikkei stocks. Finally, Table **??** includes both industry and Nikkei fixed effects. The results remain largely unchanged across specifications.

6.4 Quantification of Portfolio-Balance Effects

In this section we propose a simple back-of-the-envelope calculation to try to quantify the net aggregate portfolio balance effect of the BoJ intervention from the coefficient estimated in the cross-section. From this quantity we then derive an estimate of the aggregate elasticity of equity demand curves.

For this calculation we want to use the average effect of the policy across the two events, so we first run again our main regression in equation (21) over the pooled sample, including event fixed-effects FE_e to allow for a different intercept across the two announcements. Precisely, we estimate the following regression model for each daily horizon $h \in 1, \ldots, 252$

$$R^h_{i,e} = \beta^h \pi_{i,e} + \gamma^h X_{i,e} + \delta^h \,\mathrm{FE}_e + \varepsilon_{i,e} \tag{22}$$

where X is a vector of control variables that depends on the regression specification and $e \in (2014, 2016)$ is an index numbering the events. Since we are considering both events together, we need to rescale the π vectors to take into account the different magnitude of the announcements. Therefore we multiply π_{2014} by 3 and π_{2016} by 6 to reflect the magnitude of the target amount announced by the BoJ in the two events, respectively. We include market capitalization in each specification to control for the size factor, which is expected to become more relevant as the horizon increases. In the second specification we also control for market and Forex betas. In the third specification we additionally include each stock's Amihud ratio to control for liquidity.

Given $\hat{\beta}^h$ from the estimation, the predicted net return through the portfolio balance channel for security i is $\hat{R}^h_{i,e} = \hat{\beta}^h \pi_{i,e}$ [11]. To aggregate the effect at market level, we calculate for each

[11] Notice that the estimated $\hat{\beta}^h$ allows us in principle to compare the impact of alternative purchase schedules u' that the central bank could have implemented, conditional on the same covariance matrix Σ. In this section we are interested in the estimated portfolio balance effect of the actual purchase portfolio.

	1 week	1 month	3 months	6 months	1 year
(1) Baseline	3.54	10.10	10.28	25.17	22.32
(2) Control for market and Forex	2.23	7.72	9.53	22.08	22.45
(3) Control for market, Forex and liquidity	1.56	7.17	9.56	19.80	22.05

Table 4: Portfolio Balance Effects. The table presents the estimated net portfolio balance effect on the market, expressed in basis points per Trillion Yen invested by the central bank into the ETF purchase program. We report point estimates for the net effect impounded into prices over increasing horizons, from three models employing different sets of control variables defined in the text.

event $e \in (2014, 2016)$ the predicted market return as the value-weighted sum of security level predicted returns at every horizon

$$\hat{R}_e^h = \hat{\beta}^h \sum_i w_{i,e} \pi_{i,e} \tag{23}$$

We then divide by the capital commitment by the central bank to obtain the induced market return per trillion yen. Considering the two-year policy horizon, this amounts to 6 trillion Yen for 2014 and 12 trillion Yen for 2016, with the underlying assumptions that each announcement was completely unexpected and that investors are reacting to the announced program size. Thus, the per yen estimated average market return induced by the policy through the portfolio-balance channel is calculated as

$$\hat{R}^h = \frac{1}{2}\left(\hat{R}_{2014}^h/6 + \hat{R}_{2016}^h/12 \right) \tag{24}$$

Results of this exercise are reported in Table 4 for the three specifications. The last column shows an estimated long-term impact of about 22 basis points increase in market value per trillion yen employed. With about ¥500 trillion of total market capitalization, this implies an elasticity close to one since each yen invested translates into an increase of the market valuation by roughly one yen.

Figure 5 plots the time-series evolution of the point estimate for the third specification, showing that the portfolio balance effects are slowly impounded into prices. Consistent with the qualitative prediction of our model, a momentum-like pattern is visible over the first 100 trading days following the announcement.

Notice that the quantity we are estimating in this section is the aggregate portfolio-balance effect of the policy on the returns of stocks that are included in the TOPIX index. While the policy is expected to have additional effects through different channels, our empirical methodology allows us to identify and quantify the portfolio balance channel. In turn, this allows us to derive an estimate of the price elasticity of the demand curve for stocks, which has to be intended as local to the First Section of the TSE. We acknowledge that the policy

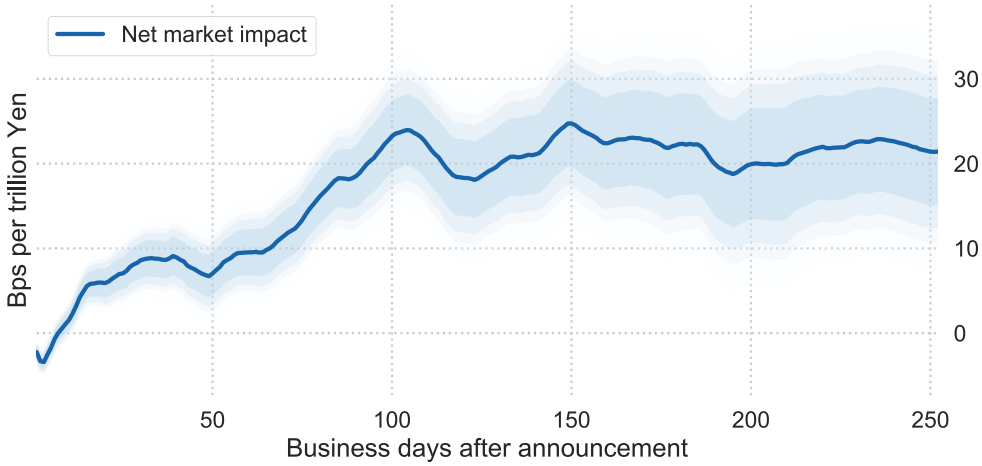


Figure 5: Portfolio Balance Effects. This figure plots the time-series evolution of the estimated portfolio balance effect induced by the BoJ purchase program, expressed in basis points per trillion Yen invested. The estimates are based on specification (3), which includes controls for stocks liquidity, market beta and exposure to the US-JPN Forex exchange rate. Thus the estimated market impact can be interpreted as the counter-factual policy effect, net of of alternative channels and confounding factors. Shaded areas denote 10%, 5% and 1% confidence intervals.

might have produced spillover to other asset classes (this is indeed a prediction of the model), but we are not considering them in this paper.[12]

The derivation of the aggregate portfolio balance effect described in this section relies on two main assumptions. First, it depends on our assumption that the representative agent in our model re-invests the proceeds from the sale of stocks to the central bank at the constant risk-free rate.[13] The key point here, is that the risk-free asset is uncorrelated with the Japanese stocks. While this is not a concern for the identification of the portfolio-balance effect, it is more problematic when we try to quantify the effect, since our approach might be providing a biased estimate of the aggregate effect if the assumption is not valid.

To understand the direction and magnitude of the potential bias, in the Internet Appendix we extend the model to allow the representative agent to re-invest the proceed in a security correlated with the targeted assets. We derive an expression of the resulting bias if we incorrectly assume the above mentioned assumption, which can be seen as a form of market

[12] Spillovers to unaffected stocks is already a key point in Greenwood (2005) and is to be expected in this setting as well. We believe that the cross-sectional heterogeneity among the stocks in the TOPIX is sufficient to support our arguments. Moreover, the TOPIX index covers all First Section companies in the Tokyo Stock Exchange (TSE), which are the majority of public companies in Japan and is by far largest section of the TSE in terms of market capitalization and trading volume.

[13] We thank the Anonymous Referee for pointing out this issue.

segmentation. The bias is a function of the market weighted average of the covariance of the omitted variable and π, where the omitted variable is the vector of covariances between the re-investment security and the stocks. The sign of the bias is ambiguous and depends on Σ, u, the market weights and the re-investment security. We therefore run simulations of the model using the parameters estimated in the data and assuming different re-investment securities, namely S&P500, 10-year US Treasury bonds and 10-year JGBs. The estimated bias is positive using the S&P500 and negative using long-term government bonds. The magnitude of the bias is relatively small, around (positive or negative) 10%. Depending on which direction the bias is going, the estimated elasticity of 1 might be slightly over- or under-estimating the true elasticity of Japanese equities.

A second reason why our approach might be delivering a biased estimate of the aggregate effect is that in our calculation we are assuming that the market is reacting to the announced size of the program. If the market is in fact reacting to expectations of a smaller program, either because investors think the BoJ will not reach the announced target or because it will soon unwind its portfolio, then the estimated elasticity of 1 represents an upper-bound for the true price elasticity. Vice versa, the estimated elasticity of 1 would be a lower-bound if investors believe that the BoJ will continue the purchase program beyond M.

7 Portfolio Rebalancing or Price Pressure?

The previous section shows that the reaction of stock prices to the upward revisions of the purchase target is consistent with a portfolio balance channel. Even at long horizons, the expected change in systematic risk is key to explain the cross-sectional variation in returns after the policy announcement. We interpret the persistence of the effect as evidence of downward sloping long-run demand curves.

An alternative explanation for the observed persistence relies on the continued pressure exercised by the BoJ through repeated purchases. If short-run demand curves are downward sloping, abnormal volumes induced by the BoJ during intervention days might push prices above fundamentals. Such effects are usually motivated by limits-to-arbitrage and are expected to revert quickly. However, as the central bank is expected to buy repeatedly, arbitrageurs may refrain from betting against mispricings and fail to bring prices back to fundamentals. If this was the case, the absence of reversal could not be interpreted as evidence for long-run demand curves sloping down. In the spirit of D'Amico and King (2013), we will refer to this kind of dynamics as *flow effect* of the policy. This naming highlights that under this explanation the price impact is caused directly by the trading volume (or *flow*) of the central bank, rather than by reduction in systematic risk which underlies the portfolio balance channel.

The repeated price pressure story implies that we should observe higher positive abnormal returns on intervention days and that these should be proportional in the cross-section to the abnormal trading volume generated by the intervention. In this section we first introduce a reduced form model that exploits the time-series and cross-sectional variation in daily purchases by the BoJ to estimate the flow effect of the policy. We then use the predicted returns from that model to remove the flow effect component from stock returns. Finally, we re-run the analysis of Section 6.2 on these *net* returns. By comparing the coefficient estimated in this way to the one in the previous section, we can assess how much of the observed price impact and its persistence is due to repeated price pressure rather than the portfolio balance mechanism.

Evaluating the relative magnitude of these two channels is essential to draw conclusions on the elasticity of long-run demand curves for stocks. The distinction between the two explanations has also practical consequences for policy makers regarding the exit strategy from the purchase program. A repeated price pressure story predicts prices to revert as soon as the buying pressure from the central bank stops, making the accumulated size of the balance sheet *de facto* irrelevant beyond that point. On the contrary, in the model of Section 4 the aggregate impact of the policy is unaffected by the timing of the purchases. In the extreme case where the central bank buys everything on the announcement day, the model predicts an immediate, complete and permanent price adjustment.

7.1 Purchase Frequency and Volumes

The QQE was announced on April 4, 2013, and the asset purchases were then gradually carried out. In its official statements, the BoJ does not commits itself to any particular purchase frequency and does not reveal in advance the days in which it will buy. Ex-post, we can see from Panel A of Figure 6 that the bank has been buying fairly consistently once to twice a week over the sample period. The blue crosses in Panel B of Figure 6 indicate intervention days. On the y-axis we report the ratio between the amount purchased and the aggregate trading volume in the underlying stock market on that day.

Since the purchase frequency remained stable over the policy horizon, the upward revisions of the annual target in October 2014 and in July 2016 translated into an increase of the daily purchased amount. However, even in the last period, the quantity purchased by the BoJ represented less than 5% of the daily market volume, a threshold which is often used by practitioners as guideline for when a trade is expected to have a significant price impact. At the stock level, the purchases by the BoJ account for more than 5% of the daily trading volume on average only for 5.2% of the targeted stocks.

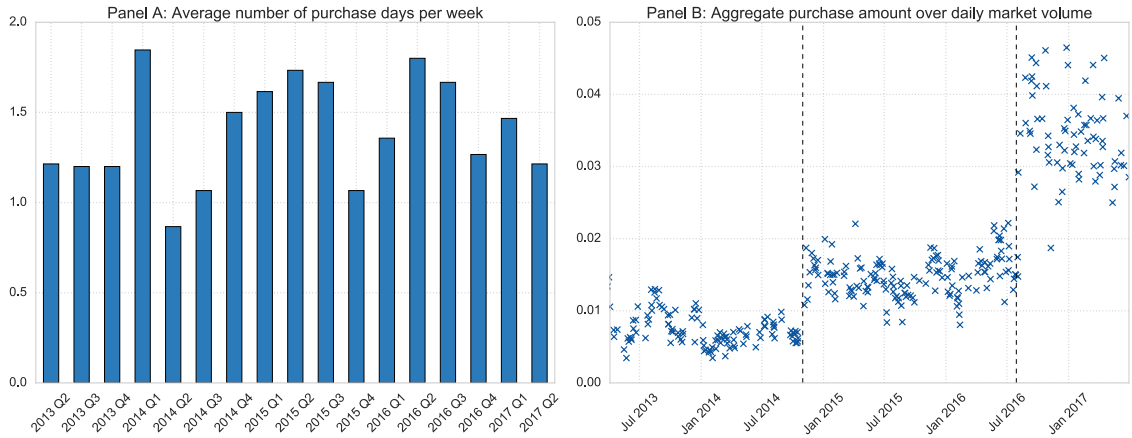


Figure 6: Purchase Frequency and Volume. Panel A plots the average number of purchase days per week at quarterly frequency. Panel B plots the ratio between the yen amount purchased by the BoJ on a given day and the aggregate trading volume in yen on that day. The aggregate trading volume is computed as the sum of the trading volume of the securities targeted by the policy.

7.2 Empirical Setup and Results

To quantify the direct price impact of purchases we estimate a dynamic model in the spirit of Eser and Schwaab (2016) that relates daily stock returns to daily flows from the central bank. The model is specified as

$$AR_{i,t} = \alpha + \beta_0 AV_{i,t} + \beta_1 AV_{i,t-1} + \beta_2 \left(\sum_{k=2}^{K} \rho^{k-2} AV_{i,t-k} \right) + \varepsilon_{i,t} \tag{25}$$

where the left-hand side variable $AR_{i,t}$ is the daily abnormal return of stock i relative to the market model, estimated following the methodology outlined in Section 5.3. On the right-hand side, the BoJ-induced abnormal volume $AV_{i,t}$ is defined by

$$AV_{i,t} := \frac{\text{BoJ Flow}_{i,t}}{\text{E}\left[\text{Volume}_{i,t}\right]} \tag{26}$$

and measures the size of the purchased amount of stock i on day t relative to the average market volume of that stock. The purchased amount BoJ $\text{Flow}_{i,t}$ is computed as $\frac{1}{2}(w_{i,T} + w_{i,N})A_t$, where $w_{i,T}$ is the weight of stock i in the TOPIX index, $w_{i,N}$ is the weight of stock i in the Nikkei 225 index and A_t is the value of ETFs purchased by the BoJ on day t. Here we assume that each trade of the BoJ in the ETF market translates into proportional shocks to the underlying basket on the same day.[14] The average daily volume $\text{E}\left[\text{Volume}_{i,t}\right]$ is estimated

[14] Underlying securities inherit shocks that occur in the ETF market both through primary market arbitrage as well as through the arbitrage that takes place continuously in the secondary market and that is carried

over a backward-looking window of six months excluding days in which the BoJ is intervening. On non-purchase days the abnormal volume is therefore zero for every stock in our sample, while it is strictly positive on purchase days.

The model includes lagged values of AV to capture the permanent component of the price pressure, net of transitory and delayed effects of purchases. The long-run effect of the flow-induced price impact can be computed from the estimated coefficients as

$$F = \beta_0 + \beta_1 + \beta_2 \left(\sum_{k=2}^{K} \rho^{k-2} \right) \tag{27}$$

The parameter $\rho \in (0, 1)$ determines how long it takes for prices to adjust following an intervention. If ρ is close to zero the dynamic of the flow effect is exhausted after two days. $F \approx 0$ implies that temporary price impacts, if any, are fully reverted. This in turn would mean that the price pressure story does not contribute to explain the persistence of the policy impact documented in Section 5. On the contrary, $F > 0$ implies that (part of) the persistence attributed to the portfolio balance mechanism might be due to the direct impact of the flow of BoJ purchases.

The identification of the direct impact of the purchases (flow effect) in this panel regression framework relies both on the exogeneity of the cross-sectional variation of the purchases and on the predetermination of the purchase amounts with respect to prices. The exogeneity in the cross-section is discussed extensively in Section 2 and mainly relies on the fact that the weighting system of the Nikkei 225 introduces significant variation in the cross-section of purchases that is unrelated to firms' fundamentals. Predetermination of the purchases is not straightforward in the current context. The criteria used by the BoJ to decide whether and how strongly to intervene on a particular day are not public information, however there are reasons to believe that the BoJ tends to intervene on days when the market is falling. In fact, the median stock return is significantly lower on intervention days (-0.6%) relative to non-intervention days (0.3%). To tackle this potential endogeneity of BoJ flows we specify the regression model in terms of abnormal returns. Given that the BoJ might be using the return on the market as a signal for whether to intervene, removing the contemporaneous return on the market should mitigate the issue. The fact that mean and median abnormal returns are

out by hedge funds and high-frequency traders (Ben-David et al., 2018). Secondary market arbitrageurs make profits by opening their positions when the price of the ETF deviates from NAV and holding them until prices converge. Our identification of the flow effect of the Policy relies on arbitrageurs trading on the same day as the BoJ. For secondary market arbitrage, this is a reasonable assumption. Competition among arbitrageurs implies that hedge funds and high-frequency traders will open their positions as soon as they observe the ETF trading at a premium over the NAV. That such arbitrage opportunities exist on the days when the BoJ buys is consistent with the evidence in Figure 3, since growth in AUM is consistent with upward pressure on ETF prices. The results reported in Table 5 provide further support for the validity of this assumption since they show that most of the price impact seem to take place on the event day and the day after.

		Panel A					Panel B		
Model	K	β_0	β_1	β_2	ρ	F	$\tilde{a}_1$	Flow Effect	Port Balance
(1)	0	0.011				0.011	138.644	5.63%	94.37%
		(9.356)					(11.175)		
(2)	1	0.004	0.015			0.019	132.288	9.96%	90.04%
		(3.314)	(11.358)				(10.662)		
(3)	2	0.004	0.015	-0.001		0.018	132.608	9.74%	90.26%
		(3.358)	(10.979)	(-0.679)			(10.688)		
(4)	5	0.004	0.015	-0.001	0.001	0.018	132.567	9.77%	90.23%
		(3.396)	(10.997)	(-0.694)	(0.035)		(10.685)		
(5)	10	0.004	0.015	-0.001	0.001	0.018	132.567	9.77%	90.23%
		(3.396)	(10.997)	(-0.694)	(0.039)		(10.685)		

Table 5: Flow Effect The table reports results from the estimation of the dynamic model described in (25), where a different value for the number of lags K is used in each specification. The models are estimated with maximum likelihood assuming normally distributed error terms and constraining the persistence parameter ρ in the unit interval. Panel A presents the estimated model parameters and the implied long-run effect F. Panel B shows OLS estimates of the coefficient $\tilde{a}_1$ resulting from a cross-sectional regression of cumulative abnormal returns, purified from the estimated flow effects, on the predicted price impact π resulting from the portfolio balance model of Section 4. The decomposition into *flow* and *portfolio balance* components is obtained by comparing $\tilde{a}_1$ with the coefficient a_1 from Section 6.2 based on standard CARs.

not significantly different from zero in both intervention and non-intervention days supports our claim.

We do not include an announcement dummy in the specification because on those days no ETF purchases were made by the BoJ. Looking at the time series of BoJ purchases, we see that the bank intervened two weeks before and one week after the first upward revision of the purchase target on October 31, 2014. Similarly, no purchases were made on July 29, 2016. Purchases are registered on the previous day and four days after.

We estimate five different specifications of model (25). We start considering only contemporaneous volumes ($K = 0$), then we augment the specification to K equal to $1, 2, 5$ or 10. The estimated parameters are reported in Panel A of Table 5 together with the implied long-run impact. The positive coefficients on β_0 and β_1 suggest that abnormal returns are significantly higher during purchase days for stocks experiencing a higher degree of buying pressure. The negative but not significant value of β_2 and a persistence parameter ρ close to zero suggest that such a price impact is not reverted in the next trading weeks and give rise to a positive long-run component F in every specification.

The results indicate a positive and persistent flow effect of the policy, which might lead to an overestimation of the portfolio balance channel in the previous section. To quantify the

consequences of not taking flows into account, we construct the flow induced returns as the fitted values of the estimated model

$$\widehat{AR}_{i,t}^{Flow} = \hat{\beta}_0 AV_{i,t} + \hat{\beta}_1 AV_{i,t-1} + \hat{\beta}_2 \left(\sum_{k=2}^{K} \hat{\rho}^{k-2} AV_{i,t-k} \right) \tag{28}$$

which we subtract from stock returns to remove the direct impact of the BoJ purchases

$$\widetilde{AR}_{i,t} = AR_{i,t} - \widehat{AR}_{i,t}^{Flow} \tag{29}$$

We then estimate our main regression (21) using $\widetilde{AR}$ instead of AR, computing the cumulative abnormal returns over a one-year horizon following the two event dates and we regress them on the predicted price impact vector π. We pool the 2014 and 2016 events together to obtain a unique estimate $\tilde{a}_1$ of the price impact of the policy through the portfolio balance channel. The ratio between $\tilde{a}_1$ and its counterpart $\hat{a}_1$ obtained estimating the model with the cumulative returns computed from AR, gives us the fraction of the estimated portfolio balance impact that might be explained by the price pressure channel.

