A high frequency analysis of the information content of trading volume

KHALADDIN RZAYEV
University of Edinburgh, United Kingdom

GBENGA IBIKUNLE[*]
University of Edinburgh, United Kingdom
European Capital Markets Cooperative Research Centre, Pescara, Italy

Abstract We propose a state space modelling approach for decomposing high frequency trading volume into liquidity-driven and information-driven components. Using a set of high frequency S&P 500 stocks data, we show that informed trading increases pricing efficiency by reducing volatility, illiquidity and toxicity/adverse selection during periods of non-aggressive trading. We observe that our estimated informed trading component of volume is a statistically significant predictor for one-second stock returns; however, it is not a significant predictor for one-minute stock returns. We show that this disparity is explained by high frequency trading activity, which eliminates pricing inefficiencies at high frequencies.

JEL Classification: G12; G14; G15

Keywords: trading volume; expected component; unexpected component; market quality; time series models; state space modelling.

[*]Corresponding author. Contact information: University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, United Kingdom; e-mail: Gbenga.Ibikunle@ed.ac.uk; phone: +441316515186.

1. Introduction

Market participants’ trades are driven by either information or the search for liquidity (see Admati and Pfleiderer, 1988). Liquidity traders do not trade on the basis of any specific information; their trading strategies are therefore not directly related to future payoffs. The trading strategies of informed traders, on the other hand, are based on private information and are directly related to future payoffs. The activities of these two fundamental types of traders have been extensively analysed in seminal papers in the larger financial markets literature, and more so in market microstructure papers. For example, Kyle (1985) predicts that the volatility of asset prices partially reflects inside information (informed trading) and is independent of liquidity-driven trading effects, while Glosten and Milgrom (1985) predict that the breadth of the bid-ask spread is primarily driven by informed trading, which incorporates adverse selection costs into the spread. In both the Kyle (1985) and the Glosten and Milgrom (1985) models, it is assumed that traders execute their trading strategies by using marker orders; thus, all traders trade aggressively in both models. More recently however, Collin-Dufresne and Fos (2016) extend Kyle's (1985) model to show that the relationship between stock price volatility and informed trading depends on the aggressiveness of traders. Furthermore, in contrast to Glosten and Milgrom's (1985) model, Collin-Dufresne and Fos (2016) predict that informed trading may be negatively correlated with adverse selection if informed traders execute their strategies using limit orders. Using a comprehensive sample of trades from Schedule 13D filings by activist investors, Collin-Dufresne and Fos (2015) show that informed traders with long-lived information tend to use limit orders, which leads to a negative correlation between adverse selection and informed trading (see also Kaniel and Liu, 2006).

This paper builds on the above predictions and findings by developing a general state space-based methodology for decomposing trading volume into unobservable liquidity-driven and information-driven components. According to Hendershott and Menkveld (2014), state space modelling is a natural tool to model an observed variable as the sum of two unobserved variables. While the application of state space modelling for decomposing price, owing to its efficiency, is very common in the finance literature (see as examples Brogaard et al., 2014; Hendershott and Menkveld, 2014; Menkveld et al., 2007), the approach has thus far not been directly applied to trading volume.[1] This is surprising given the preponderance of the literature on the strength of the relationship between price and trading volume (see as examples, Clark, 1973; Cornell, 1981; Epps and Epps, 1976; Harris, 1986, 1987; Karpoff, 1987). The heavily evidenced relationship in the literature is linked to the joint dependence of price and volume on an underlying or set of underlying variable(s); this is the 'mixture of distribution hypothesis' (MDH) (see Clark, 1973; Harris, 1986). Harris (1986) argues that the underlying variable is the rate of flow of information. Hence, as new information arrives, traders act on them by revising their positions and consequently increase trading volume. Harris (1987), using data from the NYSE, provides an empirical basis for the MDH. This implies that the theoretical basis for the application of state space modelling to price (i.e. that price reflects both information and non-information components), holds for volume.[2] However, it is important to note that while the information component of price is its permanent component, the information component of volume is transitory. This is simply because, although new information implies new permanent level of prices, it will only affect trading volume temporarily since after prices reflect this information, informed traders will no longer hold an informational advantage and will therefore cease their trading based on the exploited information (see also Fama, 1970; Chordia et al., 2002; Suominen, 2001).

As discussed by Hendershott and Menkveld (2014), the state space approach holds significant economic value over other methods that could be appropriated for variable decomposition, such as autoregressive models (see as an example, Hasbrouck, 1991). Firstly, the estimation of the model using maximum likelihood is asymptotically unbiased and efficient.

[1] McCarthy and Najand (1993) apply state space modelling to the analysis of price and volume dependence in currency futures.

[2] A second explanation for the existence of the price-volume relationship is based on the sequential information models proposed by Copeland (1976), Jennings et al. (1981) and Smirlock and Starks (1984). The models suggest that volume improves forecasts of price variability and vice versa.

Secondly, maximum efficiency in dealing with missing values is achieved due to the use of the Kalman filter, which accounts for level changes across periods with missing observations, employed in the maximum likelihood estimation. This is a critical argument in the use of state space modelling in decomposing asset prices and trading volume in a high frequency trading environment such as the one we examine, since standard estimation approaches do not deal with missing observations. For example, estimating a vector autoregression implies truncating the lag structure. Although standard approaches for decomposing trading volume may work well in a low-frequency environment, information in today's markets travel at high speeds, thus leading to those approaches potentially discarding additional information that could be obtained from high frequency data. Thirdly, following estimation, the Kalman smoother, which is basically a backward recursion after a forward recursion with the Kalman filter, facilitates a decomposition of any realised change in the series such that the estimated permanent or transitory component at any interval is estimated using all past, present, and future observations in the series. Thus, the purpose of filtering is to ensure that estimates are updated with the introduction of every additional observation (see also Durbin and Koopman, 2012).

In line with the expectation that asset price (and by extension, volume) is driven by informed trading and can therefore be decomposed into permanent and transitory components (see Brogaard et al., 2014; Menkveld et al., 2007), we demonstrate that (observable) trading volume is a sum of two unobserved series: a nonstationary series (expected component) and a stationary series (unexpected component). We argue that the unobserved expected component of trading volume is mainly driven by liquidity traders, whereas the unobserved unexpected component is primarily driven by informed traders. The expected component in the state space model is a nonstationary series and follows a random walk. Consistent with the literature, liquidity traders trade randomly (i.e. the reference to noise trading in the market microstructure literature), and thus we model the trading volume of liquidity traders as a random walk (see as examples of Admati and Pfleiderer, 1988; Kyle, 1985). In addition, in state space models, changes in the expected component affect the observable variable permanently, while changes in the unexpected component have a transitory impact on the observable variable, in this case, trading volume (see Hendershott and Menkveld, 2014), hence our argument that the expected component is driven by liquidity traders, while the unexpected component is information trades-driven.

In a test of the validity of the proposed state space-based volume decomposition approach, firstly, we use the estimated expected and unexpected components of trading volume to examine the role of liquidity and informed traders on market quality metrics, such as volatility, liquidity, and toxicity. This part of our analysis serves as a joint test of the empirical relevance of our state space model and the impact of different trader types on market quality. The relevance of our state space approach is underscored when our empirical findings are related to the model predictions in the existing relevant theoretical market microstructure literature. Secondly, we examine the predictive power of the estimated information-driven/unexpected component of trading volume on short-horizon returns. This analysis furthers our aim of demonstrating the relevance of the state space approach to decomposing trading volume into informed and liquidity components. It is also a direct test of the efficiency of the price discovery process (see Chordia et al., 2005, 2008). Similar to the order imbalance metrics employed in Chordia et al. (2008), the unexpected component also signals private information, and we expect it to be a predictor of short-horizon returns. Thirdly, we conduct a direct empirical test of the ability of the unexpected component to capture information asymmetry. The test is simple and intuitive and involves examining the behaviour of the unexpected component around earnings announcements. We focus on earnings announcements because of two reasons. The first is that there is overwhelming evidence of information leakage prior to these events (see as an example Christophe et al., 2004); it implies that earnings announcements provides ideal ground for testing the proxy for informed trading. The second reason is that testing the behaviour of informed trading proxies by using earnings announcements is a well-established and widely accepted approach in the market microstructure literature (see as examples Benos and Jochec, 2007; Easley et al., 2008).

Consistent with the aforementioned studies, we expect that the unexpected component for the days preceding earnings announcements should be significantly higher than the unexpected component for the days after earnings announcements.

All the results obtained are generally consistent with our expectations. Based on our state space-estimated information and liquidity-driven components of trading volume, we find that stock price volatility is independent of liquidity trading, but impacted by information-motivated trading (see Glosten and Milgrom, 1985; Kyle, 1985). We also find that information-motivated trading volume improves pricing efficiency by reducing price volatility and market toxicity and improving liquidity; the results are robust to alternative estimation frequencies, and volatility and liquidity proxies. This finding is in line with the theoretical model developed by Collin-Dufresne and Fos (2016), which predicts that the price volatility-informed trading relationship is influenced by two effects. On the one hand, informed trading reveals information, and this decreases uncertainty in financial markets, which reduces price volatility. On the other hand, the aggressive behaviour of informed traders could increase volatility. Thus, the net impact of informed trading on stock price volatility depends on which effect dominates. Thus, our finding in relation to volatility is linked to the period of relative calm in the S&P 500 stocks, which we examine. Furthermore, Menkveld (2013) shows that aggressive trading is not profitable during normal trading periods, i.e. trading periods are considered normal if there is no excessive aggressiveness, such as a flash crash. This implies that informed traders do not tend to use aggressive orders during periods of relative calm in financial markets; thus, their activities could lead to a reduction of volatility in the markets, as predicted by Collin-Dufresne and Fos (2016). The results are also consistent with the findings of Avramov et al. (2006) and Collin-Dufresne and Fos (2015), who find that price volatility and adverse selection are negatively correlated with informed trading. The negative relationships of informed trading with order flow toxicity and illiquidity are linked to informed traders' use of limit orders rather than (aggressive) market orders. In a large part of the market microstructure literature it is generally assumed that informed traders use only market orders, and therefore it is expected that informed traders increase aggressiveness and widen the bid-ask spread, a proxy for illiquidity and, by extension, one of its components, adverse selection (or its high frequency equivalent, market toxicity). However, Kaniel and Liu (2006), modifying Glosten and Milgrom's (1985) model, demonstrate that if there is a high probability that when exploitable information is seen as long-lived, then informed traders tend to submit limit orders. The prediction of Kaniel and Liu's (2006) model is empirically confirmed by Collin-Dufresne and Fos (2015), who find that informed traders with long-lived information tend to use limit orders, which leads to a reduction in adverse selection.[3]

Furthermore, we find that the unexpected component, as estimated using our state space approach, is a significant predictor of one-second stock returns. This implies that although financial markets are efficient in the long-term, there are short-term inefficiencies in markets because investors need time to absorb new information (see Chordia et al., 2008). However, we find that the horizon for short-term stock returns predictability has decreased substantially since the five-minute window reported by Chordia et al. (2008). The predictability of short-horizon returns now only holds on a per second basis, and no longer at the minutes-long threshold reported in earlier studies. We show that high frequency trading is the driver of this sharp reduction in length of short-term return predictability. We also find that the unexpected component captures informed trading activity around earnings announcements.

A few streams of the literature are related to this study. There are those studies delineating traders into liquidity-driven and information-driven traders (see as an example Avramov et al., 2006), and another extensive stream examining the role of the different types of traders on price volatility and liquidity (see as examples Avramov et al., 2006; Daigler and Wiley, 1999; Van Ness et al., 2016). This current paper differs from these studies in at least three respects. Firstly, the approach of decomposing trading volume using state space modelling is fundamentally different to those employed in existing studies. A major advantage

[3] The rational expectation model developed by Wang (1993) also predicts a negative correlation between informed trading and stock price volatility, but via a different mechanism. Furthermore, Admati and Pfleiderer (1988) also argue in favour of a negative relationship between adverse selection and informed trading.

of this approach, as earlier stated, is the asymptotic unbiasedness and efficiency of the estimation approach, i.e. maximum likelihood via Kalman filter (see Brogaard et al., 2014; Hendershott and Menkveld, 2014). Secondly, we examine the role of informed trading activity in the evolution of specific market quality metrics, including for a new market quality metric, market toxicity. Finally, and critically, we present new evidence on the speed of price adjustment in the presence of information-driven order flow in financial markets.

2. Theory and the previous literature

In this paper, we decompose trading volume into liquidity and information-driven components, and thereafter test the empirical relevance of our model and the role of liquidity and informed traders in the price discovery process. Our empirical analysis is based on the predictions of widely accepted theories as proposed in existing studies. Thus, this paper is related to the stream of literature investigating the impact of asymmetric information on asset prices' volatility and liquidity. Kyle (1985) presents one of the first and best-established models deriving equilibrium security prices when traders possess asymmetric information. The model assumes three types of traders in a market: a market maker, a noise trader that trades randomly, and an informed trader, and also provides a framework for determining the price impact of trading volume. The model shows that stock price volatility partially reflects inside information, which is independent of noise trading volatility. Furthermore, the model predicts that informed traders trade more actively when there is a higher level of noise trading volume in the markets, because the higher uninformed trading volume provides a "camouflage" for informed order flow. Glosten and Milgrom (1985) model the bid-ask spread and propose a then new explanation on why it arises in financial markets. The model predicts that adverse selection implies that the market maker makes losses whenever trading with insiders, and hence she is forced to impose different charges on buy and sell volumes in order to compensate for her potential losses. In other words, the model predicts that the bid-ask spread depends on informed trading activity and the independence of liquidity traders. Moreover, the model predicts that the higher the variance of prices, the greater the impact of insiders/informed traders on the bid-ask spread. Consistent with Glosten and Milgrom (1985), Easley and O'Hara (1987) also suggest that stock illiquidity should increase in the presence of informed traders, as information asymmetry increases adverse selection, which widens the spread.

In both Kyle's (1985) and Glosten and Milgrom's (1985) models, the liquidity traders trade randomly. By contrast, Admati and Pfleiderer (1988) argue that this is a strong assumption and it might be more reasonable to assume that at least some liquidity traders can select the timing of their transactions. Consistent with the literature, this model predicts that the information-motivated trades increase as liquidity driven trading volumes rise, and the variance of price changes is independent from the variance of liquidity traders. However, surprisingly, the theoretical framework predicts that adverse selection decreases with the number of informed traders. Admati and Pfleiderer (1988) argue that informed traders in possession of the same set of information will compete, and that this competition reduces adverse selection and increases benefits to liquidity traders.