Panel B of Table 5 summarizes the results of this second step, showing that the fraction of the observed cross-sectional pattern explained by the price pressure channel ranges between 5% and 10% depending on the specification. It must be noted that these figures represent upper bounds for the persistent flow effect of the policy, since this might be amplified by expectation updates consistent with the portfolio balance model, if investors learn about the commitment of the central bank through the realization of its purchases.

Taken together, the results of this section suggest that the price pressure generated by the central bank at the stock level plays a limited role in explaining the impact of the policy. We conclude that the observed cross-sectional pattern of stock returns is mostly generated by the portfolio balance channel rather than continued price-pressure arising from the central bank flows.

8 Policy Implications

In this section we discuss the policy implications of our results. Based on our theoretical framework, we show formally that the heterogeneity uncovered by our empirical analysis could be avoided if the central bank would buy the value-weighted market portfolio, since this would lead to a homogeneous reduction of firms' cost of capital in the cross-section.

Recall that in our model the cost of capital of each firm is proportional to its marginal risk contribution to the market portfolio (systematic risk). Formally, the vector of risk premia prior

to the BoJ intervention is proportional to VQ, where V is the variance-covariance matrix of fundamentals and $Q \in \mathbb{R}^n$ is the vector of shares outstanding.

As soon as the central bank purchases a quantity $q \in \mathbb{R}^n$, the cost of capital is affected and converges to $V(Q - Mq)$. In particular, firm i experiences a percentage shift in its perceived cost of capital equal to

$$\Delta k_i = \frac{(V(Q - Mq))_i}{(VQ)_i} - 1 \tag{30}$$

Notice that Δk_i is not necessarily negative, thus some firms may experience an increase in their financing costs ($\Delta k_i > 0$), even if the central bank buys some of their shares ($q_i > 0$).

It follows that a homogeneous impact on risk premia can be achieved with a vector of purchases proportional to Q. If the purchase schedule is $q^* = aQ$ for $a \in \mathbb{R}$, the effect on firm i is

$$\Delta k_i^* = \frac{(V(Q - Mq^*))_i}{(VQ)_i} - 1 = \frac{((1 - Ma)VQ)_i}{(VQ)_i} - 1 = \frac{((1 - Ma)VQ)_i}{(VQ)_i} - 1 = -Ma \tag{31}$$

which does not depend on i and is thus homogeneous across companies.

In the case of Japan, a purchase schedule q parallel to Q corresponds to the BoJ limiting its purchases of ETFs to those tracking the value-weighted TOPIX Index[15]. Buying ETFs tracking the price-weighted Nikkei 225, on the other hand, introduces a component in q which is orthogonal to Q. This, in turn, leads to heterogeneous consequences for firms financing costs, which can be interpreted as a distortion of the market allocation mechanisms. Figure **??** shows that the distortion is evident also at the industry-level.

Under the assumption that a homogeneous effect is the preferred outcome of the policy, we infer from the model that the central bank should stop buying Nikkei-indexed ETFs. More precisely, the central bank should schedule future purchases with the objective of re-shaping its equity portfolio in a value-weighted fashion.

A change of policy in this direction was solicited by a number of critics of the purchasing program, and on September 2016 the BoJ changed the guidelines for its asset purchases, reducing the share of capital flowing to ETFs tracking the Nikkei 225 Index and increasing its holdings of ETFs tracking the TOPIX. This brought the cross-sectional allocation of capital closer to what market capitalization would justify.

To date, the BoJ has not completely abandoned the price-weighted Nikkei Index, nor it is bringing its already accumulated holdings towards value-weighted proportions. According to our model, the BoJ should make sure to bring its holdings proportional to companies market capitalizations if it wants to amend the allocational side-effects of the policy.

[15] A purchase of $u' = aW_{\text{Topix}}$ in Yen corresponds to $u = aQ$ in shares, since the TOPIX is value-weighted.

9 Conclusion

In this paper we study asset pricing implications of the ETF purchase program undertaken by the BoJ since April 2013. The analysis is supported by a dynamic asset pricing model, featuring multiple assets with time-varying supply due to open market operations of the central bank.

To identify the net portfolio balance effect of the policy our empirical analysis exploits the exogeneity and the cross-sectional dimension of the BoJ's purchase schedule, which mitigates endogeneity problems characteristic of other studies.

We show that the intervention has a positive and persistent effect on domestic equity prices, thus reducing the cost of equity capital of domestic companies. We provide empirical evidence that the effect is consistent with a portfolio balance channel both in the cross-section and in the time-series.

This evidence suggests that demand curves for stocks are downward sloping in the long-run. We estimate an economically significant increase of 22 basis points in aggregate market valuation per trillion Yen invested into the program, which corresponds to a price elasticity of 1. The mechanism behind downward sloping demand curves in the model is through the change in the structure of systematic risk held by the private sector induced by the central bank's intervention. This change in the composition of risk leads to a new discount factor and consequently to price adjustments.

We also show that the outright purchases of the BoJ generate positive and persistent pressure on prices. Our estimates of the portfolio-balance channel remain significant after accounting for these flow effects of the policy.

Our results shed light on the side-effects of the LSAP, uncovering a highly heterogeneous impact in the cross-section of firms' cost of equity capital, both at the firm and at the industry level. Using our theoretical framework to evaluate the impact of arbitrary purchase schedules, we find that the observed heterogeneity in the price effects mainly arises from the weight given to the Nikkei 225 price-weighted index. Capital injections shaped according to market weights would instead induce a cross-sectionally homogeneous change in the cost of capital.

References

Andrade, P. and Ferroni, F. (2018). Delphic and odyssean monetary policy shocks: Evidence from the euro area.

Bagwell, L. S. (1992). Dutch auction repurchases: An analysis of shareholder heterogeneity. *The Journal of Finance*, 47(1):71–105.

Barberis, N. and Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, 1:1053–1128.

Ben-David, I., Franzoni, F., and Moussawi, R. (2018). Do ETFs increase volatility? *The Journal of Finance*, 73(6):2471–2535.

Bernanke, B. S. (2009). *Federal Reserve policies to ease credit and their implications for the Fed's balance sheet: a speech at the National Press Club Luncheon, National Press Club, Washington, DC, February 18, 2009*. Number 443.

BoJ (2013). Introduction of the Quantitative and Qualitative Monetary Easing, April 4th.

Brunner, K. and Meltzer, A. H. (1973). Mr. Hicks and the "Monetarists". *Economica*, 40(157):44–59.

Buraschi, A. and Whelan, P. (2015). Bond markets and unconventional monetary policy. *Handbook of Fixed-Income Securities*, First Edition.

Cochrane, J. H. and Piazzesi, M. (2002). The Fed and interest rates - A high-frequency identification. *American Economic Review*, 92(2):90–95.

Coval, J. and Stafford, E. (2007). Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics*, 86(2):479–512.

Culbertson, J. M. (1957). The term structure of interest rates. *The Quarterly Journal of Economics*, 71(4):485–517.

Curdia, V. and Woodford, M. (2011). The central-bank balance sheet as an instrument of monetary-policy. *Journal of Monetary Economics*, 58(1):54–79.

D'Amico, S. and King, T. B. (2013). Flow and stock effects of large-scale treasury purchases: Evidence on the importance of local supply. *Journal of Financial Economics*, 108(2):425–448.

Eggertsson, G. B. and Woodford, M. (2004). *Optimal monetary policy in a liquidity trap*. National Bureau of Economic Research (No. w10840).

Eser, F. and Schwaab, B. (2016). Evaluating the impact of unconventional monetary policy measures: Empirical evidence from the ECBs securities markets programme. *Journal of Financial Economics*, 119(1):147–167.

Frankel, J. A. (1985). Portfolio crowding-out, empirically estimated. *The Quarterly Journal of Economics*, 100(Supplement):1041–1065.

Gagnon, J., Raskin, M., Remache, J., and Sack, B. P. (2010). *Large-scale asset purchases by the Federal Reserve: did they work?* FRB of New York Staff Report 441.

Greenwood, R. (2005). Short-and long-term demand curves for stocks: theory and evidence on the dynamics of arbitrage. *Journal of Financial Economics*, 75(3):607–649.

Greenwood, R. (2007). Excess comovement of stock returns: Evidence from cross-sectional variation in nikkei 225 weights. *The Review of Financial Studies*, 21(3):1153–1186.

Hamilton, J. D. and Wu, J. C. (2012). The effectiveness of alternative monetary policy tools in a zero lower bound environment. *Journal of Money, Credit and Banking*, 44(s1):3–46.

Hanson, S. G. and Stein, J. C. (2015). Monetary policy and long-term real rates. *Journal of Financial Economics*, 115(3):429–448.

Harris, L. and Gurel, E. (1986). Price and volume effects associated with changes in the S&P 500 list: New evidence for the existence of price pressures. *the Journal of Finance*, 41(4):815–829.

Hau, H., Massa, M., and Peress, J. (2009). Do demand curves for currencies slope down? Evidence from the MSCI global index change. *The Review of Financial Studies*, 23(4):1681–1717.

He, Z. and Krishnamurthy, A. (2013). Intermediary asset pricing. *The American Economic Review*, 103(2):732–770.

Joyce, M., Lasaosa, A., Stevens, I., Tong, M., et al. (2011). The financial market impact of quantitative easing in the united kingdom. *International Journal of Central Banking*, 7(3):113–161.

Joyce, M., Miles, D., Scott, A., and Vayanos, D. (2012). Quantitative easing and unconventional monetary policy – an introduction. *The Economic Journal*, 122(564).

Kaul, A., Mehrotra, V., and Morck, R. (2000). Demand curves for stocks do slope down: New evidence from an index weights adjustment. *The Journal of Finance*, 55(2):893–912.

Krishnamurthy, A. and Vissing-Jorgensen, A. (2011). *The effects of quantitative easing on interest rates: channels and implications for policy.* National Bureau of Economic Research.

Krishnamurthy, A. and Vissing-Jorgensen, A. (2013). The ins and outs of LSAPs. In *Federal Reserve Bank of Kansas City's Jackson Hole Symposium on the Global Dimensions of Unconventional Monetary Policy, Jackson Hole, Wyoming, August.*

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Miles, D. and Schanz, J. (2014). The relevance or otherwise of the central bank's balance sheet. *Journal of International Economics*, 92:S103–S116.

Mitchell, M., Pulvino, T., and Stafford, E. (2004). Price pressure around mergers. *The Journal of Finance*, 59(1):31–63.

Modigliani, F. and Sutch, R. (1966). Innovations in interest rate policy. *The American Economic Review*, 56(1/2):178–197.

Nakamura, E. and Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: The information effect. *The Quarterly Journal of Economics*, 133(3):1283–1330.

Neely, C. J. et al. (2010). *The large scale asset purchases had large international effects.* Federal Reserve Bank of St. Louis, Research Division.

Petajisto, A. (2009). Why do demand curves for stocks slope down? *Journal of Financial and Quantitative Analysis*, 44(5):1013–1044.

Petajisto, A. (2011). The index premium and its hidden cost for index funds. *Journal of Empirical Finance*, 18(2):271–288.

Schnitzler, J. (2016). S&P 500 inclusions and stock supply. *Working Paper.*

Scholes, M. S. (1972). The market for securities: Substitution versus price pressure and the effects of information on share prices. *The Journal of Business*, 45(2):179–211.

Shleifer, A. (1986a). Do demand curves for stocks slope down? *The Journal of Finance*, 41(3):579–590.

Shleifer, A. (1986b). Do demand curves for stocks slope down? *The Journal of Finance*, 41(3):579–590.

Swanson, E. T. (2011). Let's twist again: a high-frequency event-study analysis of operation twist and its implications for QE2. *Brookings Papers on Economic Activity*, 2011(1):151–188.

Tobin, J. (1969). A general equilibrium approach to monetary theory. *Journal of money, credit and banking*, 1(1):15–29.

Vayanos, D. and Vila, J.-L. (2009). *A preferred-habitat model of the term structure of interest rates.* National Bureau of Economic Research (No. w15487).

Wurgler, J. and Zhuravskaya, E. (2002). Does arbitrage flatten demand curves for stocks? *The Journal of Business*, 75(4):583–608.

Appendix

A Additional Material: Figures and Tables

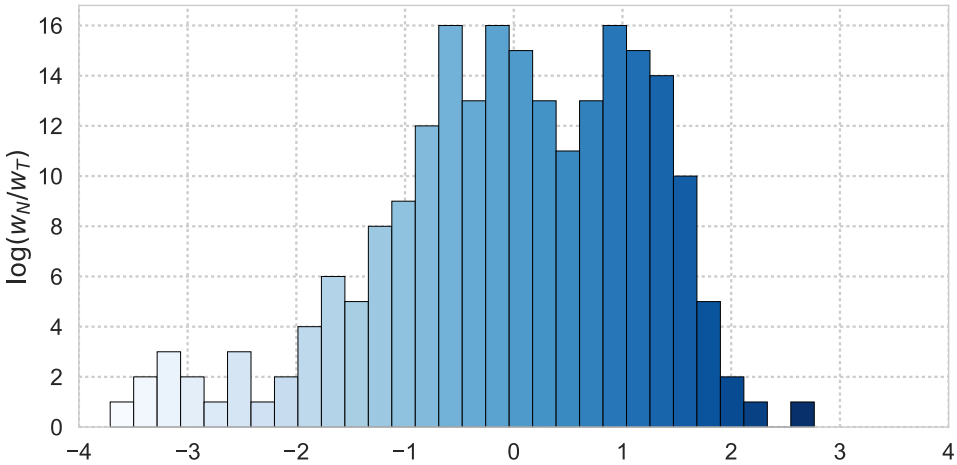


Figure A.1: Distortion. The figure plots the distribution of the log ratio between the Nikkei weight w_N and the TOPIX weight w_T for Nikkei firms only. The histogram shows a significant dispersion, confirming that Nikkei weights induce significant cross-sectional variation of purchased quantities relative to market capitalization.

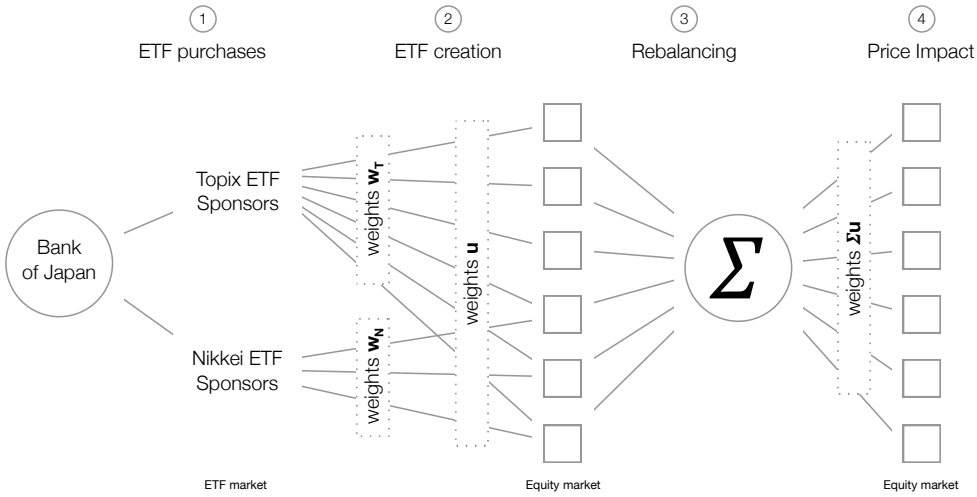


Figure A.2: From ETFs to equity This figure describes the channel though which ETF purchases of the central bank may have an impact on equity prices. As the BoJ buys TOPIX- and Nikkei-linked ETFs, these are created by ETF sponsors and/or authorized participants. The securities needed to form the ETF basket are collected by these intermediaries in the equity market, thus effectively reducing the supply of equity shares available to private investors.

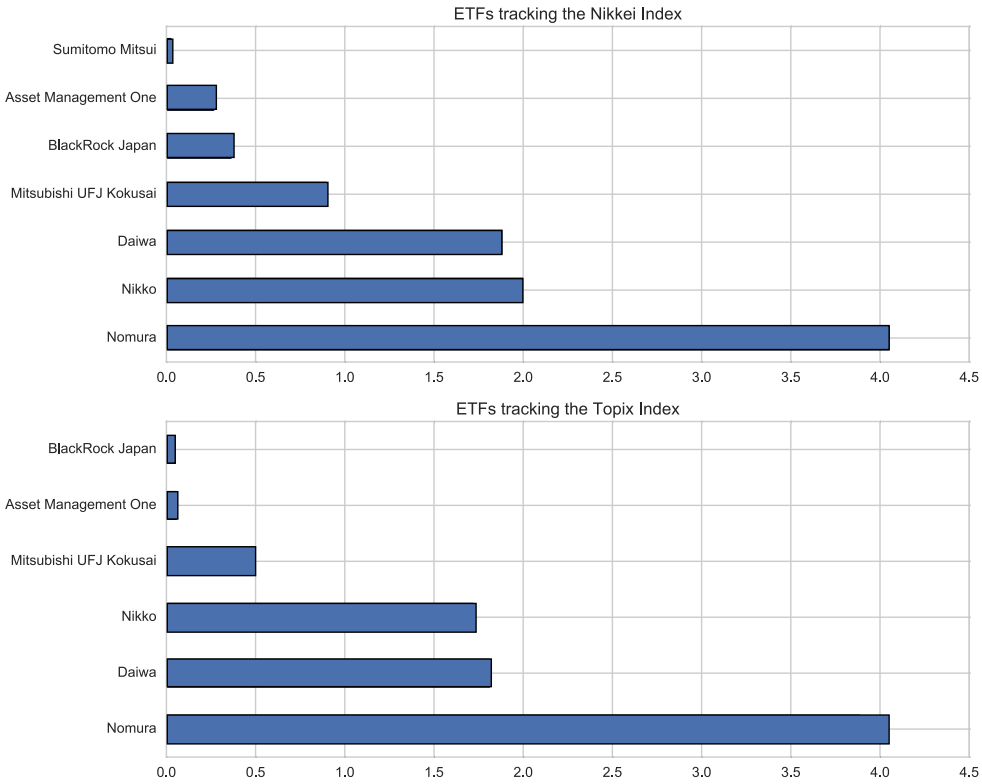
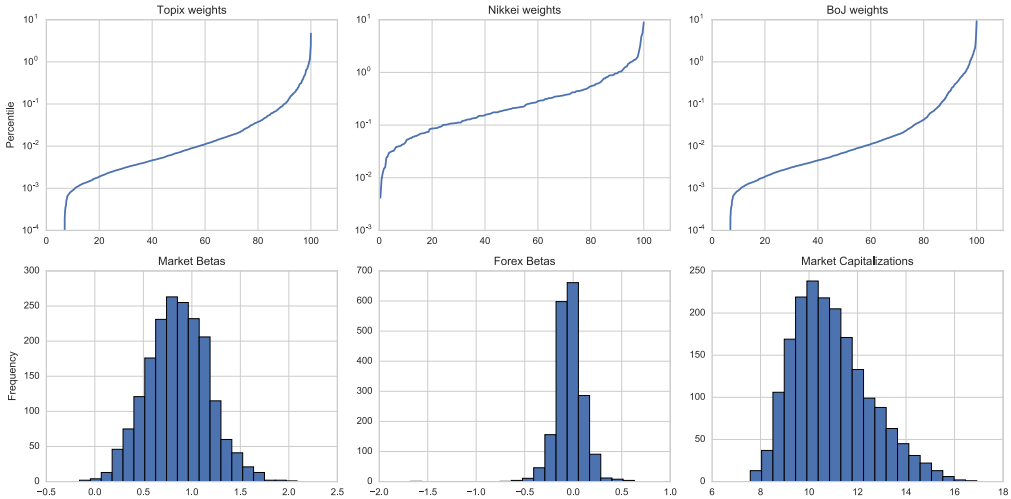


Figure A.3: Assets Under Management (AUM) by Provider (in trillion yen). This figure shows the Assets Under Management of ETFs aggregated at Provider level. The values are computed as of December 30th, 2016.

Panel A – 2014 Event



Panel B – 2016 Event

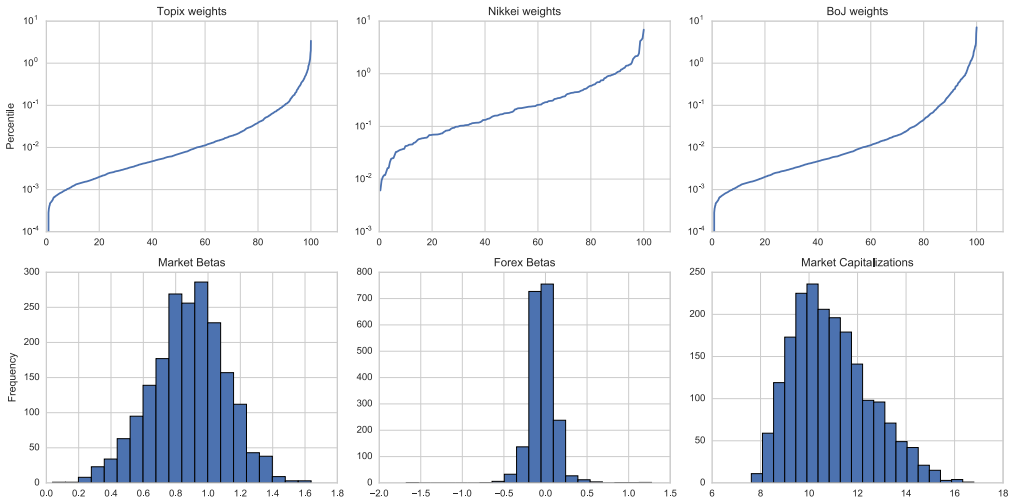


Figure A.4: Weights, betas and market capitalizations. The plots display cross-sectional heterogeneity of the variables of interest at the time of the BoJ announcements. Panel A refers to the announcement in 2014, Panel B to the announcement in 2016. The first row of each panel plots the percentile functions in logarithmic scale of the TOPIX weights (ω_T), the Nikkei weights (ω_N) and the BoJ weights(q). BoJ weights are computed as $\omega_T + \omega_N$ and correspond to the elements of the vector q in the model. The second row of each panel shows the distribution of stock-level market betas, Forex betas and market values. Market betas and Forex betas are estimated following the procedure explained in Section **??**. Companies market capitalizations are in logs.

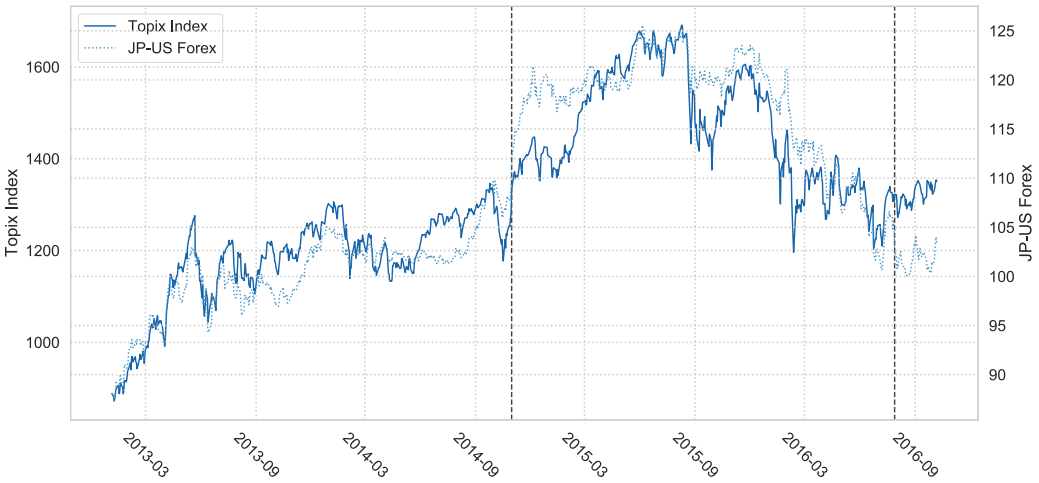


Figure A.5: TOPIX Index and JP-US Exchange Rate. This figure shows the time-series of the TOPIX Index over our sample period (green solid line, left axis) and of the exchange rate from US Dollar to Japanese Yen (purple dotted line, right axis).

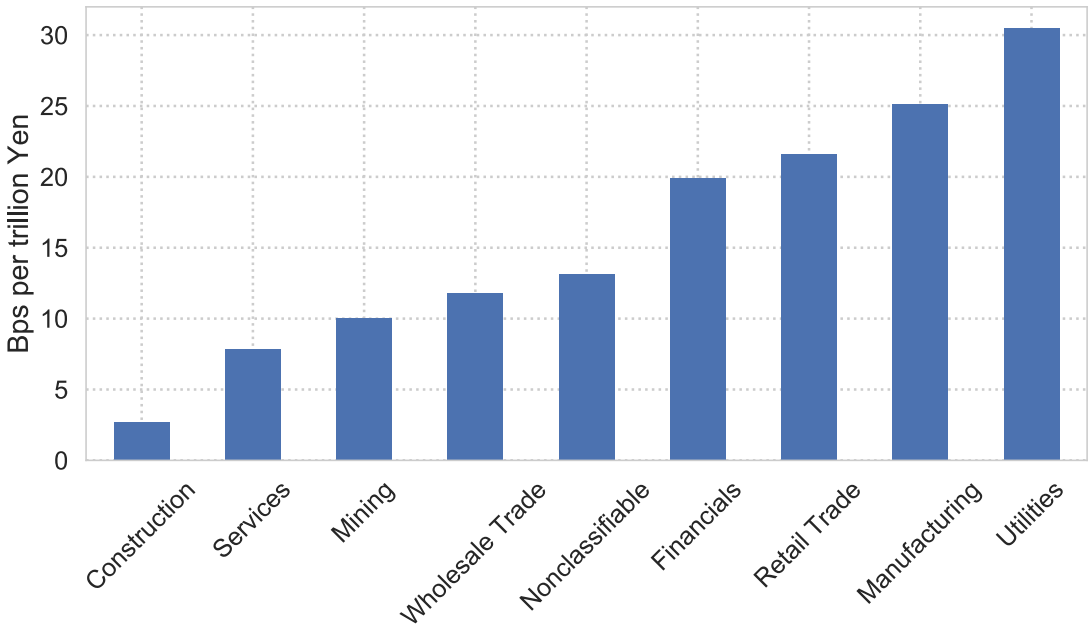


Figure A.6: Portfolio Balance Effect across Industries This figure shows the estimated portfolio balance impact of the policy, expressed in basis points per trillion Yen, computed separately for each sector.

	Mean	Std Deviation	Min	25%	50%	75%	Max	Obs
Market Cap (Billions Yen)								
TOPIX	252	853	2	17	45	148	22210	3824
Nikkei 225	1352	2110	28	294	683	1522	22210	442
Not Nikkei 225	108	251	2	15	35	94	4434	3382
Forex Beta								
TOPIX	-0.04	0.15	-1.68	-0.12	-0.04	0.04	1.27	3824
Nikkei 225	0.02	0.13	-0.39	-0.07	0.02	0.10	0.46	442
Not Nikkei 225	-0.05	0.15	-1.68	-0.12	-0.05	0.03	1.27	3382
Market Beta								
TOPIX	0.87	0.27	-0.16	0.69	0.88	1.05	2.08	3824
Nikkei 225	1.05	0.19	0.46	0.90	1.04	1.18	1.71	442
Not Nikkei 225	0.85	0.27	-0.16	0.67	0.85	1.02	2.08	3382
BoJ Weight								
TOPIX	0.05	0.22	0.00	0.00	0.00	0.01	4.65	3824
Nikkei 225	0.37	0.53	0.01	0.10	0.20	0.43	4.65	442
Not Nikkei 225	0.01	0.03	0.00	0.00	0.00	0.01	0.42	3382
Nikkei 225 Weight								
TOPIX	0.05	0.31	0.00	0.00	0.00	0.00	8.91	3824
Nikkei 225	0.45	0.82	0.00	0.09	0.20	0.45	8.91	442
Not Nikkei 225	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3382
TOPIX Weight								
TOPIX	0.05	0.18	0.00	0.00	0.01	0.03	4.70	3824
Nikkei 225	0.30	0.44	0.01	0.06	0.15	0.36	4.70	442
Not Nikkei 225	0.02	0.05	0.00	0.00	0.01	0.02	0.85	3382

Table A.1: Summary Statistics. This table provides summary statistics for various stock characteristics by index membership. TOPIX stocks represent our entire sample of stocks. Nikkei stocks are those included in the Nikkei 225 index, while Not Nikkei stocks are those that only appear in the TOPIX index. All Nikkei companies also belong to the TOPIX index. All statistics are computed using pre-event information and pooling both events together.

Panel A: October 31st, 2014

	Raw Returns				Abnormal Returns			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
π	95.77***	91.16***	39.82***	38.19***	18.13***	13.14**	47.69***	43.85***
	(16.84)	(15.62)	(4.49)	(4.32)	(3.17)	(2.24)	(5.50)	(5.09)
u		0.01***	-0.00	-0.00		0.01***	-0.00	-0.00
		(3.32)	(-0.86)	(-0.73)		(3.58)	(-0.99)	(-0.77)
Market Beta			0.02***	0.01			-0.08***	-0.09***
			(2.70)	(1.38)			(-11.04)	(-12.29)
Forex Beta			0.07***	0.07***			0.07***	0.07***
			(7.58)	(7.45)			(8.30)	(7.88)
log(Market Cap)			0.01***	0.01***			0.01***	0.01***
			(6.87)	(6.33)			(7.13)	(6.60)
Amihud			0.00	0.00			0.00	0.00
			(0.96)	(0.51)			(0.94)	(0.50)
Observations	1,851	1,851	1,807	1,701	1,851	1,851	1,807	1,701
R-squared	0.13	0.14	0.19	0.24	0.01	0.01	0.11	0.17
Industry FE	NO	NO	NO	YES	NO	NO	NO	YES

Panel B: July 29th, 2016

	Raw Returns				Abnormal Returns			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
π	18.00***	16.27***	15.80***	18.59***	18.00***	16.64***	18.59***	20.24***
	(6.29)	(5.57)	(3.59)	(4.10)	(6.11)	(5.52)	(4.10)	(4.33)
u		0.01***	0.01*	0.01**		0.01**	0.01*	0.01**
		(2.75)	(1.82)	(2.00)		(2.10)	(1.80)	(1.97)
Market Beta			-0.00	-0.02*			-0.01	-0.02**
			(-0.57)	(-1.80)			(-0.99)	(-2.22)
Forex Beta			0.02**	0.01			0.02***	0.02*
			(2.06)	(1.42)			(2.63)	(1.94)
log(Market Cap)			0.00	0.00			-0.00	0.00
			(0.99)	(1.11)			(-0.02)	(0.17)
Amihud			0.00	0.00			0.00*	0.00*
			(1.48)	(1.53)			(1.75)	(1.78)
Observations	1,905	1,905	1,839	1,734	1,905	1,905	1,839	1,734
R-squared	0.02	0.02	0.03	0.05	0.02	0.02	0.03	0.05
Industry FE	NO	NO	NO	YES	NO	NO	NO	YES

Table A.2: Robustness: Alternative Covariance Matrix Estimation The tables report results for specifications similar to those in Table **??**, but where the main explanatory variable $\pi = \Sigma u$ is constructed using the covariance matrix Σ estimated on raw returns. Regressions of event returns on the predicted price impact π are run separately for the two events. The dependent variable in columns 1-3 is the cumulative raw return, while in columns 4-6 is the cumulative abnormal return with respect to the market model estimated in the pre-event window. Cumulative returns are computed over a 10 days horizon after the announcement date. t-statistics are in parenthesis; asterisks denote conventional significance levels (***=1%, **=5%, *=10%).

	Abnormal Returns 2014						Abnormal Returns 2016					
Horizon (days)	5	10	21	63	126	252	5	10	21	63	126	252
π	11.06**	30.60***	22.51**	63.85***	152.1***	275.0***	15.52**	16.69**	30.48***	34.18**	82.20***	110.4***
	(2.11)	(4.62)	(2.52)	(4.34)	(8.50)	(10.32)	(2.96)	(2.52)	(3.42)	(2.32)	(4.60)	(4.14)
u	-0.00	-0.02***	-0.01	-0.02***	-0.02	-0.04*	0.010*	0.004	0.024**	0.014	0.045**	0.102***
	(-0.39)	(-3.20)	(-1.30)	(-2.52)	(-1.65)	(-1.99)	(2.06)	(0.67)	(2.55)	(1.35)	(2.91)	(4.70)
Market Beta	-0.04*	-0.07*	-0.08*	-0.17**	-0.32**	-0.45***	0.016	-0.01	0.009	0.012	0.007	0.041
	(-1.47)	(-1.97)	(-1.70)	(-2.32)	(-2.22)	(-2.68)	(0.59)	(-0.26)	(0.19)	(0.16)	(0.05)	(0.24)
Forex Beta	0.039**	0.041**	0.095***	0.087**	0.019	-0.12	-0.00	0.014	0.053**	0.049	0.190***	0.166
	(2.81)	(2.32)	(4.11)	(2.26)	(0.30)	(-0.03)	(-0.48)	(0.80)	(2.32)	(1.28)	(2.92)	(0.04)
log(Market Cap)	0.002	0.006	0.001	-0.00	0.005	-0.00	-0.00	-0.00	-0.01	-0.02	-0.06***	-0.10***
	(0.63)	(1.16)	(0.22)	(-0.07)	(0.26)	(-0.72)	(-1.40)	(-0.49)	(-1.60)	(-1.70)	(-3.43)	(-8.54)
Amihud	0.000	4.618	0.001	0.000	0.019***	0.005	0.000	0.000	-0.00	-0.00	-0.00**	-0.01***
	(0.66)	(0.03)	(0.75)	(0.36)	(7.19)	(1.67)	(0.18)	(0.11)	(-0.59)	(-1.31)	(-3.23)	(-3.11)
Observations	1,701	1,701	1,701	1,701	1,701	1,701	1,734	1,734	1,734	1,734	1,734	1,734
R-squared	0.114	0.160	0.102	0.120	0.180	0.141	0.071	0.050	0.101	0.111	0.203	0.191
Industry FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Table A.3: Cross-sectional regressions with industry fixed effects. The table report the coefficients of cross-sectional regressions of cumulative returns (in percentage points) computed at different horizons on the predicted price impact π and a set of control variables (standardized). In this specification we include industry fixed effects, based on the first 3 digits of the Standard Industry Classification Code (SIC-3). Regressions are run separately for the two events. The dependent variable is the cumulative abnormal return with respect to the market model estimated in the pre-event window. t-statistics from placebo regressions are in parenthesis; asterisks denote conventional significance levels (***=1%, **=5%, *=10%) based on empirical p-values.

	Abnormal Returns 2014						Abnormal Returns 2016					
	5	10	21	63	126	252	5	10	21	63	126	252
π	19.04***	42.17***	28.77***	72.38***	172.5***	300.7***	17.61**	17.34*	31.88***	38.51**	87.69***	112.9***
	(3.22)	(5.69)	(2.86)	(4.57)	(9.25)	(11.60)	(2.98)	(2.34)	(3.16)	(2.43)	(4.70)	(4.36)
u	0.008*	-0.00	-0.00	-0.03***	-0.03**	-0.11***	0.014***	0.001	-0.00	-0.02*	-0.04**	0.001
	(1.90)	(-1.32)	(-0.97)	(-3.44)	(-2.06)	(-4.64)	(3.28)	(0.27)	(-0.24)	(-2.15)	(-2.47)	(0.07)
Market Beta	-0.02	-0.05	-0.07	-0.16*	-0.29**	-0.40***	0.026	0.002	0.027	0.028	0.048	0.076
	(-0.84)	(-1.48)	(-1.37)	(-2.06)	(-2.05)	(-2.14)	(0.86)	(0.07)	(0.51)	(0.36)	(0.33)	(0.41)
Forex Beta	0.037**	0.042*	0.102***	0.097**	0.043	-0.13	-0.00	0.018	0.053*	0.060	0.196***	0.186***
	(2.20)	(2.02)	(3.81)	(2.28)	(0.61)	(-1.72)	(-0.18)	(0.90)	(2.00)	(1.43)	(2.73)	(2.33)
log(Market Cap)	0.001	0.006	0.001	0.001	0.005	-0.00	-0.00	-0.00	-0.01*	-0.03	-0.07***	-0.11***
	(0.33)	(0.94)	(0.16)	(0.06)	(0.24)	(-0.33)	(-1.48)	(-0.55)	(-1.80)	(-1.79)	(-3.58)	(-7.59)
Amihud	0.001	0.000	4.552	0.001	0.016***	0.018***	0.000	0.000	-0.00	-0.00	-0.00**	-0.00
	(0.79)	(0.25)	(0.00)	(0.67)	(6.13)	(7.14)	(0.48)	(0.39)	(-0.53)	(-0.96)	(-3.27)	(-2.45)
Nikkei	-0.01	-0.02*	-0.00	-0.00	-0.00	0.048*	-0.00	0.003	0.038*	0.049*	0.122***	0.153***
	(-1.26)	(-1.72)	(-0.27)	(-0.08)	(-0.18)	(1.54)	(-0.38)	(0.24)	(2.23)	(2.32)	(4.03)	(4.88)
Observations	1,807	1,807	1,807	1,807	1,807	1,807	1,839	1,839	1,839	1,839	1,839	1,839
R-squared	0.06	0.11	0.07	0.10	0.15	0.12	0.05	0.03	0.08	0.08	0.19	0.15
Industry FE	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO

Table A.4: Cross-sectional regressions controlling for Nikkei. The table report the coefficients of cross-sectional regressions of cumulative returns (in percentage points) computed at different horizons on the predicted price impact π and a set of control variables (standardized). In this specification we add a dummy variable *Nikkei* that indicates stocks belonging to the Nikkei 225 Index. Regressions are run separately for the two events. The dependent variable is the cumulative abnormal return with respect to the market model estimated in the pre-event window. t-statistics from placebo regressions are in parenthesis; asterisks denote conventional significance levels (***=1%, **=5%, *=10%) based on empirical p-values.

	Abnormal Returns 2014						Abnormal Returns 2016					
Horizon (days)	5	10	21	63	126	252	5	10	21	63	126	252
π	12.21**	32.64***	22.95**	63.97***	150.7***	266.7***	15.62**	16.37**	28.39**	31.48**	75.75***	102.7***
	(2.28)	(4.80)	(2.51)	(4.38)	(8.72)	(10.01)	(2.91)	(2.41)	(3.11)	(2.16)	(4.38)	(3.86)
u	0.006	-0.00	-0.00	-0.02***	-0.03**	-0.10***	0.012**	0.000	-0.00	-0.02*	-0.04**	-0.00
	(1.35)	(-1.41)	(-1.16)	(-2.89)	(-2.32)	(-4.28)	(2.72)	(0.05)	(-0.56)	(-2.42)	(-2.95)	(-0.21)
Market Beta	-0.04*	-0.07*	-0.08*	-0.17**	-0.33**	-0.46***	0.016	-0.00	0.011	0.015	0.014	0.050
	(-1.48)	(-1.98)	(-1.71)	(-2.38)	(-2.23)	(-2.28)	(0.59)	(-0.25)	(0.24)	(0.21)	(0.10)	(0.25)
Forex Beta	0.040**	0.043**	0.095***	0.087**	0.018	-0.12***	-0.00	0.012	0.044*	0.038	0.163**	0.134***
	(2.83)	(2.38)	(4.11)	(2.25)	(0.28)	(-1.93)	(-0.45)	(0.71)	(1.93)	(0.98)	(2.49)	(2.01)
log(Market Cap)	0.002	0.006	0.001	-0.00	0.004	-0.00	-0.00	-0.00	-0.01	-0.02	-0.06***	-0.10***
	(0.61)	(1.09)	(0.21)	(-0.07)	(0.24)	(-0.81)	(-1.29)	(-0.48)	(-1.61)	(-1.69)	(-3.43)	(-9.10)
Amihud	0.001	0.000	0.001	0.000	0.019***	0.005	0.000	0.000	-0.00	-0.00	-0.00**	-0.01***
	(0.67)	(0.06)	(0.70)	(0.34)	(6.97)	(1.52)	(0.18)	(0.09)	(-0.64)	(-1.32)	(-3.33)	(-3.15)
Nikkei	-0.01	-0.01	-0.00	-0.00	0.011	0.074***	-0.00	0.006	0.039*	0.050*	0.120***	0.143***
	(-1.04)	(-1.39)	(-0.22)	(-0.05)	(0.36)	(2.15)	(-0.17)	(0.46)	(2.24)	(2.29)	(3.62)	(4.17)
Observations	1,701	1,701	1,701	1,701	1,701	1,701	1,734	1,734	1,734	1,734	1,734	1,734
R-squared	0.12	0.16	0.10	0.12	0.18	0.14	0.07	0.05	0.11	0.12	0.21	0.20
Industry FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Table A.5: Cross-sectional regressions controlling for Nikkei and industry. The table report the coefficients of cross-sectional regressions of cumulative returns (in percentage points) computed at different horizons on the predicted price impact π and a set of control variables (standardized). In this specification we add a dummy variable *Nikkei* that indicates stocks belonging to the Nikkei 225 Index and industry fixed effects based on the first 3 digits of the Standard Industry Classification Code (SIC-3). Regressions are run separately for the two events. The dependent variable is the cumulative abnormal return with respect to the market model estimated in the pre-event window. t-statistics from placebo regressions are in parenthesis; asterisks denote conventional significance levels (***=1%, **=5%, *=10%) based on empirical p-values.