As already noted, generally, theoretical models examining information asymmetry in the price discovery process assume that informed traders execute their trading strategies by using market orders, i.e. they are aggressive traders (see as examples Glosten and Milgrom, 1985; Kyle, 1985). Popular models such as the probability of informed trading (PIN) model, developed by Easley et al. (1996) and Easley et al. (1997), also make this assumption. In contrast to these models, Kaniel and Liu (2006) argue that the assumption is unnecessarily strong. By extending Glosten and Milgrom's (1985) model, the authors show that informed traders with long lived information strategically tend to use limit orders instead of market orders (see also Sun and Ibikunle, 2016). Collin-Dufresne and Fos (2016) also extend Kyle's (1985) model of insider trading and show that the impact of informed trading on the price discovery process is two-fold and could be explained by two mechanisms. Firstly, informed traders reveal information, which decreases the level of price uncertainty in the market; thus, stock price volatility is negatively correlated with informed trading. Secondly, informed traders could trade aggressively, and this aggressive behaviour increases stock price volatility in financial markets; hence, stock price volatility is positively correlated with informed traders. Therefore, the relationship between market quality characteristics, such as price volatility, and informed traders depends on which effect dominates the other. The majority of market microstructure models predict positive correlations between informed trading and stock price volatility because they assume that informed traders will aim to quickly take advantage of private information by seeking to execute market orders based on such information. However, Menkveld (2013) and Rzayev and Ibikunle (2017) show that aggressive trading is not profitable for informed traders if there is no widespread aggression in the market. This implies that during calmer periods, we would expect to see a negative relationship between informed trading volume and stock price volatility (see also Collin-Dufresne and Fos, 2015; Kaniel and Liu, 2006). The negative informed trading-price volatility relationship is also predicted by rational expectations models (see as examples Hellwig, 1980; Wang, 1993).

While the relationship between informed trading volume and price volatility is nuanced, a positive relationship between aggregate trading volume (i.e. containing informed and uninformed volume) and stock price volatility is widely documented (see as an example the studies summarized in Karpoff, 1987). Generally, the impact of trading volume on stock price volatility is explained by some related theories. We mainly focus on two well-known and widely accepted theories: information theories and dispersion of beliefs theories. Information theories, such as a mixture of distributions models and sequential arrival of information models, suggest that both volatility and volume are determined by information arrivals (see Copeland, 1976, 1977; Epps and Epps, 1976). The dispersion of beliefs theory, modelled by Harris and Raviv (1993) and Shalen (1993), argues that both unusual volume and volatility are associated with the differences in traders' beliefs. To put it simply, the dispersion of beliefs model/theory incorporates the role of different types of traders into the relationship between trading volume and stock price volatility.

In most existing studies, trading activity is measured by total trading volume. However, as already noted, the dispersion of beliefs models argue that this relationship depends on the differences in traders' beliefs, and thus linking volatility to total trading volume conceals some important information (see also Chordia et al., 2002). Therefore, some studies decompose trading volume into its components and then examine the role of different trading components on market quality characteristics, such as stock price volatility and market liquidity (see as examples Avramov et al., 2006; Bessembinder and Seguin, 1993; Daigler and Wiley, 1999). Avramov et al. (2006) partition trades into two components: herding (non-informed) and contrarian (informed) trades. Consistent with the rational expectation models, Avramov et al. (2006) find that herding trades increase stock price volatility, and contrarian trades reduce it. Collin-Dufresne and Fos (2015) directly examine the role of informed traders in the pricing process by using a comprehensive sample of trades from Schedule 13D filings by activist investors, and conclude that when informed traders can select when (they could strategically trade when noise trading is high) and how (they might strategically select to use limit orders) to trade, their trading activity decreases adverse selection in financial markets.

We extend this study to examine the effects of informed trading on market toxicity, and then relate it to Van Ness et al. (2016). Van Ness et al. (2016) investigate the role of high frequency traders (HFTs) in order flow toxicity by employing the Easley et al. (2011, 2012) volume-synchronized probability of informed trading (VPIN) metric as a measure of order flow toxicity. Their study finds a negative correlation between HFT activity and order flow toxicity. It indicates that, as HFT increases, average order flow toxicity decreases. Furthermore, the authors observe a negative correlation between trading volume and order flow toxicity; specifically, as volume increases, average market toxicity decreases.

Finally, our approach for decomposing trading volume into informed and uninformed components is based on state space modelling; therefore, our paper is also related to yet another stream of the market microstructure literature, which employs state space models. Generally, the existing body of literature on market microstructure uses state space modelling only for decomposing price, rather than volume, into two components (see as examples Brogaard et al., 2014; Hendershott and Menkveld, 2014; Menkveld et al., 2007). Menkveld et al. (2007) use the approach to analyse around-the-clock price discovery for cross-listed stocks in the Amsterdam exchange and NYSE. Their study finds that NYSE plays a minor role in the price discovery process for Dutch stocks. Similar to Menkveld et al. (2007), Brogaard et al. (2014) use a state space model in order to analyse the price discovery process in the US market. More precisely, they examine the role of high frequency trading (HFT) in the price discovery process. The study reports a positive role for HFT in the price discovery process. Durbin and Koopman (2012) provide a more detailed discussion on the advantages of state space models.

3. Data and descriptive statistics

3.1 Data

The main dataset employed in this study consists of ultra-high frequency tick-by-tick data for the most active 100 S&P 500 stocks sourced from the Thomson Reuters Tick History (TRTH) database. The dataset includes data for trading days between October 2016 and September 2017. In the data, each message is recorded with a time stamp to the nearest millisecond. The following variables are included in the dataset: Reuters Identification Code (RIC), date, timestamp, price, volume, bid price, ask price, bid volume, and ask volume. We then follow Chordia et al. (2001) and Ibikunle (2015) in applying a standard set of exclusion criteria to the data, with the aim of excluding inexplicable values that may arise due to erroneous data entries. Table 1 presents the summary statistics of trading activities for the final sample of stocks.

INSERT TABLE 1 ABOUT HERE

We apply Lee and Ready's (1991) algorithm to classify trades as buyer- or seller-initiated.[4] Going by the number of transactions and nominal and dollar-denominated trading

[4] Chakrabarty et al. (2015) compare the different trades classification methods and conclude that Lee and Ready's (1991) is the most accurate method.

volume, the sell side appears to be marginally more active than the buy side over the sample period. This view is further underscored by the average trade sizes for both buys and sells. The sellers also appear to be more aggressive, based on the average sizes of their trades.

In order to execute additional out of sample tests of the validity of our state space modelling approach, we also obtain a proprietary NASDAQ-provided transactions dataset for 120 randomly selected NASDAQ and NYSE-listed stocks trading during all the trading days in 2009. The data is complementary to the main dataset we employ, because it disaggregates transactions into those executed on the basis of orders submitted by HFTs and non-HFTs. This is the same dataset described in detail by Brogaard et al. (2014). The dataset contains the following information on each transaction included in the sample: date, time (in milliseconds), transaction size (shares), price, buy-sell indicator, and liquidity nature of the two sides to each trade (HH, HN, NH and NN). HH indicates a trade based on a HFT demanding liquidity and a HFT supplying the required liquidity. HN implies that a HFT demands liquidity and a non-HFT supplies liquidity, while NH is the opposite. NN refers to trades where both counterparties are non-HFTs. We identify the sum of HH, HN and NH as HFT volume. This dataset is only employed in Section 4.5 of this paper. In that section, we present the justification for its use. Table 2 presents the summary statistics of trading activities for the NASDAQ-provided transactions dataset. The table shows that HFTs are counterparties in about 71% of all trades.