B Model Derivation

The model features a representative investor who chooses time-t demand N_t of shares to maximize its next period exponential utility subject to a standard budget constraint

$$\max_{N} \; \mathrm{E}_t\left(-\exp(-\gamma W_{t+1})\right) \tag{1}$$

$$\text{s.t.} \qquad W_{t+1} = W_t(1+r) + N_t'(p_{t+1} + D_{t+1} - p_t(1+r)) \tag{2}$$

From the first order condition it follows that

$$N_t = \frac{1}{\gamma}[\mathrm{Var}_t(p_{t+1} + D_{t+1})]^{-1}(\mathrm{E}_t[p_{t+1} + D_{t+1} - p_t(1+r)]) \tag{3}$$

We restrict our attention to the covariance stationary equilibrium. Imposing market clearing and substituting $V = \mathrm{Var}_t(p_{t+1} + D_{t+1})$ yields

$$(1+r)p_t = \mathrm{E}_t[p_{t+1} + D_{t+1}] - \gamma V Q_t \tag{4}$$

Iterating forward up to time T and applying the law of iterated expectations we get

$$(1+r)p_t = \mathrm{E}_t\left[\frac{p_T}{(1+r)^{T-t-1}}\right] + \sum_{i=0}^{T-t-1} \frac{D_t}{(1+r)^i} - \gamma V \sum_{i=0}^{T-t-1} \frac{\mathrm{E}_t[Q_{t+i}]}{(1+r)^i} \tag{5}$$

Taking the limit $T \to \infty$ and imposing the no-bubble condition yields

$$p_t = \frac{D_t}{r} - \frac{\gamma V}{(1+r)}\left(\sum_{i=0}^{\infty} \frac{\mathrm{E}_t[Q_{t+i}]}{(1+r)^i}\right) = \frac{1}{r}\left(D_t - \gamma V \Omega_t\right) \tag{6}$$

where we introduced the notation

$$\Omega_t = \frac{r}{1+r} \sum_{i=0}^{\infty} \frac{\mathrm{E}_t[Q_{t+i}]}{(1+r)^i} \tag{7}$$

which can be interpreted as the discounted time-t expected future supply of the assets. This term is crucial for our analysis, representing the channel through which the central bank is able to affect risk premia. Under no expectation of monetary policy intervention we have $\mathrm{E}_t(Q_{t+i}) = Q$, so that the resulting pricing equation collapses to

$$p_t = \frac{1}{r}\left(D_t - \gamma V Q\right) \tag{8}$$

where the vector $\gamma V Q$ can be interpreted as the cross-sectional vector of risk premia required by investors in equilibrium. In our context, this is the pricing equation that applies before the policy announcement at $t = 1$.

In the following sections we look at what happens to prices if the central bank unexpectedly commits itself to a large-scale purchase of assets over a defined period, thus affecting the expected path of future supply Ω_t.

We now solve the model in its most general form, allowing for the possibility that agents expectations on future supply change over time. We assume that for each $t \geq 1$ there exist a scalar $\lambda_t \geq 0$ such that the time-t expectation is

$$\begin{cases} \mathrm{E}_t(Q_{t+i}) = Q & \text{for } i \geq 0 \text{ and } t < 1 \\ \mathrm{E}_t(Q_{t+i}) = Q - \lambda_t(t+i)q & \text{for } i \geq 0 \text{ and } t = 1, \ldots, M \\ \mathrm{E}_t(Q_{t+i}) = Q - \lambda_t M q & \text{for } i \geq M - t \text{ and } t \geq 1 \end{cases} \tag{9}$$

The parameter λ_t can be interpreted as the degree of confidence of investors in the BoJ commitment or, in other words, as the conditional probability they attach to the continuation of the program.

Assuming that investors increase their confidence as time passes – and they observe more actual purchases by the BoJ – amounts to assume that λ_t is increasing in time.

After the BoJ announcement, for $t \geq 1$, the expected supply can be written as

$$\begin{aligned} \Omega_t &= \frac{r}{1+r} \sum_{i=0}^{\infty} \frac{\mathrm{E}_t[Q_{t+i}]}{(1+r)^i} && (10) \\ &= \frac{r}{1+r} \left(\sum_{i=0}^{M-t-1} \frac{Q - \lambda_t(t+i)q}{(1+r)^i} + \sum_{i=M-t}^{\infty} \frac{Q - \lambda_t M q}{(1+r)^i} \right) && (11) \\ &= Q - \frac{\lambda_t r}{1+r} \left(\sum_{i=0}^{M-t-1} \frac{(t+i)q}{(1+r)^i} + \sum_{i=M-t}^{\infty} \frac{Mq}{(1+r)^i} \right) && (12) \\ &= Q - \frac{\lambda_t r}{1+r} \left(\sum_{i=0}^{M-t-1} \frac{(t+i-M)q}{(1+r)^i} + \sum_{i=0}^{\infty} \frac{Mq}{(1+r)^i} \right) && (13) \\ &= Q - \lambda_t M q + \frac{\lambda_t r}{1+r} \sum_{i=0}^{M-t-1} \frac{(M-t-i)}{(1+r)^i} q && (14) \\ &= Q - \lambda_t M q + \lambda_t \varphi(t) q && (15) \end{aligned}$$

where we introduced the real-valued function

$$\varphi(t) = \frac{r}{1+r} \sum_{i=0}^{M-t-1} \frac{(M-t-i)}{(1+r)^i}, \qquad t \geq 1 \tag{16}$$

This quantity represents the residual duration of the program at time t. In the Internet Appendix we show that, for realistic values of the risk-free rate r, the function $\varphi(t)$ enjoys the following properties:

(i) $\varphi(t+1) - \varphi(t) < 0$

(ii) $\varphi(t) < M$ for $t \geq 1$

(iii) $\varphi(t) = 0$ for $t \geq M$

Given the assumption that by the end of the policy horizon $t = M$ the central bank will have purchased exactly Mq as announced and afterwards it will not engage in further market operations, it follows that the pricing equation (8) takes the form

$$\begin{cases} p_t = \frac{1}{r}\left(D_t - \gamma V Q\right) & \text{for } t < 1 \\ p_t = \frac{1}{r}\left(D_t - \gamma V(Q - \lambda_t M q + \lambda_t \varphi(t) q)\right) & \text{for } t = 1, \dots M \\ p_t = \frac{1}{r}\left(D_t - \gamma V(Q - Mq)\right) & \text{for } t \geq M \end{cases} \tag{17}$$

and the price change at the announcement day $t = 1$ can be written as

$$p_1 - p_0 = \frac{1}{r}\left(\varepsilon_1 + \lambda_1 \gamma V(Mq - \varphi(1)q)\right) \tag{18}$$

Dividing by p_0 coordinate-wise proves Proposition **??**. The equation also shows that the size of the price jump is increasing in the initial belief parameter λ_1.

On the days following the announcement, price changes depend on the time-series evolution of λ_t. Denoting the updates in beliefs by $\Delta\lambda_{t+1} = \lambda_{t+1} - \lambda_t$ we have

$$\begin{aligned} p_{t+1} - p_t &= \frac{1}{r}\left(\varepsilon_{t+1} - \gamma V((\lambda_{t+1}\varphi(t+1) - \lambda_t \varphi(t)) - \Delta\lambda_{t+1} M)q\right) && (19) \\ &= \frac{1}{r}\left(\varepsilon_{t+1} + \gamma \xi(t+1) V q\right), \qquad t = 1, \dots M && (20) \end{aligned}$$

Given $\Delta\lambda_{t+1} > 0$, the following inequalities show that $\xi(t) > 0$

$$\lambda_{t+1}\varphi(t+1) - \lambda_t\varphi(t) < \lambda_{t+1}\varphi(t) - \lambda_t\varphi(t)) = \Delta\lambda_{t+1}\varphi(t) < \Delta\lambda_{t+1} M \tag{21}$$

Therefore we conclude that if $\Delta\lambda_{t+1} > 0$ for every $t = 1, \dots, M$, then we should observe a positive relationship between Vq and the cross-section of price changes. To complete the proof of the first part of Proposition **??** we need to show that this conclusion also applies to the relationship between returns $R_{i,t+1} = (p_{i,t+1} - p_{i,t})/p_{i,t}$ and $\pi = \Sigma u$. This follows from the definitions of $\Sigma_{i,j}$, u_i and π_i

$$R_{i,t+1} = \frac{1}{r}\left(\varepsilon_{i,t+1}/p_{i,t} + \gamma\xi(t+1)(Vq)_i/p_{i,t}\right) = \frac{1}{r}\left(\varepsilon_{i,t+1}/p_{i,t} + \gamma\xi(t+1)\pi_i\right) \tag{22}$$

Finally taking the expectation of the cumulative returns we get

$$\sum_{s=1}^{t} \mathrm{E}[R_s] = \sum_{s=1}^{t} \frac{1}{r}\left(\gamma\xi(s)\pi\right) = \theta_t \pi \tag{23}$$

where $\theta_t = \sum_{s=1}^{t} \frac{\gamma}{r}\xi(s)$ is a positive and increasing function of t, which follows from $\xi(s) > 0$ for $s = 1, \dots, M$ as shown above. This concludes the proof of Proposition **??**.

C Systematic Risk in the Model

In our model the systematic risk of security i is measured as $(VQ)_i$, where V is the covariance matrix of price innovations and Q is the vector of shares outstanding. This quantity represents the covariance

of the security's price changes ε with the change in the wealth of the representative agent (i.e. the value of the market portfolio) and it therefore admits an interpretation similar to the market beta. Denoting the value of the market portfolio by MP and the covariance of the price of stock i with MP by $\beta(P)_i$ we have

$$\beta(P)_i = \mathrm{Cov}(\Delta MP, \Delta p_i) \propto \mathrm{Cov}(\varepsilon_t' Q, \varepsilon_{i,t}) = (VQ)_i \tag{24}$$

Market betas are usually defined in terms of returns, not of price changes. Thus an empirically more relevant definition of the systematic risk of security i is given by $(\Sigma W)_i$, where Σ is the covariance matrix of returns and W is the vector of percentage weights of the market portfolio. This quantity is proportional to the market beta of stock i, denoted by $\beta(R)_i$

$$\beta(R)_i = \frac{\mathrm{Cov}(R^{mkt}, R_i)}{\mathrm{Var}(R^{mkt})} \propto \mathrm{Cov}(R'W, R_i) = \sum_{i,j} \mathrm{Cov}(R_j, R_i) W_j = (\Sigma W)_i \tag{25}$$

Let $Mq = Q^{\mathrm{post}} - Q$ denote the announced change in the supply of assets and β^{post} the implied vector of market beta after the announcement. It follows immediately from the above definitions that the change in (price-level) systematic risk is proportional to the product between V and q:

$$\beta(P)_i^{\mathrm{post}} - \beta(P)_i \propto (VQ^{\mathrm{post}} - VQ)_i \propto -(Vq)_i \tag{26}$$

Similarly, from the definition of $\pi = \Sigma u$, where $u_i = p_i q_i$ is the announced change in the supply of stock i expressed in yen, it follows that

$$\beta(R)_i^{\mathrm{post}} - \beta(R)_i \propto (\Sigma W^{\mathrm{post}} - \Sigma W)_i \propto -(\Sigma u)_i = -\pi_i \tag{27}$$

For each stock i, π_i can thus be interpreted as the change in the stock beta (i.e. the systematic risk) induced by the supply shock. Notice that this change is induced by the policy through a modification of the portfolio held by the representative agent, while the fundamental covariance structure of returns is unchanged.

Focusing at High Frequency: An Attention-based Neural Network for Limit Order Books

Andrea Barbon*

Abstract

Machine learning methods deliver superior forecasting accuracy but can hardly be used to make inference. To overcome this limitation, I propose an encoder-decoder neural network augmented with an attention-based mechanism that can autonomously learn to identify the most critical regions of the input data. I first train the model using high-frequency message data from the NASDAQ and show that it outperforms other state-of-the-art models in forecasting future transaction prices. Then, I develop a methodology that uses the attention mechanism to make inference on the relative share of information content of market orders versus limit orders, concluding that the most informative events are executions of market orders while submission and cancellations of limit orders are less relevant. Finally, I test the model's behavior during the execution of real block orders from institutional investors, showing that it favors liquidity provision rather than front-running strategies.

Keywords: Limit Order Book, Machine Learning, Attention, Inference, Liquidity Provision
JEL classification: C45, C58, G13, G17

*Andrea Barbon (www.abarbon.com, andrea.barbon@usi.ch) is affiliated with USI Lugano and the Swiss Finance Institute. The author is thankful to Joshua D. Coval, Francesco Franzoni, Shimon Kogan, Fabio Pruneri, Jeremy Stein, and Adi Sunderam for fruitful comments and insightful discussions. All mistakes are mine.

1 Introduction

Artificial intelligence and machine learning methods are attracting widespread interest among the scientific community and the industry alike, due to the uncountable successful applications across a variety of different fields. In particular, deep learning methods have proven effective in the context of financial time-series forecasting. On the one hand, these methodologies are of interest for practitioners, since they deliver higher prediction accuracy with respect to traditional techniques. On the other hand, one of the major drawbacks to adopting machine learning models is that they are difficult to interpret and do not allow for inference. The black-boxiness nature of these models seems to preclude the possibility of adding them to the econometric toolkit of empirical researchers in financial economics.

However, the fact the deep neural networks are able to produce accurate out-of-sample forecasts suggests that these models can identify meaningful patterns in the data. This, in turn, opens the possibility that a fitted neural network could be used to make inference on the data on which it was trained. However, despite the intriguing potentials of these data-mining capabilities, to the best of my knowledge the literature has not yet developed any methodology to use machine learning models as tools to answer economic questions in a data-driven fashion.

This paper proposes a novel approach to increase the intelligibility of machine learning models, enabling researches to use them as inferential devices. The idea is based on a recently developed class of neural networks, featuring an attention-based mechanism that can autonomously learn to identify the most informative parts of the input data. In the context of time-series forecasting, in particular, these models learn to allocate a heterogeneous degree of attention to different time steps of the input sequence. This unique feature allows to perform inference from the trained model and delivers a higher level of interpretability relative to previously adopted architectures. Intuitively, one can use the trained attention-layer to identify the temporal regions which are most important for the model to produce its forecasts. The resulting time-series of attention levels can be interpreted as a measure of the informativeness of different events with respect to the assigned prediction task.

A natural way to showcase this idea is to apply attention-based neural networks to a realistic and well-studied forecasting problem in the context of financial markets. With such a motivation in mind, in this paper I focus on the adverse selection problem faced by market makers submitting quotes to a stock exchange. The activity of market makers consists in posting limit orders, that is, providing options for others to buy or sell at specific prices. In doing so, they face the risk of trading with informed investors endowed with more precise forecasts of the future value of the asset. Market makers can effectively limit this risk by extracting (part of) such private information from the orderflow, and updating their quotes accordingly.

My formulation of the prediction problem reproduces the challenge faced by market makers and consists in exploiting past orderflow to produce the best forecast of future transaction prices.

This setting constitutes an ideal laboratory to evaluate the performance and the interpretability of attention-based models vis-a-vis a realistic forecasting problem and, because the issue has been extensively explored by market microstructure theorists, it provides a rich set of testable hypotheses. In particular, I apply the proposed methodology to make inference on the marginal information content of limit orders versus market orders. Microstructure models typically assume that market orders are the most informative, because they are submitted by traders with superior information willing to trade fast and capitalize on their informational advantage. Inspection of the attention allocation of the trained neural network provides a data-driven method to validate or to discard such a hypothesis.

My experimental setup, inspired by the adverse selection problem faced by market makers, is further motivated by a number of additional reasons. First, given that modern market making activity takes place in electronic exchanges at very high frequency, an astonishing amount of data is available. This limits the problem of over-fitting and makes it possible to train deep neural networks featuring a large number of parameters. Another consequence of the increasing speed of market making activity has been the surge of academic interest on the impact of high-frequency traders (HFTs) on the functioning of financial markets. The debate has focused mainly on their effect on liquidity. On the one hand, HFT firms argue that their activity increases market liquidity by reducing bid-ask spreads, a claim that is supported, at least conditional on normal market conditions, by empirical evidence. On the other hand, investors and some observers in the financial press blame HFTs for back-running block trades by large institutions. One of the concerns is that, since large trades need to be split into smaller child executions over a non-trivial time frame, HFTs may quickly extract private information from the first part of the order and trade in the same direction of the institutional investors, thus increasing effective transaction costs and imposing a non-trivial externality to the originator. Alternatively, for liquidity-motivated trades, HFTs may recognize their presence in the orderflow, start trading in the same direction and revert the position before the end of the block, speculating on and adding to the temporary component of the price impact. This predatory behavior is costly for institutions, since it increases their effective cost of trading. My analysis sheds light on this important issue by studying the behavior of my model, a forecasting tool which is closely related to the algorithmic activity of HFTs, during the executions of large block trades by institutional investors.

The empirical analysis is based on high-frequency message traffic data from the NASDAQ electronic exchange. The data contain every single event in the limit order book (LOB) for a given stock, namely trade executions and the submission or cancellation of limit orders,

timestamped with micro-second precision. In particular, I use a large sample of LOB events for ten among the most liquid stocks traded in the exchange to train four distinct machine learning models, including the proposed attention-based model and alternative neural architectures which have been proven useful in this context.

It is self-evident that a necessary condition to make valid inference from the attention layer of my model is that the latter is capable of producing high-quality forecasts. If the model is not able to extract useful patterns from the data, then its attention allocation cannot be used to infer the informational content of different LOB events. Hence, in order to assess the forecasting accuracy of my attention-based network, I compare it to the benchmark models through different measures of out-of-sample performance.

Once the forecasting power of the attention-based model is validated, I proceed by carrying out a systematic analysis of the attention layer. First, I inspect the output of the attention-layer during every example in the test set, consisting of a sequence of LOB events and a target variable representing future transaction prices, to extract the level of attention assigned by the trained model to each step of the input time-series. Then, to systematically asses how the model focuses its temporal attention, I run panel regressions of the resulting attention levels on dummy variables indicating different types of LOB events, namely submissions and cancellations of limit orders and execution of market orders. The estimated coefficients from these regressions can be interpreted as the level of informativeness of each event type, with respect to the forecasting of future execution prices.

For the third part of the analysis, focused on liquidity, I complement order book data with transaction-level data for institutional investors provided by AbelNoser/Ancerno. This database reports transactions in the US equity market from large mutual funds and hedge funds, recording the identity of the funds with unique identifiers. This allows me to identify block orders, searching for multiple sequential executions on a single stock by a single institution summing up to a large volume.

I then train my model on LOB data from time windows preceding each of the identified block-order and use it to produce predictions during the event days. Then, to understand whether my model fosters back-running or predatory trading, I check if the sign of these predictions are systematically related to the side of the block orders.

My first set of results shows that my attention-based model delivers a significantly higher out-of-sample performance relative to the competing models, across different stocks and multiple trading days. Importantly, I provide evidence of a significant marginal contribution of the attention mechanism in improving the forecasting accuracy relative to an alternative specification of the model in which the attention layer is bypassed. This result demonstrates that the attention mechanism can effectively improve the information extraction from LOB events,

thus validating the identification assumption that the attention level assigned by the model to each input time-step can be used as a proxy for the informativeness of the underlying event. This is a necessary condition to justify the use the model to make inference and, therefore, it allows to proceed with the central part of the analysis.

The second set of results, obtained from the above-described panel regression approach, indicates that attention levels peak during trade executions, are mid-range during submission of new limit orders, and are significantly lower on cancellations of outstanding quotes. The coefficient estimates are robust to the inclusion of additional explanatory variables which are expected to be related to information, such as the speed of trading activity and the magnitude of mid-price changes. These findings are consistent with the standard assumption of theoretical microstructure models featuring asymmetric information, that is, that market orders are submitted by informed traders while limit orders come from uninformed market makers.

The third set of results shows that my model, in fact, behaves systematically differently in the presence of block orders. In particular, the sign of its predictions are negatively related to the side of the block, indicating that the model forecasts a higher profit for the market maker if she take the opposite side and provides liquidity to the institutional investor. This evidence is consistent with previous research showing that algorithmic trading improves liquidity and, further, it complements them focusing on the realized trading costs associated to institutional block orders rather then measuring only its effects on bid-ask spreads.

Finally, I demonstrate how the attention layer can be used to visually inspect the model's decisions on a case-by-case basis, adding an extra layer of interpretability to the output of the model. This can be done by superimposing the time-series of attention levels to a plot representing the sequence of input variables. Even though limited by its qualitative nature, this method can provide some insights on the inner workings of the model, both in situations where the forecast is correct and those where it is mistaken.

Related Literature

My attention-based model, which takes as input past orderflow data and produces accurate forecasts of future transaction prices, represents an instance of the broader class of technological innovations in the processing of demand data. More efficient information extraction from the orderflow allows to better distinguish between trades motivated by fundamentals and those driven by exogenous liquidity shocks. Uninformed demand shocks introduce a layer of noise in prices, obstructing the information aggregation role of financial markets (Black, 1986, Shleifer and Vishny, 1997). Information extraction technologies, and my model in particular, could help market participants filtering out the noisy component of the orderflow, leading to a more efficient extraction of fundamental information and increasing price informativeness.