INSERT TABLE 2 ABOUT HERE

3.2 Main Variables

One aim of this study is to examine the role of informed and liquidity traders in the evolution of price volatility, liquidity, and market toxicity. This inevitably translates into a joint test of the empirical relevance of the state space model we employ and the impact of the different types of traders on several market quality metrics. Specifically, we build a set of predictive regressions to test the impact of the expected and unexpected components of traded volume on price volatility, liquidity, and market toxicity. Thus, apart from the state space-estimated unobservable variables, our volatility, liquidity, and market toxicity measures are the main variables of interest.

Consistent with the literature, we use absolute price change as a proxy for stock price volatility. For robustness, we also use the standard deviation of stock returns (see as examples Karpoff, 1987; Lamoureux and Lastrapes, 1990) as a proxy for stock price volatility. Absolute price change is defined as the absolute value of the differences between prices at times *t* and t_{-1}, and we use one-second, one-minute and one-hour intervals to compute the absolute price change(s). To compute the standard deviation of stock returns, firstly we employ the midpoint of the bid and ask quotes corresponding to every transaction.[5] For robustness, we also compute the standard deviation of stock returns by computing the returns from the execution price for each transaction rather than the midpoint of the prevailing quotes.

For robustness, we employ three spread measures as proxies for liquidity; the spread metrics are the effective spread, quoted spread, and relative spread. The relative and quoted spread measures are computed using the best bid and ask prices for each interval, *t*, which corresponds to one second, minute, or hour.[6] The relative bid-ask spread is obtained by dividing the difference between ask and bid prices by the midpoint of both prices, while the quoted spread is simply the difference between the ask and bid prices. The effective spread is twice the absolute value of the difference between the last transaction price in an interval, *t*, which corresponds to one second, minute, or hour, and the midpoint of the prevailing bid and ask prices.

We use the order imbalance (*OIB#*) metric proposed by Chordia et al. (2008) as a proxy for the level of order toxicity in the market. This is because existing order toxicity measures, such as the volume synchronised probability of informed trading (VPIN - see Easley et al., 2012), essentially capture the essence of order imbalance in the market and thus are highly correlated with *OIB#*. *OIB#* is computed as the absolute value of the number of buyer-initiated

[5] Chordia et al. (2008) and Avramov et al. (2006) employ midpoint returns to reduce bid-ask bounce.
[6] For robustness, we also employ the last bid and ask quotes for each interval.

trades minus the number of seller-initiated trades divided by the total number of trades during the interval, *t*. We employ the one-minute and one-hour intervals to compute market toxicity, because it is challenging to obtain enough trading volume for the lower volume stocks to compute unbiased order imbalance metrics within a one-second interval.

Apart from the main variables discussed above, there are a few other variables that are critical to our analysis. In our state space model, trading volume is an observable variable, which is decomposed into two unobservable variables – the expected/uninformed/liquidity and unexpected/informed components. Thus, the unexpected and expected components should be mechanically correlated with trading activity and volume. This implies that we need to include at least one proxy for trading volume and activity in our secondary models to control for volume. To this end, we employ the natural logarithm of trading volume as the first and main control for trading volume, since the state space-estimated components are driven by the evolution of trading volume (see also Chordia et al., 2002).[7] Our second trading activity-related proxy is the absolute value of buyer-less seller-initiated trades, which should adequately proxy trading activity because of Chordia et al.'s (2002) argument that the metric strongly affects prices and liquidity (see also Collin-Dufresne and Fos, 2015). Table 3 presents the summary statistics for the above variables. The descriptive statistics for the estimated unobservable expected and unexpected components of trading volume are presented in Table 4 and discussed in Section 4. The methodological approach form estimating the unobservable variables is also motivated in Section 4.

INSERT TABLE 3 ABOUT HERE

Table 3 presents the descriptive statistics for measures of liquidity, volatility and return computed over one-second intervals; market toxicity, constructed over one-minute intervals, is also presented. The average effective, relative, and quoted spreads are about 0.009, 0.0004, and 0.018, respectively. Average returns are weakly negative from October 2016 to September

[7] For robustness, we also use the level of trading volume as a proxy for trading activities and obtain completely consistent results.

2017. The mean and median for the absolute price change are about 0.0092 and 0.009 respectively. Average market toxicity (based on the order imbalance measure developed by Chordia et al. (2008)) is high at 0.54067, since it is computed over one-minute intervals.

4. Trading volume and the state space model

4.1. Motivating the application of state space modelling to trading volume

Transactions in financial markets are motivated either by the need for liquidity or the need to exploit information (see Admati and Pfleiderer, 1988). As predicted by the theoretical models of Kyle (1985) and Glosten and Milgrom (1985), liquidity and informed order flows have different impacts on price changes and the bid-ask spread (see also Collin-Dufresne and Fos, 2016; Wang, 1993). Avramov et al. (2006) empirically measure the relative impact of informed and liquidity traders on financial instruments and document the different impacts of these traders (see also Collin-Dufresne and Fos, 2015). In this paper, we aim to disentangle liquidity and informed trading volume using state space modelling and examine their relative impacts on price volatility, liquidity, and market toxicity. State space models are a natural tool for modelling an observed variable as the sum of two unobserved variables, and the asymptotic unbiasedness and efficiency of the models' estimation, i.e. maximum likelihood using Kalman filter (see Brogaard et al., 2014; Hendershott and Menkveld, 2014), make them best suited to analysing high frequency time series.

In our setting, the local level model decomposes trading volume into two parts. The first is a smoothed (level, constant) component of trading volume, which is driven by liquidity-seeking order flow, while the second is an irregular component of trading volume, which deviates from the smoothed (level) component and is therefore driven by informed order flow. This is a natural starting point, since the volume of liquidity order flow can be expected to remain relatively constant, while informed traders are not inclined to trade smoothly. These expectations are consistent with the predictions of the models of Easley and O'Hara (1992), Kyle (1985) and Huberman and Stanzl (2005). Firstly, in the Easley and O'Hara (1992) model, liquidity traders trade with a constant level of intensity; however, informed traders only trade when new information enters the market. These predictions demonstrate that liquidity trading can be modelled as a smoothed (constant) part of trading volume, and informed trading can be modelled as a deviation from this smoothed level; this is exactly what state space estimation does efficiently. Similar to Easley and O'Hara (1992), Huberman and Stanzl (2005) show that liquidity traders tend to trade a fairly fixed (constant) number of shares, because it helps them to minimise the mean and variance of the costs of trading; this again implies that liquidity-motivated trading volume can be considered as the smoothed (constant) part of trading volume. Furthermore, in Kyle's (1985) model, informed traders are not predisposed to trade smoothly when their trades would have no effect on execution price, which means that informed traders would not normally trade a constant number of shares. Therefore, their trading activity can be modelled as a deviation from the smoothed component of the trading volume. These factors strongly suggest that indeed, state space models are a natural and efficient tool for decomposing trading volume into liquidity and informed components.

Our approach involves modelling observable high frequency trading volume series as the sum of an unobservable nonstationary series (the expected component) and a stationary series (the unexpected component). Our argument that the expected component is primarily driven by liquidity trades and the unexpected component is mainly driven by information-motivated trades is based on the following reasons. Firstly, consistent with the literature, liquidity-motivated traders trade randomly (see as examples Glosten and Milgrom, 1985; Kyle, 1985). In the state space representation, the expected component is modelled as a (nonstationary) random walk, and hence it is reasonable to argue that liquidity traders drive the expected component, since if the random walk holds, all available information would have been incorporated into stock prices. Secondly, in state space models, the nonstationary (random walk) series, or the expected component, has a permanent impact on the observable variable. This implies that in our setting, liquidity-seeking order flow constitutes the permanent component of trading volume. While trading may not be informationally efficient in the absence of informed trades, they can still occur. This is not the case when liquidity-seeking order flow is lacking in the market. This is in line with the literature. The permanent character of liquidity order flow is confirmed by the no trade theorems. For example, both the Kyle (1985) and the Glosten and Milgrom (1985) models predict a breakdown of the price discovery process in the absence of liquidity traders, or when there is an excessive level of informed traders in the market relative to liquidity traders. This is simply because when there is a dearth of liquidity traders, market makers will aim to protect themselves against being adversely selected by setting the bid price low and the ask price high enough so as to preclude any trade. Furthermore, high levels of informed orders relative to liquidity orders implies that orders will cluster on one side of the order book, leading to no trade scenarios. In a related study, Morris (1994) shows that in order to solve no trade problems, the priority is to add liquidity traders to the market. This implies that without liquidity traders, trading in markets breaks down. Hence, liquidity traders are a critical permanent feature of financial markets (see also Brunnermeier, 2001; Milgrom and Stokey, 1982).

Information-motivated traders drive the unexpected component of trading volume for two reasons. Firstly, the information arrival process is an ‘unexpected’ process, and hence simple intuition suggests that information-motivated trades should be modelled as an unexpected component of trading volume. Secondly, according to Chordia et al. (2002), private information impacts liquidity temporarily in financial markets. Although information is a permanent component of stock prices (see Menkveld et al., 2007), it has a temporary impact on trading volume. One reason is that, according to the Efficient Market Hypothesis (EMH), any new information is simultaneously absorbed by traders, and hence it can only cause transitory (short-term) changes in trading volume (see Fama, 1970). The temporary character of informed traders is also predicted by the theoretical model of Suominen (2001). Suominen (2001) shows that after trading reveals the private information held by informed traders, liquidity traders will revise their pricing and thus become more cautious. This may result in a decrease of informed trading in the market. Thus, any changes in the information-driven component of trading volume, while having a durable impact on price, should only affect trading volume temporarily. Specifically, the implication here is that in our state space representation, the unexpected component has a transitory impact on the observable trading volume variable (see Hendershott and Menkveld, 2014).

Another stream of the finance literature lends further support to our methodological approach. When investigating trading behaviour in financial markets, modelling may focus on the duration between transactions as a means of capturing trading intentions, such that the time stamp may be used as an explanatory variable in the mean function of durations. In addition, a cubic spline may be used to smooth out huge variations in the duration effects. Such a model is often regarded as a state space counterpart of the autoregressive conditional duration (ACD) model of Engle and Russell (1998) (see also Durbin and Koopman, 2012).[8] The ACD is suitable for analysing trading data with transactions at irregular intervals, and the model is extensively used in the market microstructure literature to test hypotheses about duration and transaction clustering. In our state space representation, the permanent characteristics of the expected component imply constant duration, whereas the transitory structure of the unexpected component requires non-constant duration between transactions. Since the expected and unexpected components of trading volume are motivated by liquidity and information trades respectively, there should be constant (non-constant) duration in liquidity (informed) trading activity. For example, as transactions duration decreases, we would expect an increase in the speed of price adjustment to new information (see Dufour and Engle, 2000). Specifically, if indeed our state space representation is empirically relevant, then we would expect that non-constant duration or duration clustering is driven by informed trading. The empirical findings in the literature (see as examples Dufour and Engle, 2000; Engle, 2000; Russell and Engle, 2005; Zhang et al., 2001) are in line with this expectation, and therefore provide an additional set of robust arguments to further underscore the empirical relevance of our state space

[8] Pacurar (2008) provides a review of the duration modelling literature.

approach. The empirical findings can be explained by the predictions of the Easley and O'Hara (1992) model. However, ultimately, the ACD is an autoregressive model and is therefore less efficient for decomposing an observed variable into unobserved components than the state space modelling approach with maximum likelihood estimation and using Kalman filter (see Brogaard et al., 2014; Durbin and Koopman, 2012; Hendershott and Menkveld, 2014).

4.2. The state space equation and estimation

We model trading volume as the sum of a non-stationary expected (liquidity-driven) component and a stationary unexpected (information-driven) component.[9] In its simplest form, the structure of the state space model for trading volume can be expressed as:

$$v_{it} = m_{it} + s_{it} \tag{1}$$

and

$$m_{it} = m_{it-1} + u_{it} \tag{2}$$

where

$$v_{it} = ln(TVolume_{it}), \tag{3}$$

$TVolume_{it}$ is the volume traded in stock *i* at time *t*, m_{it} is a non-stationary expected component of the volume traded in stock *i* at time *t*, s_{it} is a stationary unexpected component of the volume traded in stock *i* at time *t,* and u_{it} is an idiosyncratic disturbance error. s_{it} and u_{it} are assumed to be mutually uncorrelated and normally distributed. Time, *t*, equals one-second, one-minute or one-hour. Although a one-second interval is a suitable frequency to investigate high-frequency trading activity, it is a very short interval for trade-based measures such as trading volume, hence we employ one-minute and one-hour interval analysis for robustness. Furthermore, any interval that has fewer than three transactions is excluded from the sample.

[9] In addition to modelling natural logarithm of trading volume as an observable variable in the state space representation, for robustness, we also employ percentage changes in trading volume and simple changes in trading volume. Our inferences are unchanged irrespective of the approach we employ, indeed all the estimates obtained are qualitatively similar.

The structure of the model shows that only changes on u_{it} affect the trading volume permanently; s_{it} is temporary because it affects trading volume only at a particular time. By using maximum likelihood (likelihood is constructed using the Kalman filter),[10] we can easily estimate σ_{it}^{2u} and σ_{it}^{2s}. According to the structure of our state space model, the expected component of trading volume is due to the activity of the fraction of the market populated by liquidity traders, while the other fraction of the market populated by informed traders reflect the unexpected component of trading volume. It implies that our estimated coefficients (σ_{it}^{2u} and σ_{it}^{2s}) can be used as proxies for the two fractions of the market's trading volume, i.e. σ_{it}^{2u} is a proxy for liquidity-motivated traders and σ_{it}^{2s} is a proxy for information-motivated traders. In order to jointly test the empirical relevance of the state space model and the role of informed and liquidity traders in functionality and the efficiency of financial markets, we employ predictive multivariate regressions as presented in the next section.