Extraction of orderflow information is also related to market liquidity through the adverse selection channel. The bid-ask spread can be seen as the compensation required by market makers for the risk of trading against agents endowed with superior information on the fundamental value of the asset. For this reason, a standard prediction of microstructure models is that the width of the bid-ask spread is proportional to the level of information asymmetry between market makers and other market participants. But information can leak through prices, thus market makers have an incentive to track the orderflow in real time and try to deduce fundamental information, in an effort to reduce the asymmetry. My model is specifically trained to address this problem and could be of great value for market makers in their quest to reduce the level of information asymmetry. Therefore, to the extent that its usage marginally increases the ability of real-world market makers to forecast future transaction prices, the adoption of my attention-based model could lead to a reduction of bid-ask spreads and an improvement in market liquidity.

The speed in delivering forecasts based on real-time data represents a key factor to determine whether the model is applicable to real exchanges. To understand why this is the case, notice that the submission of limit orders by market makers effectively provides options for others to trade at the posted prices. If liquidity demanders can identify profitable in-the-money options arising from stale limit orders, they can pick them off and profit at the expense of the market maker. Assuming the pick-off risk is non-negligible and market makers require a compensation for it, this has the effect of widening bid-ask spreads and increasing execution costs. Consequently, if algorithms can reduce the cost of free trading options implicit in limit orders, then the level of adverse selection depends on the quality and efficiency of the algorithms employed by market makers (Foucault et al., 2013). Training a deep neural network on historical data takes a non-negligible amount of time and hardware resources, because of the large number of required data points and the computational cost of back-propagation passes. However, once the training process is over, the calculation of forecasts based on real-time data is extremely fast even on a desktop machine. Because they can be employed in high frequency contests, neural network models providing high-quality forecasts of future transaction prices have the potential to help liquidity providers to efficiently update their quotes and reduce the risk of being picked off. My attention-based model can therefore reduce adverse selection and increase market liquidity through this channel, allowing market makers to post narrower spreads in equilibrium.

The idea that technological innovation leads to more informative prices and improves market liquidity is supported by empirical evidence. For instance Hendershott et al. (2011) use a neat identification strategy, based on the introduction of automated quote dissemination in the New York Stock Exchange, to provide causal evidence supporting this view. The same conclusion arises from the empirical analysis of Chaboud et al. (2014), who exploit a long time

series of high-frequency data from the foreign exchange market. Nevertheless, the argument for liquidity improvements and price efficiency, often used by high frequency traders (HFTs) to justify their increasing presence in markets, has been widely criticized by academics and observers in the financial press. One concern is that the focus on orderflow analysis may lead to a reduction of investment in fundamental research in the long run, which in turn would decrease price informativeness and impede the efficient allocation of capital across firms. The technology growth model proposed by Farboodi and Veldkamp (2018) speaks to this issue, investigating a setting in which agents allocate their budget between fundamental and orderflow information. Their model implies that in the long-run, as information technology increases the total amount of information available to the agents (high-technology limit), fundamental analysis is not crowed-out by investment in orderflow information extraction and that price informativeness should consequently increase over time.

From a theoretical standpoint, the link between orderflow information extraction and liquidity relies on the assumption of information asymmetry between market markers and privately informed investors. Standard microstructure models, starting from the seminal Glosten and Milgrom (1985) and Kyle (1989), assume that market makers post limit orders based on the publicly available information while informed traders, to avoid the risk of losing their informational advantage, submit market orders which guarantee immediate execution. An implication of such an assumption is that market orders are more informative than limit orders in forecasting future price levels. This fact is challenged by the findings of Brogaard et al. (2016), who analyze LOB data from the Canadian equity market and, using a vector autoregression (VAR) model, show that price discovery occurs predominantly through limit orders. The authors argue that the reason why limit orders provide the majority of price discovery is that they are far more numerous than executions (market orders represent less than 5% of messages). In contrast, even though submissions and deletions of limit orders constitute the vast majority of events in the dataset, my analysis reveals that the attention-based model chooses to focus significantly more on execution of market orders. This result suggests that the information content of executed trades is significantly more important as a driver of LOB dynamics, consistent with the view that market orders reveal private information. I conjecture that the discrepancy between the findings and that of the above mentioned work may be due to the limitation of the VAR model.

This view is supported by the work of Sirignano and Cont (2018), who propose a deep neural network model with three LSTM layers, trained on order book data for a large cross-section of US equity stocks, to predict the sign of future price movements. They show that the deep learning approach delivers superior forecast accuracy compared to linear VAR models, uncovering the importance of nonlinear relations between state variables and price changes. Moreover, their results provide evidence of path-dependence in price dynamics and suggest

the existence of a universal and stationary price formation mechanism shared among different stocks.

In a recent stream of literature, machine learning methods have been applied to the problem of forecasting mid-price changes at short horizons using LOB data. A variety of model architectures have been proposed and studied, based on Support Vector Machines (SVMs), on recurrent neural networks (RNNs) like the long short-term memory (LSTM), on convolutional neural networks (CNNs) or a combination of a CNN layer followed by a RNN layer. [1]

More closely related to my approach, a limited number of recent works explore architectures augmented with attention mechanisms. Among these, Tran et al. (2018) propose a bilinear network augmented with an attention-based mechanism to predict mid-price changes for ten stocks traded in the Finnish stock exchange. While the attention layer they implement is similar to the one employed in this work, my model is different because it is based on an encoder-decoder architecture with an RNN as encoder, as described in Section 2. An alternative attention-based model trained on LOB data has been used by Mäkinen et al. (2018) to predict high frequency stock price jumps identified using the non-parametric test proposed by Lee and Mykland (2007). Their results show that an attention-based model achieves a higher accuracy than linear models, convolutional neural networks and plain LSTM in this task.

A different problem has been explored by le Calvez and Cliff (2018), who train a deep neural network on simulated LOB data to learn the trading behavior of a profitable algorithmic trader. They demonstrate that the model can fully reproduce the trading strategy of the trader and can generate even higher profits. Another related work is that of Nevmyvaka et al. (2006), who use millisecond LOB data from NASDAQ and a reinforcement learning model to address the problem of optimized trade execution. Their results show that such an approach to the problem can significantly reduce the transaction costs associated to the liquidation of a large position by an informed trader. My paper addresses a similar problem, but seen from the perspective of market markers providing liquidity to the market and facing adverse selection.

Differently from the objective of the work, however, the research outlined above has tended to focus on forecasting performance rather than interpretability and inference. One key con-

[1] The design of Zhang et al. (2019), for instance, is based on a CNN layer which allows the model to aggregate information from different levels of the LOB, computing weighted-averages with autonomously learned weights. As a second layer they employ an Inception Module (Szegedy et al., 2015) to extract local interactions over different time horizon. The resulting features are finally passed to an LSTM layer, which captures the dynamic temporal behavior. The work by Sirignano (2019) proposes a new neural network architecture for spatial distributions and uses it to model the joint conditional distribution of the future best bid and ask prices. Taking advantage of the local spatial structure of the LOB, the proposed spatial neural network based on bounded hidden units outperforms both a standard neural network and a logistic regression model.

tribution of my work is to provide a novel motivation for the application of machine learning methods in the context of academic research in finance, showing how attention-based models can effectively be used to make data-driven inference.

The concept of *attention* for neural network models was first introduced in the context of neural machine translation by the seminal paper of Bahdanau et al. (2014). The authors augment their network with a novel attention layer to allow the model to dynamically focus, while producing the translation for a specific word, on a subset of related words in the input phrase. Building on this idea, Vaswani et al. (2017) take the attention paradigm to the next level proposing a groundbreaking machine translation approach. Differently from previous state-of-the-art models for that task, their model is based entirely on attention mechanisms and does not include any recurrent nor convolutional layer. They show that the so called Transformer model achieves unprecedented levels of accuracy in English-to-French translation and proves significantly faster to train. Outside the context of machine translation, the attention-based approach has proven useful in a number of different tasks related to textual analysis[2], time-series prediction, and medical diagnosis, achieving high levels of both performance and interpretability[3].

Structure of the Paper

The rest of the paper is organized as follows. Section 2 describes the architecture of the attention-based model and briefly outlines the benchmark models. Section 3 defines the experimental setting, the input features, the target variable, and the training and evaluation procedures. Section 4 compares the model with the competing models with respect to forecasting accuracy. Section 5 uses the attention-layer to make inference on the marginal information content of LOB events and to improve visualizations of the model's decisions during specific examples. Section 6 presents the analysis on the model's behavior in the presence of block orders by institutional investors. Finally, Section 7 contains concluding remarks.

2 Model Architecture

In this section I propose and formalize a neural encoder-decoder model augmented with an attention-based mechanism, which I will apply to the liquidity provision problem described in Section 3.4. More specifically, the proposed model features an intra-attention (or self-

[2] For instance, the attention paradigm has been applied to text comprehension (Cheng et al., 2016, Parikh et al., 2016), textual entailment (Lin et al., 2017), relation classification (Zhou et al., 2016) and text summarization (Paulus et al., 2017).

[3] See for example Qin et al. (2017) and Riemer et al. (2016) for time-series and Choi et al. (2016) for medical diagnosis.

attention) layer, that is, an attention mechanism in which different positions of the same sequence are related to each others with the objective of producing a meaningful representation of the input time-series.

As depicted in Figure 1, the model consists of four layers. The input and the recurrent layers can be though as the *encoder*, while the *decoder* is composed by the attention and the output layers.

The encoder is a recurrent neural network (RNN) which aims to capture long-term temporal dependence in the sequence of LOB events. At each time step, the RNN is trained to produce a representation $\mathbf{h}_t$ of the current state. The so called *hidden state*[4] contains information on the most recent event together with information from past events. To generate the next hidden state $\mathbf{h}_{t+1}$, the RNN decides which part of the information to retain and which to forget. Ideally, the final hidden state $\mathbf{h}_T$ should contain all the information needed to predict the target variable. However, this forces the model to condense all the relevant information in a single vector of fixed length, potentially resulting into a bottleneck.

This problem can be tackled by introducing a decoder featuring an attention layer. This allows the model to additionally perform a soft-search on the entire sequence of the hidden states generated by the encoder, using the final hidden state $\mathbf{h}_T$ as key for the query. Intuitively, the decoder first looks at $\mathbf{h}_T$ to be aware of the current state. Then, before generating the final prediction, it looks *back in time* and extracts the additional information required to put the current state into context.

In the following sections I provide a formal description of the encoder and decoder layers. I implemented the model in python leveraging on the latest version of the TensorFlow framework (Girija, 2016) with eager execution.

2.1 Encoder

Given the input sequence $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)$, with $\mathbf{X}_t \in \mathbb{R}^N$ where N is the number of features and T is the number of lags, the encoder learns a representation mapping

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{X}_t) \tag{1}$$

where $\mathbf{h}_s$ is the hidden state of the encoder at time s and f is a nonlinear activation function which could be estimated by any recurrent neural network (RNN). In this work I use the gated recurrent unit (GRU) proposed by Cho et al. (2014). As an alternative, I also experimented with the long short-term memory (LSTM) network of Hochreiter and Schmidhuber (1997), obtaining very similar results.

[4] The term *hidden state* in the context of recurrent neural networks refers to a distinct concept than in the theory of Markov chains and should therefore not being confused with the latter.

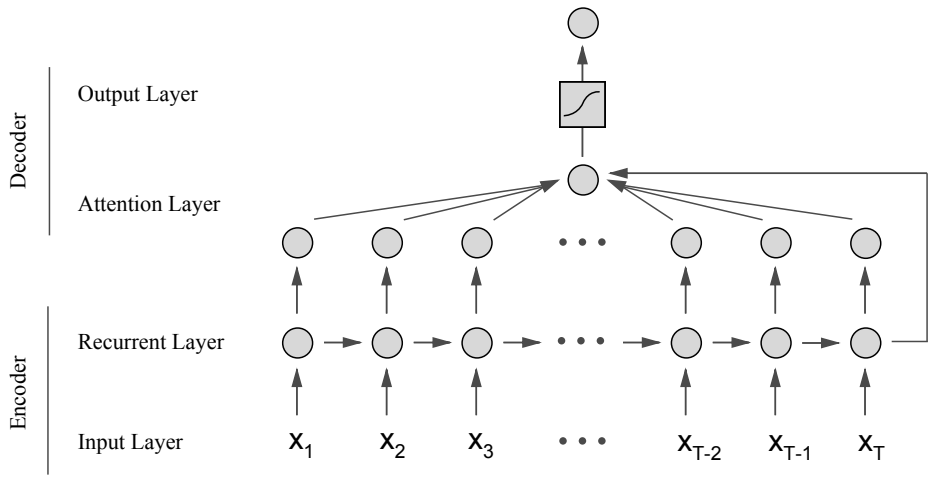


Figure 1: Model Architecture. The Figure provides a schematic for the architecture of my attention-based model, consisting of four layers. The input and the recurrent layers can be though as the *encoder*, while the *decoder* is composed by the attention and the output layers. The encoder is a recurrent neural network, which captures long term temporal dependence in the sequence of LOB events and encodes the input data into a sequence of hidden states. The decoder first employs an attention-mechanism and runs a soft-search on the hidden states generated by the encoder, using the last hidden state as key for the query. Then it uses the resulting attention distribution to construct the *context vector* as the attention-weighted average of the hidden states. Finally, to compute the prediction, the context vector and the last hidden states are fed to a fully-connected layer with hyperbolic tangent activation.

Even though my approach is robust to the choice of the encoder's RNN, both GRU and LSTM are a natural choice since they do not suffer from the vanishing gradient problem described in Hochreiter (1998) and Hochreiter et al. (2001). These RNNs, therefore, allow for the long-term time-series dependencies which are expected to be found in limit order book dynamics.

My GRU implementation is based on the first version of Cho et al. (2014), which has the advantage of being compatible with the cuDNN library proposed by Chetlur et al. (2014) and can thus be trained on a GPU. The activation of the step-t hidden state $\mathbf{h}_t \in \mathbb{R}^H$ is described as follows. First, the *reset* gates $\mathbf{r}_t = (r_t^1, \ldots, r_t^H)$ are computed as

$$\mathbf{r}_t = \sigma\left(\mathbf{W}^r \mathbf{X}_t + \mathbf{U}^r \mathbf{h}_{t-1} + \mathbf{b}^r\right) \tag{2}$$

where σ denotes the sigmoid function applied component-wise, while $\mathbf{W}^r$, $\mathbf{U}^r$ and $\mathbf{b}^r$ are parameters to be learned. Similarly, the *update* gates $\mathbf{z}_t = (z_t^1, \ldots, z_t^H)$ are computed as

$$\mathbf{z}_t = \sigma\left(\mathbf{W}^z \mathbf{X}_t + \mathbf{U}^z \mathbf{h}_{t-1} + \mathbf{b}^z\right) \tag{3}$$

where the matrices $\mathbf{W}^z$, $\mathbf{U}^z$ and the vector $\mathbf{b}^z$ are parameters to be learned. Finally, the

hidden state $\mathbf{h}_t$ of the GRU unit is computed as

$$\mathbf{h}_t = \mathbf{z}_t \circ \mathbf{h}_{t-1} + (\mathbf{1} - \mathbf{z}_t) \circ \widetilde{\mathbf{h}}_t \tag{4}$$

$$\widetilde{\mathbf{h}}_t = \tanh\left(\mathbf{W}^h \mathbf{X}_t + r_t \circ \mathbf{U}^h \mathbf{h}_{t-1} + \mathbf{b}^h\right) \tag{5}$$

where $\circ$ denotes the Hadamard product and $\mathbf{W}^h$, $\mathbf{U}^h$ and $\mathbf{b}^h$ are parameters to be learned.

2.2 Decoder

The decoder takes as input the last cell state $\mathbf{s}_T$ and the entire sequence of hidden states $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_T)$ generated by the encoder[5]. These are processed through a temporal attention mechanism, defined as follows, which selects the most relevant encoder hidden states across time. The vector $\mathbf{e}$, containing the temporal attention weights associated to each time step, is a non-linear transformation of the last encoder cell state $\mathbf{s}_T$ and the full sequence of hidden states. The resulting vector is then standardized using the component-wise softmax function, obtaining the attention-distribution $\boldsymbol{\alpha}$

$$\mathbf{e} = \mathbf{v}^\top \tanh(\mathbf{W}^a \mathbf{s}_T + \mathbf{U}^a \mathbf{h} + \mathbf{b}^a)\,, \tag{6}$$

$$\boldsymbol{\alpha}_t = \frac{\exp(\mathbf{e}_t)}{\sum_{s=1}^{T} \exp(\mathbf{e}_s)}, \quad t = 1, \ldots, T\,, \tag{7}$$

where the vectors $\mathbf{v}$ and $\mathbf{b}^a$ and the matrices $\mathbf{W}^a$ and $\mathbf{U}^a$ are parameters to be learned.

Next, the context vector $\mathbf{c}$ is calculated as the attention-weighted average of the hidden state vectors, where each state $\mathbf{h}_s$ is weighted by the corresponding attention level $\boldsymbol{\alpha}_s$

$$\mathbf{c} = \sum_{s=1}^{T} \boldsymbol{\alpha}_s \mathbf{h}_s\,. \tag{8}$$

The attention weight $\boldsymbol{\alpha}_t$ can thus be interpreted as the importance, in determining the prediction, of the information content of the encoder hidden state at time t. Finally, the prediction is computed feeding the context vector $\mathbf{c}$ and the last hidden state $\mathbf{h}_T$ to a fully-connected network with one layer and hyperbolic tangent activation function

$$\hat{y} = \tanh(\mathbf{F}(\mathbf{c}, \mathbf{h}_T) + \mathbf{b}), \tag{9}$$

where the matrix $\mathbf{F}$ and the vector $\mathbf{b}$ are parameters to be learned.

The attention mechanism is useful in this context because it does not force the model to encode a whole sequence of LOB events into a single fixed-length vector. Rather, it encodes the input

[5] Notice that in the case, since I use a GRU as the RNN layer, I have $\mathbf{s}_T = \mathbf{h}_T$.

into a sequence of vectors and adaptively over-weights a subset of these vectors to produce the prediction. This frees the model from having to squash all the information contained in the order book dynamics into a fixed-length vector. In a sense, the models learns to choose what parts of the input sequence are most meaningful as a context to the current state, and selectively predicts an output based mostly on those parts. Moreover, once the model is trained, the attention mechanism allows for a post-hoc analysis of the model's decision on each particular situation. In Section 5 I exploit this feature to systematically asses what types of LOB events are most relevant for the price formation process and to qualitatively analyze the model's predictions on specific examples.

2.3 Benchmark Models

I compare the proposed encoder-decoder attention-based model with alternative architectures proposed by the literature. These alternative models have been applied to the task of predicting future mid-price movements using sequences of LOB events, a slightly different problem from the one I study in this work, where the target variable includes also information about liquidity. Nevertheless, since previous research proved these architectures effective at extracting information from past order book dynamics, they constitute a natural benchmark to evaluate the performance of my model.

Plain-Vanilla Neural Network
A fully-connected neural network with a single layer and hyperbolic tangent activation function. The input is a $(25 \times T)$-dimensional array, containing the entire panel of lagged features over the previous T events flattened into a single vector. This model is the simplest among those I consider, as it does not feature a recurrent layer and thus is not expected to handle time-series dependence efficiently.

CNN-LSTM (DeepLOB)
This model, proposed by Zhang et al. (2019), features a convolutional layer aimed at capturing the spatial structure of the LOB followed by a recurrent layer for long-run time dependency. The convolutional layer is composed by a number of sub-layers, each containing 16 filters of different sizes and strides, followed by an inception module with blocks of 32 filters. The output is then passed to the recurrent layer, consisting of a LSTM network with 64 hidden units. I replicate the model using the TensorFlow backend and replace the softmax activation of the final fully-connected layer with a tanh function, since the task analyzed in this paper is a regression and not a classification problem.

Encoder-Decoder With No Attention
This model is nested into my encoder-decoder attention-based model, the difference being that the attention layer is completely bypassed. Hence the encoder is exactly the same, while the

decoder computes the final activation simply as $\hat{y} = \tanh(\mathbf{F}\mathbf{h}_T + \mathbf{b})$ rather than using equation (9). Including this models to the benchmark allows to infer the marginal improvement in performance due to the attention layer.

3 Experimental Design

This section describes the experimental setting underlying this work. I provide information on the employed dataset, outline the construction of the feature and the target variables, provide a formal statement of the prediction problem and finally describe the training and evaluation procedures. The experimental results on the out-of-sample performance the four models described in Section 2.3 are reported in Section 4.

3.1 Dataset

I collected data for 10 of the most liquid stocks traded in the NASDAQ exchange[6], for the period from February 2019 to March 2019. The data is provided by Lobster Data and is based on the NASDAQ Historical TotalView-ITCH database. Even though the sample spans only a relatively period in calendar time, given the high-frequency nature of the LOB dynamics it contains more than 47 million observations.

Each entry of the dataset represents an event (or *message*) regarding the first five levels of the LOB, on both the ask and the bid side, time-stamped with microsecond precision and classified into three main types: *deletion* events are recorded when a limit order is (fully or partially) canceled by the originator and removed from the LOB; *execution* describe trades, that is, when a limit order is matched with an incoming market order and executed; *submission* events indicate that a new limit order is posted or an existing limit order is updated. Each event represents a change in the LOB and, accordingly, the dataset provides a snapshot of the new state implied by that event, including the price levels and the aggregate sizes of the limit orders outstanding in the first five levels of the LOB.