The model captured in Equations (1) – (3) is a special case of the general state space representation. The standard state space model is formulated for a vector of time series $\boldsymbol{v_t}$ with a frequency/time period t and is given by:

$$\boldsymbol{v_t} = \boldsymbol{W_t}\boldsymbol{\delta} + \boldsymbol{Z_t}\boldsymbol{m_t} + \boldsymbol{s_t}, \qquad \boldsymbol{m_{t+1}} = \boldsymbol{D_t}\boldsymbol{m_t} + \boldsymbol{R_t}\boldsymbol{u_t}, \ \ t = 1, .., N, \tag{4}$$

where disturbances $\boldsymbol{s_t} \sim N(\boldsymbol{0},\ \boldsymbol{S_t})$ and $\boldsymbol{u_t} \sim N(\boldsymbol{0},\ \boldsymbol{U_t})$ are mutually and serially uncorrelated (we ignore the stock notation i for simplicity). Furthermore, the initial state vector $\boldsymbol{m_1} \sim N(\boldsymbol{a},\ \boldsymbol{P})$ is uncorrelated with the disturbances. The mean vector $\boldsymbol{a}$ and variance matrix $\boldsymbol{P}$ are usually implied by the dynamic process for $\boldsymbol{m_t}$ in Equation (4) (see Durbin and Koopman, 2012). The remaining terms, $\boldsymbol{W_t}, \boldsymbol{Z_t}, \boldsymbol{D_t}, \boldsymbol{R_t}, \boldsymbol{S_t}$ and $\boldsymbol{U_t}$ are called system matrices and generally are

[10] The Kalman filter evaluates the conditional mean and variances of the state vector $\boldsymbol{m_t}$ given past observations $V_{t-1} = \{\boldsymbol{v_1}, .., \boldsymbol{v_{t-1}}\}$:

$$\boldsymbol{a_{t|t-1}} = \mathrm{E}(\boldsymbol{m_t}|V_{t-1}), \qquad \boldsymbol{P_{t|t-1}} = \mathrm{var}(\boldsymbol{m_t}|V_{t-1}), \quad t = 1, .., N.$$

In order to initialize the Kalman filter, we further have $\boldsymbol{a_{1|0}} = \boldsymbol{a}$ and $\boldsymbol{P_{1|0}} = \boldsymbol{P}$, where $\boldsymbol{m_1} \sim N(\boldsymbol{a},\ \boldsymbol{P})$. This initialization works only if $\boldsymbol{m_t}$ is a stationary process. However, as in our case, often $\boldsymbol{m_t}$ is not a stationary process. Hence, "diffuse initialization" should be done and estimated by numerically maximizing the log-likelihood, which may be evaluated by the Kalman filter due to prediction error decompositions. It can be shown that when the model is correctly specified the standardized prediction errors are normally and independently distributed with a unit variance (see Durbin and Koopman, 2012 for further details).

assumed fixed for $t = 1, .., N$. The elements of these system matrices are usually known; however, some elements that are functions of the fixed parameter vector need to be estimated. Our basic equations, (1) and (2), can be represented as the state space Equation (4) by choosing $\boldsymbol{v_t}$ as a single time series, where $\boldsymbol{W_t} = 0$, $\boldsymbol{Z_t} = \boldsymbol{D_t} = \boldsymbol{R_t} = 1$, $\boldsymbol{S_t} = \sigma_s{}^2$, and $\boldsymbol{U_t} = \sigma_u{}^2$. We observe that $\sigma_s{}^2$ and $\sigma_u{}^2$ vary for each frequency t for $t = 1, .., N$.

Unlike standard variable decomposition approaches, this model naturally deals with missing observations since the Kalman filter is used for its estimation. This is critical in a high frequency analysis. Recall the estimation principles of the Kalman filter (see Footnote 10): the estimation process in the case of missing observations is similar to that when we estimate with a full dataset. However, some adjustments are required. As can be deduced from Equations (1) – (2), the state space model consists of the measurement equation (Equation 1) and the transition equation (Equation 2). When we have missing observations in $\boldsymbol{v_t}$ the Kalman filter is not able to use the measurement equation, however the transition equation can be used since it depends on the previous estimated state ($\boldsymbol{m_{t+1}}$depends on $\boldsymbol{m_t}$**)** (see Equation 2). Specifically, the Kalman filtering implies that with missing observations in $\boldsymbol{v_t}$, the best estimation for $\boldsymbol{m_t}$ is simply the evaluation of the transition equation (see Durbin and Koopman, 2012).

We now report the estimates of the general state space model as presented in Equations (1) – (3).

INSERT TABLE 4 ABOUT HERE

Table 4 presents the cross-sectional mean estimated values of the expected (liquidity-driven) and unexpected (information-driven) components of trading volume as decomposed using the state space model. The results are presented for mean estimates based on one-second, one-minute, and one-hour estimations. For clarity, we divide our sample into quartiles according to their level of trading activity/activeness; trading activity is measured by trading volume. The stocks in Quartile 1 are the least active ones, whereas Quartile 4 contains the most active stocks. As expected, the mean of the variance of the unexpected component is consistently higher than the mean variance of the expected component, irrespective of the data frequency the model is estimated with. In addition, the estimates for the unexpected component's variance in each quartile is higher than the corresponding estimates for the expected component. There are at least two reasons for this distribution in the estimates. Firstly, consistent with the structure of our state space approach, informed trades are more informative than the liquidity trades. Secondly, according to the literature, liquidity traders tend to trade consistently for liquidity reasons (see as examples Easley and O'Hara, 1992; Huberman and Stanzl, 2005). By contrast, informed traders are likely to trade only if they have an informational advantage over other traders. It implies a higher variance for informed traders and our results are consistent with this expectation.

Informed traders strategically trade more actively when trading volume and liquidity trading is high, as higher trading volumes provide better "camouflage" for informed trades. The estimates presented in Table 4 are consistent with this widely held view in the market microstructure literature. The mean variance of liquidity-motivated trades in Quartile 4 is higher than the mean variance of liquidity traders in all of the other quartiles and is lowest in Quartile 1. This suggests that informed traders should be more active in Quartile 4 and least active in Quartile 1; the unexpected component estimates in Table 4 are completely in line with this expectation. The mean variance of the unexpected component in Quartile 4 are 1.51, 1.88 and 1.96 for the one-second, one-minute and one-hour estimations respectively. These estimates are 48%, 55.37% and 46.27% larger than the one-second, one-minute, and one-hour frequencies mean estimated values for Quartile 1 stocks at 1.02, 1.21, and 1.34 respectively. The above estimates underscore the significance of liquidity and informed order flows in financial markets. Inferring from the Kyle (1985) and Glosten and Milgrom (1985) models, when uninformed traders are scarce in the market, the price discovery process becomes impaired or even breaks down. When the opportunities of being compensated for gathering information are reduced, as happens in the market environment with few uninformed traders, fewer than optimal potential informed traders are incentivised to acquire information. The absence of informed traders in the markets impairs the price discovery process, since their trades convey information to the market. Thus, both liquidity and informed traders are critical to the price discovery process. An approach that allows us to directly estimate the proportion of trading volume that can be attributed to both types of traders is therefore valuable in several contexts, not least in market reporting activities, investment management, and policy/regulations development. For example, firm managers' responses to the so-called speeding ticket (Price and Volume Query) often issued by some exchanges, such as the Australian Securities Exchange, focuses mainly on explaining the evolution of trading volume, rather than attempt explanations of the information drivers of price.

4.3. A joint test of the empirical relevance of the state space model and the impact of trading volume components on market quality

In order to establish the empirical relevance of our state space approach for decomposing trading volume, we estimate a series of multivariate regressions to test whether the estimated components of trading volume's impact on market quality variables are consistent with the predicted and established patterns in the literature.