Table 1 presents the average number of events per minute for each of the sample stocks, broken-down by event type. Microsoft is the most active stock, with more than 3000 events fired on average for each minute of the trading day. Other stocks, e.g. Tesla or Adobe, are significantly less active but nevertheless present more than 500 events per minute. I note that only a small fraction of events represents executions, while the vast majority refers to submission and cancellation of new orders.

[6] American Airlines (AAL), Apple (AAPL), Adobe (ADBE), Comcast (CMCSA), Ebay (EBAY), Facebook (FB), Microsoft (MSFT), Nvidia (NVDA), Pepsico (PEP) and Tesla (TSLA).

	AAL	AAPL	ADBE	CMCSA	EBAY	FB	MSFT	NVDA	PEP	TSLA
Submissions per Minute	438.2	1074.9	264.2	583.4	499.6	643.8	1472.8	378.5	304.0	256.5
Deletions per Minute	449.7	1053.8	245.2	621.0	525.4	616.5	1532.8	362.7	306.0	214.4
Executions per Minute	16.1	62.1	24.0	14.2	13.7	65.4	53.4	49.4	20.2	75.8

Table 1: Frequency of Events by Type. The table presents the average number of LOB events per minute, by event type. In *deletion* events, a limit order is (fully or partially) canceled by the originator and removed from the LOB; in *execution* events, a limit order is matched with a market order and executed; in *submission* events, a new limit order is posted in the LOB or an existing limit order is updated.

3.2 Features

In this section I describe the procedure to construct the 25 features variable from the raw LOB data fed to the benchmark supervised learning models. My objective is to avoid polluting the features set with ratios between variables, moving averages or other arbitrarily constructed quantities. On the contrary, I select the minimum set of variables which in principle allows a full description of the LOB dynamics. I make sure that each feature has a straightforward interpretation, with the objective to improve the intelligibility of the considered models, in the sense of Lipton (2016). Moreover, this choice allows me to perform a post-hoc visual analysis of the proposed attention-based model in Section 5.

As in most machine learning settings involving time-series data, an important aspect of pre-processing is to ensure the stationary of the employed features. This is a necessary condition to allow models to generalize the patterns learned in the training set and to achieve a satisfactory out-of-sample performance. The issue is particularly severe when the objective is to forecast financial data, given the high level of non-stationarity and the multi-modal nature of the variables involved.

I develop a novel approach to guarantee the informational quality of the resulting features. First of all, following Tsantekidis et al. (2018), I avoid using z-score standardization for price and size variables. This choice prevents me to run into forward-looking bias (if full-sample statistics are used) or to be forced to choose arbitrary windows (if backward-looking statistics are computed from rolling-windows). Moreover using past observations to standardize future quantities may lead to a distortion of the latter. For instance, suppose the outstanding liquidity is particularly high on day $d-1$. If one uses the previous day average to standardize order sizes on the next day, this would result into under-stated levels and variability of the features representing order sizes for day d.

An alternative approach to standardization is proposed by Passalis et al. (2019), who develop a novel neural layer with the specific objective of dealing with this problem. Their adaptive layer learns to efficiently normalize the raw LOB data to be passed as input to a plain

or convolutional neural network. My approach is different in that I do not build an input normalization layer into the model.

I construct features representing the relative distribution of available liquidity across the first 5 levels of the LOB[7]. For each event t, let $as(\ell, t)$ and $bs(\ell, t)$ denote the number of shares on offer in level ℓ of the ask and the bid side of the LOB, respectively. I define the corresponding features as the ratio to the total amount of outstanding liquidity

$$AS(\ell, t) = \frac{as(\ell, t)}{\sum_{l=1}^{5} as(l, t) + bs(l, t)} \tag{10}$$

$$BS(\ell, t) = \frac{bs(\ell, t)}{\sum_{l=1}^{5} as(l, t) + bs(l, t)} \tag{11}$$

This results into 10 features with unitary sum, describing the shape of the distribution of available liquidity in the LOB at each point in time.

I complement the above information with 10 additional features describing the offer price for each order book level. Let $ap(\ell, t)$ and $bp(\ell, t)$ be the raw offer prices associated to level ℓ of the ask and bid sides of the LOB, respectively. Further, let $m(t) = (ap(1, t) + bp(1, t))/2$ denote the prevailing mid-price at time t. I define the price-level features as the relative distance of each offer price from the mid price, expressed in percentage points

$$AP(\ell, t) = \frac{ap(\ell, t) - m(t)}{m(t)} \times 10^2 \tag{12}$$

$$BP(\ell, t) = \frac{m(t) - bp(\ell, t)}{m(t)} \times 10^2 \tag{13}$$

The resulting set of 10 features convey information on the level-ℓ *spreads* posted by market makers in the LOB at each point in time. Hence, their first differences can be used to infer the widening or shrinking of the spreads over time. However, these features fail to capture information about the price level. I address this issue by adding an additional feature describing the mid-price movement relative to the previous message, expressed in basis points

$$\text{PriceChange}(t) = \frac{m(t) - m(t-1)}{m(t-1)} \times 10^4 \tag{14}$$

Next, I construct two dummy variables categorizing message types. The first dummy, 'Execution', indicates a transaction being executed, that is, when (part of) an incoming market order is matched with an outstanding limit order. As in Zhang et al. (2019), I do not distinguish between the execution of *visible* and *hidden* limit orders. The dummy 'Deletion' indicates the

[7]It turns out that augmenting the depth to 10 or more levels does not significantly increase predictive power, consistent to the results of Zhang et al. (2019).

removal or the cancellation (partial removal) of a limit order from the outstanding LOB. The complementary set of messages corresponds to the submission of new limit orders.[8]

I further add to the set of features the binary variable 'Direction', taking values in $\{-1, 1\}$ and providing the directionally of the LOB message, i.e positive (negative) values indicate a change in the bid (ask) side of the order book. In particular, for executions, positive (negative) values of 'Direction' indicate seller-initiated (buyer-initiated) trades.

Finally, to provide information on the frequency of the trading activity in real time, which are expected to be relevant given the significant clustering of trading activity over time observed in the data and reported by Menkveld (2018). I hence define the feature 'Elapsed Time' as $\log(1+\tau)$, where τ measures the seconds in calendar time which have passed from the previous message.

Summary statistics on the above described features are reported in Table 6, separately for each of the sample stocks. I notice that the distribution of time intervals between consecutive events is significantly right-skewed. This signals the above mentioned clusters of high-frequency trading activity in the data. Even if not visible from the table, I report here that the median time interval is less than 1 millisecond for every stock and that the fastest events unfold at the order of micro-seconds. As a final remark, I highlight the fact that only a small fraction of the messages are triggered by executions. In fact, the share of executed transactions ranges between 1% and 6% across all sample stocks but TSLA, exhibiting a significantly higher fraction of executions.

3.3 Target Variable: Market Maker's Expected Profit

In electronic limit order markets, like the NASDAQ, market makers post ask and bid quotes in the LOB. These can later be crossed with incoming market orders, giving rise to transactions. Doing so the market maker provides liquidity to other market participants and she is compensated by earning the bid/ask spread. However she faces the adverse selection problem arising from orders submitted by agents trading on private information (Glosten and Milgrom, 1985, Kyle, 1989). Intuitively, posting a sell (buy) limit order is similar to underwriting a call (put) option with strike price equal to the quote. Informed speculators may exercise these options and "pick off limit orders" when these become stale after the arrival of new information (Copeland and Galai, 1983). It is therefore key for market makers to efficiently extract information from LOB events to quickly update their quotes.

The target variable I define below is tailored to capture the payoff of a strategic market maker

[8]This is the case because I drop observations with positive 'trading halt' indicator. I do not add any indicator for submission events to avoid redundancy.

posting buy market orders and facing adverse selection. More specifically, it measures the profit earned (or the loss suffered) by a market maker whose bid is matched to an incoming sell market order, posing that she wants to minimize her risk exposure by closing the long position on the stock as fast as possible. I assume the market maker (passively) buys the stock at the prevailing bid price and, immediately afterwards, she tries to unload the position by posting sell limit orders at the prevailing ask price. In case she finds a counterpart buying the stock at a reasonably high ask price, she makes a profit equal on average to the bid/ask spread. Otherwise, if she does not find a buyer within some pre-determined time range, I assume she actively unloads the position submitting a market order at the bid price.

Formally, for each ordered sequence of LOB events $[t_0, t_1]$ let $\mathcal{B}(t_0, t_1)$ denote the subset of buyer-initiated executions during that interval, that is,

$$\mathcal{B}(t_0, t_1) = \{\tau \in [t_0, t_1] \text{ s.t. } \text{Direction}(\tau) \times \text{Execution}(\tau) = -1\} \tag{15}$$

Next recall that $ap(1, t)$ and $bp(1, t)$ denote the best offer prices on the ask and bid sides of the LOB, respectively. I define the *minimal execution price*, that is, a proxy for the price at which the market maker can actively close a long position, as

$$\mathcal{P}_1(t_0, t_1) = \begin{cases} \min\limits_{\tau \in \mathcal{B}(t_0,t_1)} & ap(1, \tau) \quad \text{if } \mathcal{B}(t_0, t_1) \neq \varnothing \\ & bp(1, t_1) \quad \text{otherwise} \end{cases} \tag{16}$$

Finally, setting the prediction horizon to $H > 0$, I define the time-t target variable as

$$\mathscr{T}(t, H) = \frac{\mathcal{P}_1(t, t + H) - bp(1, t)}{bp(1, t)} \times 10^2 \tag{17}$$

Notice that $\mathscr{T}(t, H)$ provides a lower-bound for the percentage return earned by a market maker who provides liquidity to an incoming sell order at time t and closes the resulting long position within the H following LOB events. In reality, a sophisticated market maker could increase her profit by selling at a higher price, e.g. at the *maximum* executed sell order rather that at the *minimum* as in equation (16). I nevertheless decide to be conservative, so that $\mathscr{T}(t, H)$ is more effective at measuring the down-side risk of the market-making strategy.

3.4 Problem Statement

The prediction problem I analyze in this paper is motivated by the situation faced by a human market maker providing liquidity in a limit order market. More specifically I focus on one side of the decision-making process, assuming that the market maker acts as a counterpart to *sell* market orders. Even though the symmetrical problem of liquidity provision to *buy* market orders is equally interesting, I choose to focus on the sell side because of its connection with the issue of flash crashes (Kirilenko et al., 2017, Easley et al., 2011).

As noted in the above section it is crucial for a market maker to extract information from LOB events in a fast and efficient fashion, to reduce the risk of being "picked off" by speculators in case of the arrival of new information. My objective is to apply supervised learning models to help dealing with such a task. The input to the model is the time-series describing the most recent order book dynamics, represented by the features defined in Section 3.2. The models are trained to use this information set to predict the future return earned by the market maker, as defined in Section 3.3.

I now formally state the prediction problem studied in the paper. Let $F(\tau) \in \mathbb{R}^{25}$ denote the vector of features described in Section 3.2, at time τ. Setting the number of lagged features to $T > 0$, the sequence of features that will be used as input to produce a prediction for time t is given by $\mathscr{F}(t, T) = [F(t - T), \ldots, F(t - 1)]$. The task of the candidate machine learning model is predict the target variable $\mathscr{T}$ using the backward-looking information set $\mathscr{F}$. In other words I want the model to best approximate the function $\mathscr{F}(t, T) \mapsto \mathscr{T}(t, H)$ for each example indexed by t in the dataset. For the experimental results described in Section 4, I set the number of lags to $T = 50$ and the prediction horizon to $H = 100$.

Notice that the setting is fundamentally different from the problems explored by the recent literature on machine learning methods applied to LOB data, in two dimensions. First the target variable I try to predict is not based only on the future change in the mid-price as in Lee and Mykland (2007), Sirignano (2019), Passalis et al. (2019). In particular, the target variable contains information on both price levels and executions, taking into account the future orderflow hitting the LOB. This aspect is key for any predictive model which aims to be helpful for real-world market makers. Second the target variable is continuous, which allows to state the prediction task as a regression problem rather than a classification problem as in the above cited works. In this regard, the approach has the advantage to deliver a more informative evaluation of the models' performance and, further, it does not require to choose arbitrary thresholds to define categorical variables.

3.5 Training Procedure

This section describes the training and evaluation procedures, employed for each of the four supervised learning models considered in the study.

The training period ranges from the 22nd to the 27th of February 2019, while the 28th of February is used as validation. The test period goes from the 1st to the 7th of March 2019. I use only data recorded during trading hours, from 9:30 AM to 16:00 PM. To be conservative, I further remove the first and the last minute of each trading day. This ensures that lagged features do not belong to the previous trading day.

All models are trained separately for each sample stock using stochastic gradient descend (SGD) and the Adam optimizer proposed by Kingma and Ba (2014). A training epoch consists of 50 batches, with each batch containing 1024 examples. I use the same random seed to generate the training batches, ensuring that differences in the models' performance are not driven by random differences in the training process.

Since all the models are smooth and differentiable, parameters can be learned by standard back propagation with mean squared error (MSE) as the loss function

$$\mathscr{L} = \frac{1}{N} \sum_{t} (y_t - \hat{y}_t)^2 \tag{18}$$

where N is the sample size. To limit over-fitting, the end of the training is triggered by an early stopping rule (Prechelt, 1998) based on the validation MSE, with patience parameter set to 5 epochs. The training and prediction procedures are run on a machine featuring a *NVIDIA Tesla P4* GPU unit, which allows a considerable speed-up relative to CPUs. The process for the 10 sample stocks took about 5 hours to run.

4 Experimental Results

In this section I report experimental results on the prediction problem described in Section 3.4. As explained above, the target variable is a proxy for the profits earned by a market maker who provides liquidity to incoming sell market orders. I compare the out-of-sample performance of the four models outlined in Section 2, using data for 10 stocks traded in the NASDAQ exchange. The test period spans one trading week, from the 1st to the 7th of March 2019, and comprises roughly 26 million test samples.

I compare each model's performance using two standard evaluation metrics for regression problems. The first metric is the Root Mean Squared Error, defined as

$$RMSE = \left(\frac{1}{N} \sum_{i=i}^{N} (y_i - \hat{y}_i)^2 \right)^{1/2} \tag{19}$$

where y_i is the target variable described in Section 3.3, expressed in basis points, $\hat{y}_i$ is the model's prediction based on backward-looking features and i runs over the samples of the test set for a given stock. The second metric is the Fraction of Variance Unexplained $FVU = 1 - R^2$, where R^2 is the coefficient of determination of a linear regression of the realized target variable y_i on the model's prediction $\hat{y}_i$. Regressions are run at the stock-level over the test data using a standard OLS estimator.

Results are reported in Table 2, where the two panels display the RMSE the FVU, respectively. Panel (a) shows that my attention-based model achieves, on average, lower out-of-sample

	Attention (1)	No Attention (2)	DeepLOB (3)	Neural Net (4)
AAL	**3.334**	3.350	3.346	3.464
AAPL	**0.986**	0.988	0.989	1.063
ADBE	**3.044**	3.123	3.095	3.111
CMCSA	1.661	**1.658**	1.668	1.767
EBAY	**2.007**	2.055	2.018	2.151
FB	**1.599**	1.607	1.604	1.621
MSFT	**0.970**	0.976	0.979	1.045
NVDA	**3.009**	3.038	3.068	3.078
PEP	**1.456**	1.460	1.466	1.509
TSLA	**4.454**	4.590	4.594	4.623
Average	**2.252**	2.284	2.283	2.343

(a) Root Mean Squared Error

	Attention (1)	No Attention (2)	DeepLOB (3)	Neural Net (4)
AAL	**93.85%**	94.83%	94.45%	97.57%
AAPL	**97.49%**	97.82%	98.15%	99.41%
ADBE	**93.62%**	98.94%	96.81%	97.04%
CMCSA	79.30%	**79.14%**	79.72%	88.07%
EBAY	**83.78%**	87.89%	84.73%	92.51%
FB	**96.73%**	97.77%	97.30%	96.85%
MSFT	**91.76%**	93.15%	93.73%	97.23%
NVDA	**95.89%**	97.13%	98.52%	97.70%
PEP	97.67%	98.21%	99.02%	**97.04%**
TSLA	**91.90%**	97.59%	97.71%	95.92%
Average	**92.20%**	94.25%	94.01%	95.93%

(b) Fraction of Variance Unexplained

Table 2: Experimental Results. The table reports out-of-sample experimental results for 10 stocks traded in the NASDAQ exchange. Four predictive models are compared: (1) 'Attention' is the attention-based encoder-decoder model proposed in this paper; (2) 'No'Attention' is similar to the encoder-decoder model (i), but with the attention layer bypassed; (3) 'DeepLOB' is the model proposed in Zhang et al. (2019), based on a convolutional layer followed by a LSTM layer; (4) 'Neural Network' is a plain-vanilla fully-connected neural network. The root mean squared error computed as $RMSE = (\sum_i (y_i - \hat{y}_i)^2/N)^{1/2}$, where y_i is the target variable described in Section 3.3, expressed in basis points, and i runs over the samples of the test set. The fraction of variance unexplained is computed as $FVU = 1 - R^2$, where R^2 is the coefficient of the determination from a linear regression of the target variable on the model's predictions, estimated on the test set.

prediction errors relative to the benchmark models. This is the case for the vast majority of the sample stocks, with the exception of CMCSA for which the lowest RMSE is obtained with the plain encoder-decoder model. I notice that the CNN-LSTM model (DeepLOB) achieves an accuracy which is virtually identical to that of the encoder-decoder model without attention layer, while the plain-vanilla neural network performs significantly worse.

A similar pattern is observed in Panel (b) for the FVU metric. The attention-based model is is able to explain a larger fraction of the variability in the target variable for most of the sample stocks, relative to the benchmark models. As for the RMSE the average performance of the DeepLOB model is comparable to that of the encoder-decoder model with no attention layer, while the plain-vanilla neural network performs significantly worse.

For some of the sample stocks in particular, the increase in prediction accuracy is substantial. Considering e.g. the ADBE stock, the rightmost panel of Table 2 shows that the attention-based model is able to explain more than 6% of the variation in the target variable. Remarkably, this translates roughly into a two-times increase relative to the amount of variation explained by the second-best model, that is the CNN-LSTM network by Zhang et al. (2019).

These results show that my attention-based model is superior, in terms of prediction accuracy, to the benchmark models with respect to the regression problem considered in this paper. In particular, they demonstrate that the addition of the attention-layer to the encoder-decoder network generates a significant improvement. This result supports the intuition that the attention layer provides more flexibility to the model, freeing it from the need to condensate the relevant information on past LOB events in a single vector of fixed dimension. Moreover it implies that the model has learned how to allocate its attention to the most relevant events of the input time-series, thus justifying the identification assumption at the base of the inference exercise described in the next section.

5 Making Use of Attention

This section explores the intelligibility improvements resulting from the inclusion of an attention layer to deep neural network models. I first present a novel methodology which, in principle, allows to make data-driven inference from any neural network trained on time-series data and augmented with an attention-layer. I then provide a showcase application in the context of market microstructure, in which the trained model is used to make inference on the relative information content of market orders and limit orders. Finally, I further exploit the attention layer to improve the visualization of the neural network's decision-making process.

5.1 Inference Methodology

In this Section I describe my novel inferential methodology and apply it to the context of information dissemination though orderflow. The idea is based on the following observation on the functioning of neural networks augmented with an attention layer. In order to forecast a given target variable using time-series data, the model has to *jointly* learn two distinct tasks. The first one is to identify in the input data the most relevant information with respect to the forecasting problem, while the second one is to process such information to actually produce the forecast.

Standard deep neural networks[9] perform these two tasks by first finding a sequence of non-linear transformations of the input space, one for each of the layers, and then using the transformed data to generate the final prediction. Deep networks have been empirically proven to be highly efficient and, in fact, the ability to transform input data over multiple steps is one of the most prominent rationalization for the adoption of a deep architecture. The drawback

[9] Neural network are considered *deep* if they contain at least four hidden layers of neurons between input and output.

is that the large number of required neurons and the fully connected network linking them together generate a high level of complexity, making it very difficult for humans to comprehend and make sense of the transformations chosen by the network. The situation becomes even more complicated when dealing with time-series data, where recurrent neural networks have been showed to be the best-performing architecture.

In an attempt to tackle this intelligibility problem, I argue that adding an attention layer to neural networks makes it possible to extract at least some information on their inner configuration. The reason is that attention-based neural networks applied to time-series forecasting are designed to highlight the most relevant parts of the input sequence by assigning to them higher levels of attention. Even though it does not help us to shed light on the transformations performed by the deep part of the network, the attention layer provides insights on the relative importance of each time-step in a clean and intelligible fashion.

Such a feature of attention-based models provides a straightforward way to make inference on the relative information content of different events in the context of time series forecasting. First, the researcher recovers the attention vector $\boldsymbol{\alpha}$ defined in equation (6) and generated by the trained model for each sample e of the test set. This gives a set of attention levels $y_{e,t}$ associated to time-step t of the input of e. Second, she constructs a set of categorical (or continuous) variables $X_{e,t}$ describing any characteristic of interest associated to time-step t of the input sequence of e. Finally, the researcher pools the resulting data from the entire test set and runs a panel regression of the attention levels $y_{e,t}$ on the characteristics $X_{e,t}$. The estimated beta coefficients can be interpreted as the relative importance of each of the characteristics of interest, thus making it possible to make inference on the value of their informational content.

One of the advantages of this approach is that the inference is based on linear regression estimators. First of all, results from a panel regression are relatively easy to interpret and communicate. Moreover, regressions provide out-of-the-box confidence intervals and p-values (potentially based on clustered standard errors) and a number of other standard metrics. Further, it allows to easily perform multi-variate analysis and include specifications with fixed effects, thus enabling the researcher to properly perform inference.

5.2 Market Orders versus Limit Orders

I now apply the inference methodology described in Section 5.1 to test if certain kinds of LOB event are more informative than others for price discovery, that is if they are more useful to construct a forecast of future prices. The list of LOB event types I consider is exhaustive and includes: (i) the execution of a market order; (ii) the submission of a new limit order or the update of an outstanding limit order; (iii) the full or partial cancellation of an outstanding limit

order. As argued in the introduction, standard microstructure models typically assume that market orders are the most informative events for price discovery, because they are submitted by traders holding private information and, therefore, willing to trade fast and capitalize on their informational advantage. The analysis of this section uses the the attention allocation of my attention-based model to provide a data-driven method to validate or reject such an assumption. Formally, I want to test validity of the following hypothesis:

H0: *Market-order executions are more informative than the submission or cancellation of limit orders with respect to price discovery, that is to the forecasting of future prices.*

In the empirical microstructure literature, a standard approach to test the above hypothesis is to fit a Vector Auto Regression (VAR) model to LOB data and perform a variance-decomposition exercise (Hasbrouck, 1991a,b, Brogaard et al., 2014, 2016). A limitation of such a methodology arises from the fact that the VAR is a linear model and, most importantly, that it has been shown by Sirignano and Cont (2018) to under-perform the forecasting ability of non-linear neural network models. My methodology provides an alternative approach based on neural networks, thus overcoming the limitations of the VAR model and allowing for an arbitrary data-driven parametrization of the relationship between LOB dynamics and future prices.

To test the null hypothesis **H0**, I first recover the attention levels produced by the trained model for each of the 51200 examples in the test set. I then construct three dummy variables indicating the three event types, thus categorizing each of the 50 input time-steps of each example as *execution*, *submission* or *cancellation*. Per-minute summary statistics of such a classification are reported in Table 1. Finally, I run the panel regressions described by

$$\begin{aligned} \text{Attention}_{e,t} \;=\; \alpha \;&+\; \beta_0 \text{ Event Type}_{e,t} \\ &+\; \beta_1 \text{ Elapsed Time}_{e,t} \\ &+\; \beta_2 \mid \text{Price Change}_{e,t} \mid \;+\; \varepsilon_{e,t} \end{aligned}$$

where $\text{Attention}_{e,t}$ is the attention level associated to time-step t of the input sequence of example e and $\text{Event Type}_{e,t}$ is one the above described dummies, depending on the specification. Controls include $\text{Elapsed Time}_{e,t}$ and $\mid \text{Price Change}_{e,t} \mid$, measuring the time elapsed and the absolute mid-price change, respectively, from the previous LOB event. Standard errors are conservatively clustered by Day $\times$ Stock to account for potential serial correlation in residuals.

Results, reported in Table 3, show that the coefficient on the dummy indicating executions is positive and highly statistically significant. This implies that attention levels are significantly higher during the executions of market orders (+6%) and, in turn, that the model over-weights

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Attention Level	Attention Level	Attention Level	Attention Level	Attention Level	Attention Level
Execution	6.00***			5.93***		
	(5.63)			(5.59)		
Cancellation		-1.46***			-1.45***	
		(-7.96)			(-7.98)	
Submission			0.88***			0.88***
			(5.72)			(5.71)
Elapsed Time				1.81***	2.06***	2.08***
				(3.62)	(4.22)	(4.26)
|Price Change|				0.24***	0.56***	0.72***
				(3.35)	(4.17)	(4.82)
Constant	1.85***	2.71***	1.57***	1.78***	2.62***	1.47***
	(66.26)	(30.11)	(20.93)	(58.83)	(30.01)	(18.14)
Observations	2,560,000	2,560,000	2,560,000	2,560,000	2,560,000	2,560,000
R-squared	0.03	0.02	0.01	0.03	0.02	0.01
SEs Clustered By	Date×Stock	Date×Stock	Date×Stock	Date×Stock	Date×Stock	Date×Stock

Table 3: Attention Drivers. The table reports results from a panel regressions at the example-event level of attention levels onto dummy variables indicating three types of LOB events (executions, submissions and cancellations). For each example in the test dataset, attention levels for each time-step of the input sequence are extracted from the attention-based model described in Section 2 trained on the training set as described in Section 6.2. In columns (4), (5) and (6) controls for the time elapsed and the absolute price change from the previous LOB event are added to the regression. t-stats are reported in parentheses, based on standard errors clustered at the Day × Stock to account for potential serial correlation in residuals. Asterisks denote significance levels (***= 1%, **= 5%, *= 10%).

the corresponding time-steps to produce the final predictions. On the contrary, the negative coefficient on the cancellation dummy implies that the model allocates less attention to cancellation events. Submission events are somewhat in between, with a slight but significant increase in attention of less than 1 percentage points relative to the average.

Importantly, these coefficients are stable when controls are added to regression in specifications (4), (5) and (6). These results ensure that the heterogeneity in attention levels is not solely motivated by bid-ask bounce effects or by illiquidity. I can therefore rule out an alternative explanation based on the fact that executions may mechanically induce predictability because of the bid-ask bounce.

All in all, the results of this analysis cannot reject the null hypothesis **H0**. In other words, executions of market orders are by far the most informative events in LOB dynamics to forecast future transaction prices. Even though submission of limit orders are also deemed

informative, the model assigns a level of attention that is an order of magnitude lower with respect to executions. Cancellations of outstanding limit orders, instead, seem to play a limited informational role.

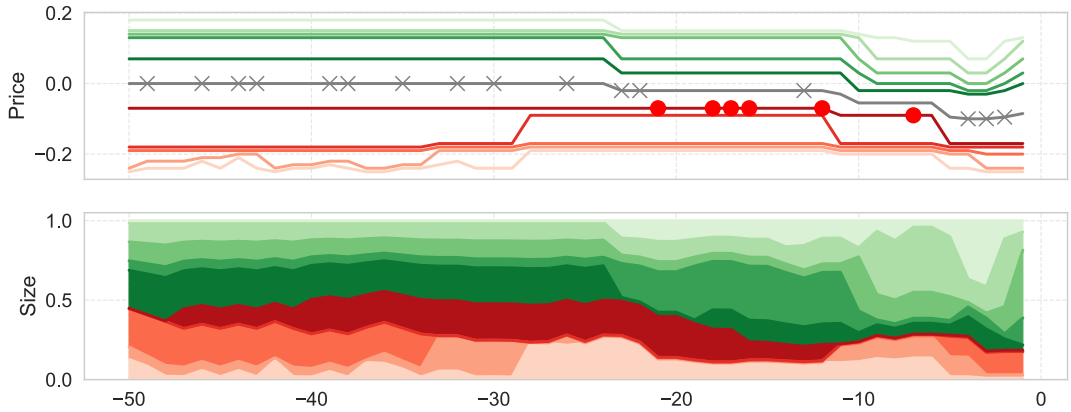
5.3 Visualization

Machine learning models are heavily criticized for their black-boxiness nature and their lack of interpretability. However, as argued by Lipton (2016), the interpretability of a supervised learning model is far from being a well-defined concept, as it refers to several distinct ideas and tries to address the need for desiderata from the model distinct from prediction accuracy. One characteristic to expect form an *ideal* model is a high level of informativeness, that is, its ability to convey information to a human decision-maker through its outputs or additional procedures. I argue that my attention-based model is preferable to alternative model architectures in this dimension, because its attention layer allows to produce visualizations that convey more intuition about the decisions taken by the model.

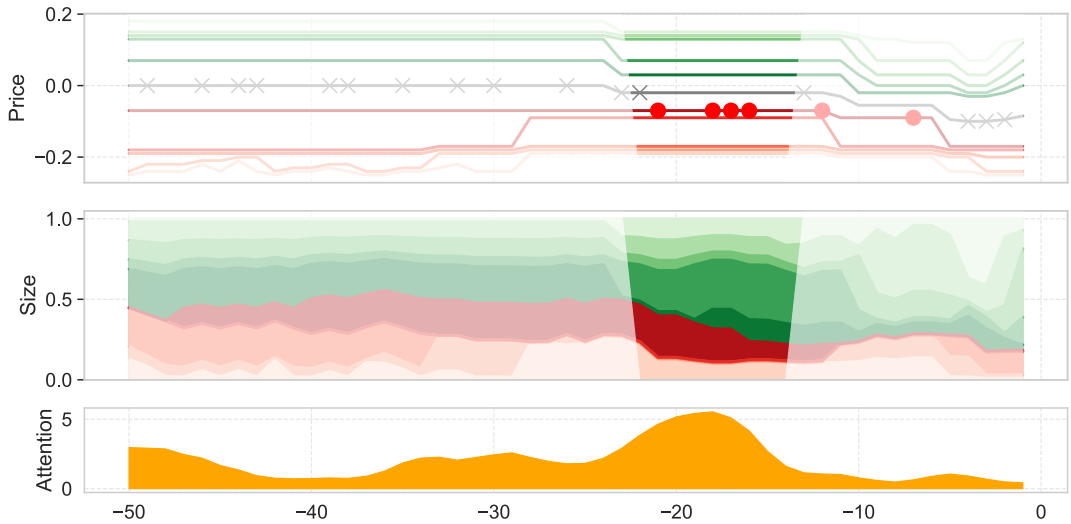
The computer science literature provide a large number of studies in which attention has been used to improve visualization, leading to useful insights of the inner working of neural network models (Bahdanau et al., 2014, Choi et al., 2016). A widely cited work by Xu et al. (2015), for instance, devise a neural model with visual attention and train it to generate textual captions from images. The attention layer helps the model to achieve higher accuracy relative to previous architectures and, moreover, it allows to visually inspect on which part of the image the model focuses to generate each word of the caption. Interestingly, the authors analyze a sample of model's mistakes, and use attention to gain intuition into what the model erroneously saw.

To showcase how the attention layer can be helpful in producing insightful and informative visualizations in the context of this work, consider Figure 2, representing the time-series of input features for a specific example in the test set of the Tesla stock, for which the predicted return is highly positive. The top panel displays the evolution of the mid-price (in gray) and five price levels of outstanding limit orders on the ask side (green) and the bid side (red), in percentage distance from the mid-price. Green (red) dots mark buyer (seller) initiated executions, while gray crosses represent order deletions. The second panel displays the evolution of the ask (in green) and bid (in red) sizes, presented as a percentage of the total liquidity available in the first five levels of the LOB. These two panels surely contain a significant amount of information but, unfortunately, a human decision maker cannot easily understand what is the rationale behind the model's decision to produce a highly positive forecast for future price movements.

The situation can be improved using the attention distribution $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T$ associated to this



(a) Visualization without Attention



(b) Visualization with Attention

Figure 2: Visualization The figure provides a graphical representation of the time-series of features for a specific event in the test set of the Tesla stock, for which the predicted return is highly positive. The first plot of both panels (a) and (b) display the evolution of the mid-price (in gray) and five levels on the ask side (green) and the bid side (red), in percentage distance from the mid-price. Green (red) dots mark buyer (seller) initiated executions, while gray crosses represent order deletions. The second plot of both panels (a) and (b) display the evolution of the relative ask (in green) and bid (in red) sizes, presented as a percentage of the total liquidity in shares available in the first five levels of the LOB. Panel (b) includes an additional third plot, displaying the temporal attention distribution (in percentage) employed by the model to make the prediction, that is, the attention vector $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T$ associated to this particular event. Using this distribution, the time steps in the other two plots corresponding to a level of attention below 2% are shaded, while the region where the model focuses the most are highlighted.

example, which is displayed in the bottom panel. I highlighted the time steps on which the model has decided to focus, shading away those for which the level of attention is below the 2% level. This adjustment makes the plot more informative for a human viewer, allowing her to focus on the most relevant part of the image. In this case, the positive forecast of the model seems to depend on a series of four large sell executions that consume a significant portion of liquidity on the bid side of the book.

6 Liquidity Provision to Institutional Block Orders

The rise of electronic trading platforms resulted into an increased presence of high-frequency traders (HFTs) in equity markets. Members of this new category of traders are described by the Securities and Exchange Commission (SEC) as "professional traders" using "extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders". These new market participants and their impact on market functioning have been investigated by a number of academic studies, the vast majority of which finds that HFTs activity results into reduced bid-ask spreads and improved price efficiency (Hendershott et al., 2011, Chaboud et al., 2014, Boehmer et al., 2014, Brogaard et al., 2015).

Observers in the financial press and practitioners in the industry, however, claim that algorithmic traders may harm market liquidity by front-running block orders of institutional investors[10]. Block orders are large trades which are split over time into multiple executions (child orders) in an attempt to minimize their market impact. One of the main concerns is that algorithms may identify such orders recognizing patterns in the initial part of their execution and, then, front-run the remainder of the block – a strategy usually referred to as *back-running*(Van Kervel and Menkveld, 2019). This issue is of first importance to institutional investors, since their effective trading costs do not depend only on bid-ask spreads but rather on *implementation shortfall*, that is, the cumulative price impact of their block orders.

In principle, the neural network model described in this paper could learn from order book data to recognize blocks before they are fully executed and autonomously engage in back-running strategies, imposing externalities on institutions. Assuming my model – or a similar one – is used by real-world market makers and HFTs, such an automatic tendency to front-run may result in a coordinated action during block trades and induce a systematic reduction of liquidity exactly when institutions need it the most. This dynamics could increase trading costs for institutional investors and, moreover, it may pose significant threats to market stability and

[10] See, for instance, *Wealth Fund Cautions against Costs Exacted by High-Speed Trading*, The New York Times, October 2013; *High-Frequency Trading is Basically Evil* Berkshire Munger, May 2013; and *Institutional Investors Air HFT Concerns* Financial Times, September 2011.

Aggregate Statistics		Event-level Statistics	
Number of events	716	Average block duration (minutes)	39.91
Number of institutions	51	Average number of trades	22.89
Number of stocks	78	Average dollar volume (million $)	6.61
Number of days	234	Average volume ratio wrt CRSP	2.36%

Table 4: Block-Trade Events Summary Statistics. The table reports summary statistics on the identified block-trade events, including aggregate information and event-level averages by minute. The events are defined as collections of executions by an institutional investor such that: (i) the collection consists of at least 5 distinct trades; (ii) the trades are about a single stock j; (iii) the trades are on the same side (buy or sell); (iv) the trades are executed on a single trading day t; (v) the trades are executed during a period of at least 5 minutes and at most 1 hour; (vi) the aggregate volume generated by the trades is at least 1% of the trading volume for stock j reported in CRSP on day t.

lead to excess volatility and even to flash crashes. Understanding how these kind of models behave in the presence of block orders is therefore a relevant issue for both institutional investors and regulatory agencies.

To shed light on this question, I test the behavior of my model during real block orders by institutions. The empirical strategy is based on the ex-post identification of block orders in the AbelNoser/Ancerno dataset, containing transaction-level data from mutual funds and other large institutional investors. In particular, I use data from the year 2014 and impose the conditions outlined in Section 6.1, resulting into a large number of distinct block-order events. Once these large transactions are identified, for each event I train my model in the preceding days and then produce out-of-sample predictions during the actual block order execution and during placebo periods. To conclude the analysis, I run a panel regression of the out-of-sample predictions of the model onto a categorical variable indicating the presence and the side of the block orders. Regression results provide evidence both on the ability of the model to detect block orders from the orderflow and, most importantly, on its relative tendency to engage in front-running or liquidity provision strategies.

6.1 Block Trades Identification

The AbelNoser/Ancerno database contains execution-level data for a large number of institutional investors in the US equity market. Among other information, each row in the dataset reports the institution, the traded stock, the share volume, and the execution price and timing. This structure makes it possible to identify block orders, by searching for multiple repeated executions by a single institution for the same stock. To give the model enough time to rec-

ognize the block, I impose that the sequence of executions is spread over at least 5 minutes[11]. However, I also impose a limit of 1 hour on the duration of each block, to guarantee enough precision of the categorical variable indicating the presence of a block, used in the regression analysis of Section 6.3. Formally, I define a block order $B(i, j, t)$ by institution i on stock j during day t as a collection of trades satisfying the following conditions:

(i) The collection consists of at least 5 distinct trades;

(ii) The trades are about a single stock j;

(iii) The trades are on the same side (buy or sell);

(iv) The trades are executed on a single trading day t;

(v) The trades are executed during a period of at least 5 minutes and at most 1 hour;

(vi) The aggregate volume generated by the trades is at least 1% of the trading volume for stock j reported in CRSP on day t.

The universe of stocks potentially subject to a block order is restricted to the 100 stocks with the highest trading volume in the Nasdaq Stock Exchange as of August 2019. This ensures that the corresponding LOB data is available in the LOBster dataset.

The procedure results in the identification of 716 block trades by 51 different institutions on 78 distinct stocks[12]. Summary statistics, reported in Table 4, show that the identified block-trade events are sizeable in dollar terms and as a fraction of the stock daily volume. Moreover the blocks are split significantly and carried over long periods of time, as they comprise on average more than 20 child executions and last for about 40 minutes on average. Finally, the fact that they are spread overs 234 trading days ensures that they do not arise from specific calendar time events but, rather, these kind of block trades are a systematic feature of US equity market.

6.2 Training Procedure and Predictions

For each block-trade event $B(i, j, t)$ on stock j during day t, I collect data describing the LOB events of that stock during days $t-3$, $t-2$, $t-1$ and t (in business-day units). I use these data to re-train the attention-based model from scratch for each event on the prediction task described in 3.4, that is, to forecast the target variable described in Section 3.3, representing

[11] Even though this may seem a short time frame, Table 1 shows that a window of 5 minutes corresponds on average to more than 1000 LOB events.

[12] The list of stocks subject to a least one block order is the following: AAL, ADBE, ADI, ADSK, ALGN, ALXN, AMAT, AMGN, ATVI, BIIB, BMRN, CDNS, CELG, CERN, CHTR, CMCSA, COST, CSCO, CSX, CTAS, CTSH, DLTR, EA, EBAY, EXPE, FISV, FOX, FOXA, GILD, GOOG, GOOGL, HAS, HSIC, IDXX, ILMN, INCY, INTC, INTU, ISRG, JBHT, KLAC, LRCX, LULU, MAR, MCHP, MDLZ, MELI, MNST, MSFT, MXIM, MYL, NFLX, NTAP, NVDA, ORLY, PAYX, PCAR, PEP, QCOM, REGN, ROST, SBUX, SIRI, SNPS, SWKS, SYMC, TMUS, TSLA, TTWO, TXN, UAL, ULTA, VRSK, VRTX, WDAY, WDC, XEL, XLNX

the expected return of a market maker providing liquidity to the market. In particular, the three business days preceding each block trade are used as training set ($t-3$ and $t-2$) and validation set ($t-1$), using the same procedure as that described in Section 3.5.

Once the model is trained, I use it to produce predictions for each event of the block-trade day. These predictions are then averaged at the minute-frequency and standardized using the mean and standard deviation of the distribution of predictions for each stock.

6.3 Regression Results

Once the event-minute level predictions described in the previous Section are linked to the panel of block-trade events, I construct the categorical variable "Block Order Side", defined at the event-minute level and taking values in $\{-1, 0, +1\}$. In particular, "Block Order Side(B, m)" takes the value of -1 (respectively $+1$) when the corresponding block order B is executing during minute m and it represents a *sell* (respectively *buy*) order by the institutional investor. In those minutes in which no block order was identified by the procedure described in Section 6.1, instead, the variable is set to 0. Notice that by construction the "Block Order Side" variable is equal to zero in the majority of the data points, since block orders are required to last at most for one trading hour. Indeed, in the final panel data, it takes non-trivial values in about 8% of the observations.

To test whether my attention-based model has a tendency to front-run or back-run institutional investors' block orders, I run a regression of the model's predictions on the "Block Order Side". Formally, we estimate the baseline regression model

$$\text{Model Prediction}(B, m) = \alpha + \beta_0 \,\text{Block Order Side}(B, m) + \varepsilon(B, m), \tag{20}$$

and more stringent specifications with stock and day fixed effects, to control for stock-specific and date-specific events that could bias the model's prediction in a specific direction. The last specification includes day×stock fixed effects, which allows to compare the predictions of the model within each stock-day, focusing on the differences between different minutes conditional on the presence of a block order.