Kyle (1985) presents a theoretical model for deriving equilibrium security prices when traders' information sets are asymmetric. The model predicts that price volatility depends only on the informed trading volume and is independent of liquidity-based trading volume. In an associated work, Collin-Dufresne and Fos (2016) extend and generalize Kyle's (1985) model to show that informed trading-induced price volatility depends on the aggressiveness of informed traders. Thus, motivated by the predictions of the above-mentioned models, we jointly test the empirical relevance of the state space model and the roles of informed and liquidity traders in inducing price volatility by estimating the following regression:[11]

[11] Although we employ Pooled OLS (with panel corrected standard errors) for the primary estimations, for robustness, we also use fixed effects (stock and date) estimations with qualitatively similar outcomes/results obtained.

$$|\Delta p_{it}| = \alpha + \beta_1 Espread_{it-1} + \beta_2 TV_{it-1} + \beta_3 BSI_{it-1} + \beta_4 \sigma^{2s}_{it-1} + \beta_5 \sigma^{2u}_{it-1} + \varepsilon_{i,t} \qquad (5)$$

where $|\Delta p_{it}|$ is the absolute value of price changes for stock i at time t_{-1}, $Espread_{it-1}$ is the effective spread, measured as twice the absolute value of the difference between the last transaction price at time t_{-1} minus the prevailing bid-ask spread at the transaction time for stock i at time t_{-1}, $TV_{it-1}$is the natural logarithm of trading volume for stock i at time t-1, BSI_{it-1} is the absolute difference between buyer- and seller-initiated trades for stock i at time t_{-1}. σ^{2s}_{it-1} is the proxy for informed trading volume for stock i at time t_{-1} and σ^{2u}_{it-1} is the proxy for liquidity trading volume for stock i at time t_{-1}; both variables are obtained by maximum likelihood and from the state space estimation described in Section 4.2. The model is estimated at one-second, one-minute, and one-hour intervals. Consistent with literature, we use absolute price changes to measure price volatility and employ effective spread for controlling liquidity. As stated, we use the natural logarithm of trading volume as the observable variable in the state space model. This implies that our proxies for informed and liquidity trading are mechanically correlated with trading volume. We control for trading volume in the framework; the correlation coefficients in Table 5 show that the inclusion of the variable does not lead to multicollinearity concerns. Chordia et al. (2002) argue that prices and liquidity in financial markets are strongly affected by the difference between buyer- and seller-initiated trades. Therefore, we use the absolute difference between buyer- and seller-initiated trades as the additional proxy to control for the effect of trading volume, in addition to the natural logarithm of trading volume. σ^{2s}_{it-1} and σ^{2u}_{it-1} are the most important variables in the regression. If indeed our state space model correctly decomposes trading volume into liquidity and informed traders, we expect to see an insignificant relationship between σ^{2u}_{it-1} and price volatility after controlling for volume and liquidity, as Kyle (1985) argues that price volatility is not affected by liquidity traders. σ^{2s}_{it-1} on the other hand should be negatively and significantly correlated with price volatility, due to the absence of excessive aggressiveness in our sample period (see Collin-Dufresne and Fos, 2016). We also use a second proxy for volatility, i.e. the standard deviation of stock returns, which is a widely employed proxy in the literature (see as an example Lamoureux and Lastrapes, 1990). Consistent with the literature, we include the lagged value of the standard deviation of stock returns as an additional explanatory variable (see as examples, Justiniano and Primiceri, 2008; Schwert, 1989):

$$\sigma_{it}^{p} = \alpha + \ \beta_1 \sigma_{it-1}^{p} + \beta_2 Espread_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5 \sigma_{it-1}^{2s} + \beta_6 \sigma_{it-1}^{2u} + \varepsilon_{i,t} \quad (6)$$

Glosten and Milgrom's (1985) model is based on the idea that the extent of the adverse selection problem facing specialists when they trade with informed traders is one of the factors influencing the bid-ask spread. The model predicts that the bid-ask spread is positively correlated with informed trading; however, it is independent of the liquidity trading. The model is based on the assumption that informed traders exploit their information sets through the submission of market orders, i.e. they trade aggressively. However, Kaniel and Liu (2006) modify the Glosten and Milgrom (1985) model and show that informed traders with long-lived information tend to use limit orders rather than market orders (see also Menkveld, 2013). This implies that by submitting limit orders, informed traders might improve liquidity. In addition, the theoretical model presented by Admati and Pfleiderer (1988) shows that informed traders who observe the same signal will compete against each other in exploiting the information signal, and this may lead to the market maker facing a smaller adverse selection problem. When faced with reduced adverse selection risk, market makers will respond with tighter spreads. Motivated by the predictions of the above-mentioned theoretical models, we jointly test the empirical relevance of the state space model and the role of informed and liquidity traders in liquidity provision by using the following regression:

$$Spread_{it} = \alpha + \ \beta_1 \sigma_{it-1}^{p} + \beta_2 TV_{it-1} + \beta_3 BSI_{it-1} + \beta_4 \sigma_{it-1}^{2s} + \beta_5 \sigma_{it-1}^{2u} + \varepsilon_{i,t} \qquad (7)$$

where $Spread_{it}$ corresponds to one of relative, quoted, or effective spread. Quoted spread is the last ask price minus the last bid price at time *t*, while the relative spread is the quoted spread divided by the last mid-point at time *t*. Effective spread, TV_{it-1}, BSI_{it-1}, and σ_{it-1}^{p} are as previously defined and included to control for trading volume, order flow dynamics and volatility respectively. σ_{it-1}^{2s} and σ_{it-1}^{2u} are the key variables in the model and are as previously defined. The model is estimated at one-second, one-minute, and one-hour intervals. If indeed our state space model correctly decomposes trading volume into liquidity and informed traders, then we would expect to see no significant relationship between σ_{it-1}^{2u} and the various bid-ask spread metrics we use as dependent variables after controlling for volume, since Glosten and Milgrom (1985) argue that the bid-ask spread is not affected by liquidity-induced trading activity. By contrast, σ_{it-1}^{2s} should be significantly and negatively related to the bid-ask spread variables, because informed trading induces adverse selection, which is the major determinant of how wide the market maker spread is. A negative relationship between σ_{it-1}^{2s} and the spread is expected also because there is no evidence of excessive aggressiveness in our sample period (see Collin-Dufresne and Fos, 2015, 2016; Menkveld, 2013).

Finally, we investigate the role the informed trader plays in the creation of a toxic trading environment in the market. This is because the relationship between informed trading and market toxicity is a flipside question of the impact of informed traders on the functionality and efficiency of financial markets. In other words, questions about the role of informed traders in the inducement of market efficiency and the impact of informed traders on market toxicity are natural extensions of each other and one may not be fully explored without the other. Thus, we employ the following model to examine the relationship between market toxicity and informed trading:

$$MT_{it} = \alpha + \beta_1 Espread_{it-1} + \beta_2 TV_{it-1} + \beta_3 BSI_{it-1} + \beta_4 \sigma_{it-1}^{2s} + \beta_5 \sigma_{it-1}^{2u} + \varepsilon_{i,t} \quad (8)$$

where MT_{it} is the proxy for market toxicity and all of the other variables are as previously defined. We use the nominal order imbalance (OIB#) developed by Chordia et al. (2008), which captures buying and selling pressure, as a proxy for order flow toxicity. The Lee and Ready (1991) algorithm is used to classify trading volume into buys and sells. Thus, MT_{it} is calculated as the absolute value of the difference between the numbers of buy and sell trades, divided by the total number of trades:

$$MT = \frac{|\#Buy\ Trades - \#Sell\ Trades\ |}{\#Buy\ Trades + \#Sell\ Trades} \tag{9}$$

In a departure from the other models already presented, we estimate this model only at the one-minute and one-hour frequencies. This is because it is difficult to obtain enough trading volume to compute MT_{it} within the one-second intervals in an unbiased manner. According to Collin-Dufresne and Fos (2016) and Kaniel and Liu (2006), informed traders strategically choose to trade more when noise in trading is high. They also execute their trading strategies by submitting limit orders (passive orders) (see also Menkveld, 2013), which leads to a negative relationship between informed trading volume and market toxicity during normal trading sessions (see also Admati and Pfleiderer, 1988). Thus, we expect to see a negative correlation between σ^{2s}_{it-1} and market toxicity (see also Collin-Dufresne and Fos, 2015).