Results reported in Table 5 show that the coefficient on β_0 highly significantly different from zero across all specifications, proving that the model's predictions are strongly correlated with the presence of executing block orders. This suggests that the model is able to identify block-trades solely from anonymous orderflow information provided by LOB dynamics. This is a striking result, given that the model has no direct information on institutional investors trades and has not been trained specifically to detect them. Moreover, the fact that the coefficient is negative indicates that the model is significantly more likely to forecast a profit by taking the opposite side of the block order. In other words, the attention-based model suggests the

	(1)	(2)	(3)	(4)
Dependent Variable	Model Prediction	Model Prediction	Model Prediction	Model Prediction
Block Order Side	-0.08***	-0.08***	-0.06***	-0.05**
	(-3.39)	(-3.30)	(-2.74)	(-2.57)
Constant	-0.01			
	(-0.64)			
Observations	277,112	277,112	277,112	277,112
R-squared	0.01	0.01	0.13	0.14
Stock Fixed Effects		Yes		Yes
Day Fixed Effects			Yes	Yes
SEs Clustered By	Stock-Day	Stock-Day	Stock-Day	Stock-Day

Table 5: Model's Predictions during Block Orders. The table reports results from panel regressions of minute-level average predictions of the attention-based model's onto the categorical variable Block Order Side(B, m), taking values in $\{-1, 0, +1\}$ and indicating the presence of a sell block order (-1), of a buy block order $(+1)$ or of no block order during that minute. Block orders are identified imposing the conditions described in Section 6.1 and predictions are generated by the attention-based model trained from scratch and validated on the three business days preceding each block order event. t-statistics are reported in parentheses, based on two-ways standard errors clustered by stock and by day to account for potential serial correlation in residuals. Asterisks denote significance levels (***= 1%, **= 5%, *= 10%).

market maker to buy the stock during liquidations by an institutional investor and, on the contrary, to sell it (or simply not to buy it) when the institution is building up a position on the stock.

All in all these results suggest that the attention-based neural network model studied in this paper, if employed by real world market makers, could potentially improve available liquidity for institutional investors and reduce their trading costs. It is worth remarking that this positive externality for institutions does not imply that market makers are worse off. On the contrary, assuming that the model produces a valuable forecast of their expected profits, market makers should benefit by using its predictions as a signal to adjust their limit orders. This result is consistent with the view that market makers can profit from block orders by engaging in liquidity provision strategies and being compensated for immediacy.

Is liquidity provision the optimal strategy available to market makers? This is an open question, whose answer crucially depends on the information set they have at their disposal. A recent work by Barbon et al. (2019) provides empirical evidence of orderflow leakage by brokers during fire sales of their institutional clients. The authors also show that the market players

who receive the information engage in predatory strategies and earn significant profits. A potential interpretation of the results from this section is that the attention-based model is simply not able to perfectly re-construct the first-hand information available to brokers and, consequently, it cannot forecast the total size and duration of block orders.

7 Concluding Remarks

In this paper I argue that machine learning models can be used to make inference on economic questions, providing a showcase example based on an attention-based neural network. In the first step of the analysis I use LOB data to show that my model is more effective than other architectures in providing accurate forecasts of future transaction prices. I then leverage on this result to justify the use of the attention distributions generated by the model as a proxy for the informational content of different LOB events with respect to price discovery. My results provide empirical evidence of the importance of market order executions for price discovery, while showing that limit orders dynamics play a limited role.

The idea of using attention-based neural network models as inferential device to explore economic issues is novel and could be further explored by future research. For instance, the same method can be applied to different settings to answer different economic questions. Moreover, the attention mechanism could be applied to the cross-sectional rather then the time-series dimension, providing a non-linear way of assessing the relative importance of different factors.

In the second part of the paper, I study how the model behave when faced with the execution of block orders from institutional investors. Results from this exercise suggest that the model favors liquidity provision rather than front-running strategies. However, to better understand the impact that algorithms of this kind can have on market liquidity, further research is needed to analyze the behavior of alternative model architectures and alternative training strategies. For example, one could think of adopting a reinforcement-learning approach and, further, to train the model to directly identify block orders. This, in principle, could be done by constructing a training set with examples of ex-post identified block-order events.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barbon, A., Di Maggio, M., Franzoni, F., and Landier, A. (2019). Brokers and order flow leakage: Evidence from fire sales. *The Journal of Finance*, Forthcoming.

Black, F. (1986). Noise. *The journal of finance*, 41(3):528–543.

Boehmer, E., Fong, K., and Wu, J. (2014). International evidence on algorithmic trading, singapore management university. Technical report, unpublished working paper.

Brogaard, J., Hagströmer, B., Nordén, L., and Riordan, R. (2015). Trading fast and slow: Colocation and liquidity. *The Review of Financial Studies*, 28(12):3407–3443.

Brogaard, J., Hendershott, T., and Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306.

Brogaard, J., Hendershott, T., and Riordan, R. (2016). Price discovery without trading: Evidence from limit orders. *The Journal of Finance*.

Chaboud, A. P., Chiquoine, B., Hjalmarsson, E., and Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, 69(5):2045–2084.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., and Shelhamer, E. (2014). cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512.

Copeland, T. E. and Galai, D. (1983). Information effects on the bid-ask spread. *the Journal of Finance*, 38(5):1457–1469.

Easley, D., De Prado, M. L., and O'Hara, M. (2011). The microstructure of the flash crash: Flow toxicity, liquidity crashes and the probability of informed trading. *Journal of Portfolio Management*, 37(2):118–128.

Farboodi, M. and Veldkamp, L. (2018). *Long Run Growth of Financial Data Technology*. Centre for Economic Policy Research.

Foucault, T., Kadan, O., and Kandel, E. (2013). Liquidity cycles and make/take fees in electronic markets. *The Journal of Finance*, 68(1):299–341.

Girija, S. S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of financial economics*, 14(1):71–100.

Hasbrouck, J. (1991a). Measuring the information content of stock trades. *The Journal of Finance*, 46(1):179–207.

Hasbrouck, J. (1991b). The summary informativeness of stock trades: An econometric analysis. *The Review of Financial Studies*, 4(3):571–595.

Hendershott, T., Jones, C. M., and Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1):1–33.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Kyle, A. S. (1989). Informed speculation with imperfect competition. *The Review of Economic Studies*, 56(3):317–355.

le Calvez, A. and Cliff, D. (2018). Deep learning can replicate adaptive traders in a limit-order-book financial market. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1876–1883. IEEE.

Lee, S. S. and Mykland, P. A. (2007). Jumps in financial markets: A new nonparametric test and jump dynamics. *The Review of Financial Studies*, 21(6):2535–2563.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Mäkinen, M., Kanniainen, J., Gabbouj, M., and Iosifidis, A. (2018). Forecasting of jump arrivals in stock prices: New attention-based network architecture using limit order book data. *arXiv preprint arXiv:1810.10845*.

Menkveld, A. J. (2018). High-frequency trading as viewed through an electron microscope. *Financial Analysts Journal*, 74(2):24–31.

Nevmyvaka, Y., Feng, Y., and Kearns, M. (2006). Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning*, pages 673–680. ACM.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., and Iosifidis, A. (2019). Deep adaptive input normalization for price forecasting using limit order book data. *arXiv preprint arXiv:1902.07892*.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Prechelt, L. (1998). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., and Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*.

Riemer, M., Vempaty, A., Calmon, F., Heath, F., Hull, R., and Khabiri, E. (2016). Correcting forecasts with multifactor neural attention. In *International Conference on Machine Learning*, pages 3010–3019.

Shleifer, A. and Vishny, R. W. (1997). The limits of arbitrage. *The Journal of finance*, 52(1):35–55.

Sirignano, J. and Cont, R. (2018). Universal features of price formation in financial markets: perspectives from deep learning. *Available at SSRN 3141294*.

Sirignano, J. A. (2019). Deep learning for limit order books. *Quantitative Finance*, 19(4):549–570.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tran, D. T., Iosifidis, A., Kanniainen, J., and Gabbouj, M. (2018). Temporal attention-augmented bilinear network for financial time-series data analysis. *IEEE transactions on neural networks and learning systems*.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., and Iosifidis, A. (2018). Using deep learning for price prediction by exploiting stationary limit order book features. *arXiv preprint arXiv:1810.09965*.

Van Kervel, V. and Menkveld, A. J. (2019). High-frequency trading around large institutional orders. *The Journal of Finance*, 74(3):1091–1137.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Zhang, Z., Zohren, S., and Roberts, S. (2019). Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2:207–212.

8 Appendix

	AAL, N=3,486,402					AAPL, N=8,444,983					ADBE, N=2,047,355					CMCSA, N=4,696,675					EBAY, N=4,004,569				
	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%
AS(1)	0.07	0.05	0.00	0.06	0.25	0.05	0.05	0.00	0.04	0.20	0.09	0.07	0.00	0.08	0.31	0.09	0.06	0.00	0.08	0.27	0.10	0.07	0.01	0.08	0.33
AS(2)	0.10	0.05	0.02	0.09	0.30	0.09	0.05	0.01	0.08	0.25	0.10	0.07	0.00	0.09	0.32	0.11	0.04	0.05	0.10	0.25	0.11	0.05	0.03	0.10	0.31
AS(3)	0.11	0.06	0.02	0.09	0.38	0.11	0.05	0.03	0.11	0.29	0.10	0.06	0.00	0.09	0.32	0.11	0.04	0.05	0.10	0.22	0.10	0.05	0.03	0.09	0.26
AS(4)	0.11	0.06	0.03	0.10	0.35	0.13	0.05	0.04	0.12	0.32	0.11	0.06	0.00	0.09	0.32	0.10	0.04	0.05	0.10	0.21	0.10	0.05	0.04	0.10	0.27
AS(5)	0.11	0.06	0.03	0.10	0.32	0.13	0.06	0.04	0.12	0.34	0.11	0.07	0.00	0.09	0.34	0.10	0.04	0.04	0.09	0.21	0.10	0.05	0.03	0.09	0.27
BS(1)	0.07	0.06	0.00	0.06	0.26	0.05	0.04	0.00	0.04	0.18	0.08	0.07	0.00	0.07	0.29	0.09	0.06	0.00	0.08	0.27	0.10	0.06	0.00	0.08	0.30
BS(2)	0.11	0.06	0.02	0.09	0.33	0.08	0.04	0.01	0.08	0.21	0.09	0.07	0.00	0.08	0.30	0.11	0.04	0.05	0.10	0.24	0.10	0.04	0.03	0.09	0.26
BS(3)	0.11	0.07	0.02	0.10	0.37	0.11	0.04	0.03	0.10	0.24	0.10	0.06	0.00	0.09	0.31	0.11	0.03	0.05	0.10	0.22	0.09	0.04	0.03	0.09	0.22
BS(4)	0.11	0.07	0.03	0.10	0.37	0.12	0.04	0.03	0.12	0.26	0.11	0.06	0.00	0.09	0.32	0.10	0.03	0.05	0.10	0.18	0.10	0.04	0.03	0.09	0.23
BS(5)	0.11	0.07	0.03	0.10	0.38	0.13	0.05	0.03	0.12	0.27	0.11	0.07	0.00	0.10	0.33	0.09	0.03	0.04	0.09	0.17	0.10	0.04	0.03	0.09	0.24
AP(1)	0.02	0.01	0.01	0.01	0.04	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.02	0.06	0.01	0.00	0.01	0.01	0.03	0.01	0.00	0.01	0.01	0.03
AP(2)	0.05	0.01	0.04	0.04	0.07	0.01	0.00	0.01	0.01	0.02	0.03	0.01	0.01	0.02	0.08	0.04	0.00	0.04	0.04	0.05	0.04	0.00	0.04	0.04	0.05
AP(3)	0.08	0.01	0.07	0.07	0.10	0.02	0.00	0.01	0.01	0.02	0.03	0.02	0.01	0.03	0.10	0.07	0.00	0.06	0.06	0.08	0.07	0.00	0.06	0.07	0.08
AP(4)	0.10	0.01	0.10	0.10	0.13	0.02	0.00	0.02	0.02	0.03	0.04	0.02	0.02	0.03	0.11	0.09	0.00	0.09	0.09	0.10	0.10	0.00	0.09	0.09	0.11
AP(5)	0.13	0.01	0.12	0.13	0.16	0.03	0.00	0.03	0.03	0.03	0.04	0.02	0.02	0.04	0.12	0.12	0.00	0.11	0.12	0.13	0.12	0.00	0.12	0.12	0.14
BP(1)	0.02	0.01	0.01	0.01	0.04	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.02	0.06	0.01	0.00	0.01	0.01	0.03	0.01	0.00	0.01	0.01	0.03
BP(2)	0.05	0.01	0.04	0.04	0.07	0.01	0.00	0.01	0.01	0.02	0.03	0.01	0.01	0.02	0.08	0.04	0.00	0.04	0.04	0.05	0.04	0.00	0.04	0.04	0.05
BP(3)	0.08	0.01	0.07	0.07	0.10	0.02	0.00	0.01	0.01	0.02	0.03	0.02	0.01	0.03	0.10	0.07	0.00	0.06	0.06	0.08	0.07	0.00	0.06	0.07	0.08
BP(4)	0.10	0.01	0.10	0.10	0.13	0.02	0.00	0.02	0.02	0.03	0.04	0.02	0.02	0.03	0.11	0.09	0.00	0.09	0.09	0.10	0.10	0.00	0.09	0.09	0.11
BP(5)	0.13	0.01	0.12	0.13	0.16	0.03	0.00	0.03	0.03	0.03	0.04	0.02	0.02	0.04	0.12	0.12	0.00	0.11	0.12	0.13	0.12	0.00	0.12	0.12	0.14
Execution	0.02	0.13	0.00	0.00	1.00	0.03	0.16	0.00	0.00	1.00	0.04	0.21	0.00	0.00	1.00	0.01	0.11	0.00	0.00	1.00	0.01	0.11	0.00	0.00	1.00
Deletion	0.50	0.50	0.00	0.00	1.00	0.48	0.50	0.00	0.00	1.00	0.46	0.50	0.00	0.00	1.00	0.51	0.50	0.00	1.00	1.00	0.51	0.50	0.00	1.00	1.00
Direction	-0.04	1.00	-1.00	-1.00	1.00	-0.02	1.00	-1.00	-1.00	1.00	-0.01	1.00	-1.00	-1.00	1.00	-0.00	1.00	-1.00	-1.00	1.00	0.04	1.00	-1.00	1.00	1.00
Elapsed Time	0.07	0.26	0.00	0.00	1.14	0.03	0.09	0.00	0.00	0.41	0.11	0.43	0.00	0.00	2.02	0.05	0.20	0.00	0.00	0.92	0.06	0.24	0.00	0.00	1.09
Price Change	-0.00	0.27	-1.41	0.00	1.41	-0.00	0.09	-0.29	0.00	0.29	-0.00	0.27	-0.96	0.00	0.95	-0.00	0.13	0.00	0.00	0.00	-0.00	0.16	0.00	0.00	0.00

	FB, N=5,090,807					MSFT, N=11,800,393					NVDA, N=3,031,962					PEP, N=2,415,737					TSLA, N=2,097,592				
	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%	Mean	Std	1%	50%	99%
AS(1)	0.06	0.06	0.00	0.04	0.27	0.06	0.05	0.00	0.06	0.22	0.08	0.08	0.00	0.06	0.39	0.07	0.05	0.00	0.06	0.25	0.10	0.11	0.00	0.08	0.52
AS(2)	0.09	0.07	0.00	0.08	0.31	0.09	0.04	0.03	0.09	0.24	0.09	0.08	0.00	0.07	0.40	0.10	0.05	0.02	0.09	0.25	0.10	0.11	0.00	0.08	0.56
AS(3)	0.11	0.07	0.01	0.10	0.34	0.10	0.05	0.03	0.10	0.26	0.10	0.08	0.00	0.08	0.41	0.10	0.05	0.02	0.10	0.26	0.10	0.11	0.00	0.08	0.57
AS(4)	0.12	0.07	0.02	0.11	0.37	0.11	0.05	0.04	0.11	0.29	0.11	0.08	0.00	0.09	0.43	0.12	0.05	0.02	0.11	0.28	0.10	0.11	0.00	0.07	0.57
AS(5)	0.13	0.07	0.02	0.12	0.39	0.12	0.06	0.04	0.11	0.32	0.12	0.08	0.01	0.11	0.44	0.12	0.06	0.03	0.11	0.30	0.10	0.11	0.00	0.07	0.58
BS(1)	0.06	0.06	0.00	0.04	0.26	0.06	0.05	0.00	0.06	0.21	0.08	0.08	0.00	0.06	0.37	0.07	0.05	0.00	0.06	0.25	0.10	0.11	0.00	0.07	0.59
BS(2)	0.09	0.06	0.00	0.08	0.29	0.10	0.04	0.03	0.09	0.23	0.09	0.08	0.00	0.07	0.40	0.09	0.05	0.02	0.09	0.24	0.10	0.12	0.00	0.07	0.61
BS(3)	0.11	0.06	0.01	0.10	0.32	0.11	0.04	0.04	0.10	0.24	0.10	0.08	0.00	0.08	0.41	0.10	0.05	0.02	0.09	0.24	0.10	0.12	0.00	0.07	0.63
BS(4)	0.12	0.06	0.01	0.11	0.34	0.12	0.05	0.04	0.11	0.26	0.11	0.08	0.00	0.10	0.43	0.11	0.05	0.02	0.10	0.27	0.10	0.12	0.00	0.07	0.64
BS(5)	0.13	0.06	0.02	0.12	0.36	0.12	0.05	0.04	0.12	0.28	0.12	0.08	0.00	0.11	0.44	0.12	0.06	0.03	0.11	0.30	0.10	0.12	0.00	0.07	0.66
AP(1)	0.01	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.04	0.01	0.00	0.00	0.00	0.03	0.02	0.01	0.00	0.02	0.07
AP(2)	0.02	0.01	0.01	0.01	0.04	0.01	0.00	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.06	0.02	0.01	0.01	0.01	0.04	0.04	0.02	0.01	0.03	0.11
AP(3)	0.02	0.01	0.01	0.02	0.04	0.02	0.00	0.02	0.02	0.03	0.03	0.01	0.02	0.03	0.07	0.02	0.01	0.02	0.02	0.05	0.05	0.03	0.01	0.05	0.14
AP(4)	0.03	0.01	0.02	0.03	0.05	0.03	0.00	0.03	0.03	0.04	0.04	0.01	0.02	0.04	0.08	0.03	0.01	0.03	0.03	0.06	0.06	0.03	0.02	0.06	0.16
AP(5)	0.03	0.01	0.03	0.03	0.06	0.04	0.00	0.04	0.04	0.05	0.05	0.01	0.03	0.04	0.09	0.04	0.01	0.04	0.04	0.07	0.07	0.03	0.02	0.07	0.19
BP(1)	0.01	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.04	0.01	0.00	0.00	0.00	0.03	0.02	0.01	0.00	0.02	0.07
BP(2)	0.02	0.01	0.01	0.01	0.04	0.01	0.00	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.06	0.02	0.01	0.01	0.01	0.04	0.04	0.02	0.01	0.03	0.11
BP(3)	0.02	0.01	0.01	0.02	0.04	0.02	0.00	0.02	0.02	0.03	0.03	0.01	0.02	0.03	0.07	0.02	0.01	0.02	0.02	0.05	0.05	0.03	0.01	0.04	0.14
BP(4)	0.03	0.01	0.02	0.03	0.05	0.03	0.00	0.03	0.03	0.04	0.04	0.01	0.02	0.04	0.09	0.03	0.01	0.03	0.03	0.07	0.06	0.03	0.02	0.05	0.16
BP(5)	0.03	0.01	0.03	0.03	0.06	0.04	0.00	0.04	0.04	0.05	0.05	0.01	0.03	0.04	0.09	0.04	0.01	0.04	0.04	0.08	0.07	0.03	0.02	0.07	0.18
Execution	0.05	0.21	0.00	0.00	1.00	0.02	0.13	0.00	0.00	1.00	0.06	0.24	0.00	0.00	1.00	0.03	0.17	0.00	0.00	1.00	0.14	0.34	0.00	0.00	1.00
Deletion	0.47	0.50	0.00	0.00	1.00	0.50	0.50	0.00	1.00	1.00	0.46	0.50	0.00	0.00	1.00	0.49	0.50	0.00	0.00	1.00	0.39	0.49	0.00	0.00	1.00
Direction	0.01	1.00	-1.00	1.00	1.00	0.01	1.00	-1.00	1.00	1.00	0.02	1.00	-1.00	1.00	1.00	0.01	1.00	-1.00	1.00	1.00	0.06	1.00	-1.00	1.00	1.00
Elapsed Time	0.05	0.17	0.00	0.00	0.77	0.02	0.08	0.00	0.00	0.36	0.08	0.26	0.00	0.00	1.23	0.10	0.32	0.00	0.00	1.48	0.11	0.37	0.00	0.00	1.72
Price Change	0.00	0.17	-0.60	0.00	0.60	0.00	0.09	-0.45	0.00	0.45	-0.00	0.28	-0.96	0.00	0.96	-0.00	0.15	-0.43	0.00	0.43	-0.00	0.64	-2.17	0.00	2.19

Table 6: Features Summary Statistics. The table presents stock-level summary statistics for the feature variables used as input to the machine learning models. The 'Ask Price', 'Bid Price', 'Ask Size' and 'Bid Size' variables are generated from the raw dataset by following the standardization procedure outlined in Section 3.2. The 'Execution' and 'Deletion' features are dummy variables for LOB messages indicating executions and order cancellations, respectively (the complementary set of observations constitutes submission of new limit orders). 'Direction' is a binary variable taking values in $\{-1, 1\}$ and providing the directionally of the LOB message, i.e positive (negative) values indicate a change in the bid (ask) side of the order book. In particular, for executions, positive (negative) values of 'Direction' indicate seller-initiated (buyer-initiated) trades. 'Elapsed Time' measures the time from the previous message, in seconds. 'Price Change' denotes the percentage mid-price change from the previous message.