INSERT TABLE 5 ABOUT HERE

Table 5 presents a correlation matrix with all the variables featured in the above-presented models. The low correlation coefficient estimates among the variables (except for the liquidity proxies, which is expected) suggest that we do not have multicollinearity issues with the regression models. The results obtained from the estimation of Equations (5) – (8) are presented in Tables 6 – 8. Firstly, we discuss the one-second, one-minute, and one-hour frequency regression estimates for Equations (5) and (6). These are presented in Table 6.

INSERT TABLE 6 ABOUT HERE

The inferences drawn from the estimates in Table 6 are consistent across all frequency estimations. The coefficient estimates show that the lagged unexpected (information-driven) component of trading volume is a significant predictor of absolute price changes; all coefficients are statistically significant at the 0.01 level. In contrast, the liquidity/expected component is not a significant predictor of absolute price changes once we control for volume and liquidity. This is unsurprising since the latter component is liquidity driven and it is ‘expected’ in the sense that the trading activity generating it is based on information already incorporated into the price of the traded financial instruments. The results hold for both measures of price volatility that we employ, i.e. absolute price changes (presented in Panel A) and the standard deviation of stock returns (Panel B), although the unexpected component coefficient is generally larger in Panel A across all estimated frequencies. The negative coefficient estimates indicate that increases in information-motivated trades reduces price volatility in financial markets. This result is consistent with the result of the empirical study of Avramov et al. (2006), who find that stock price volatility is negatively correlated with informed traders. The significant unexpected component and the insignificant expected component estimates imply a validation of the empirical relevance of our state space approach to decomposing trading volume into informed and liquidity-driven components. As predicted by Kyle's (1985) model, the informed trading volume captured by our state space approach is significantly related to price volatility, however, the liquidity trading component is not.

We also note that while the coefficient estimates are consistent for all estimation frequencies across both panels, the impact of the unexpected component is stronger for lower frequencies. For example, in Panel A (B), the effect of the unexpected component on volatility proxies for the one-hour frequency estimation is 6.65 (124.28) and 98.80 (1,249) times larger than that of the one-minute and one-second frequency estimations respectively. These differences are due to more information being typically released over longer durations. It is plausible to expect that the market learns more about the developments relevant to an instrument over an hour than over a second or a minute, or at the very least, comes to terms more with the series of information over a longer time horizon. The estimated coefficients for all the other explanatory variables are consistent with the existing literature; trading volume and the effective spread are both positively and significantly correlated with price volatility (see Epps and Epps, 1976; Glosten and Milgrom, 1985).

The above-outlined results are consistent with the model presented by Collin-Dufresne and Fos (2016). The relationship between informed trading and price volatility is impacted by two factors. Firstly, informed traders' activity in the market leads to the revelation of information and this new information reduces price uncertainty in financial markets. The reduction in price uncertainty in turn spurs a reduction in price volatility. Secondly, informed traders may trade aggressively in a liquidity-constrained environment and thereby increase aggressiveness in financial markets, and this may increase price volatility. Thus, the relationship between informed traders and price volatility depends on the aggressiveness of informed traders. The relationship will be positive if informed traders use aggressive orders (market orders) and create excessive aggressiveness in the market. Interestingly, in related papers, Menkveld (2013) and Rzayev and Ibikunle (2017) show that aggressive orders are not profitable during normal trading periods, i.e. if there is no extreme volatility in financial markets, then the use of aggressive market orders offers no trading advantage to informed traders. The implication here is that informed traders seldom submit aggressive orders during normal trading days. Hence, as we do not observe any instance of excessive aggressiveness in our sample for the period we focus on, we would expect to find the negative impact of informed trading on stock price volatility reported in Table 6 (see also Wang, 1993).

The explanatory powers of the one-second regressions are low, with the *Adjusted R^2* being only about 0.40% for absolute price changes in Panel A and 0.92% for standard deviation of stock returns in Panel B. This is unsurprising and is due to our employment of a one-second frequency for the models' estimations, with very little information being released throughout the duration (see Chordia et al., 2008). Consequently, the *Adjusted R^2* estimates are larger for the one-minute and one-hour frequencies, which are 1.71% and 5.27% respectively in Panel B.

INSERT TABLE 7 ABOUT HERE

We now turn to the relationship between liquidity and the decomposed trading volume components. We estimate Equation (7) for this purpose. In Table 7, we present the model's estimates, and Panels A, B, and C show the results with relative, quoted, and effective spread measures as respective proxies for liquidity. For each liquidity proxy, we estimate Equation (7) at one-second, one-minute, and one-hour frequencies. The estimates show that, consistent with the predictions of Glosten and Milgrom's (1985) model predictions, the lagged unexpected component is a significant predictor of liquidity. The estimates for the lagged unexpected component of trading volume are negative and statistically significant at the 0.01 level irrespective of which liquidity proxy we employ. By contrast, the expected component is not significantly related with bid-ask spread after controlling for volume and order flow dynamics. The results in all of Table 7's panels indicate that the state space model we employ in this study appropriately decomposes trading volume into liquidity- and information-driven components. Consistent with the results in Table 6, our results show that the information-driven component is negatively (positively) correlated with the bid-ask spread (liquidity). Negative coefficients indicate that informed traders are more likely to consume liquidity in financial markets rather than provide it; in this case, they are liquidity consumers. The results are consistent with the findings of Collin-Dufresne and Fos (2015). The coefficients of all control variables are in line with the consistent literature. Similar to the price volatility model, *Adjusted R^2* values in Panels A, B, and C are generally small for the one-second and one-minute high frequency estimations, with estimates ranging from 0.49% to 1.45%. The low *Adjusted R^2* values are due to the estimation frequencies. Hence, the one-hour frequency models have much higher levels of explanatory powers. In Panels A, B, and C, the *Adjusted R^2* values are 14.01%, 11.15% and 10.18% respectively.

INSERT TABLE 8 ABOUT HERE

Finally, in this section, we examine the regression estimates based on an investigation of the impact of liquidity and informed traders on market toxicity (as shown in Equation 8). Table 8 presents the estimated coefficients for the model estimated at one-minute and one-hour frequencies. Consistent with the results in Tables 6 and 7, the lagged unexpected component of trading volume is negatively and statistically significantly related with market toxicity at the

0.01 level; however, the expected component is not, after we control for volume and liquidity. The inverse relationship between the market toxicity proxy and the unexpected component suggests that information-motivated trading volume reduces order flow toxicity in financial markets, even after controlling for the overall impact of trading volume and liquidity. At least two mechanisms could explain this observed effect. Firstly, theoretical models like that of Glosten and Milgrom (1985) assume that informed traders use aggressive orders (market orders) to execute their trading strategies, and hence they increase the bid-ask spread and induce adverse selection risk/market toxicity. However, upon the modification of Glosten and Milgrom's (1985) model, Kaniel and Liu (2006) show that informed traders with long-loved information tend to use limit orders rather than market orders during normal trading periods (see also Menkveld, 2013). The prediction of Kaniel and Liu's (2006) model is empirically confirmed by Collin-Dufresne and Fos (2015). Thus, informed traders might use limit orders, which contribute to a reduction of the bid-ask spread by removing uncertainty in instruments' prices, as long as the trading period is not aggressive. In addition, the theoretical model presented by Admati and Pfleiderer (1988) shows that informed traders who observe the same signal will compete against each other in exploiting the information signal, and this may lead to the market maker facing a smaller adverse selection problem. When faced with reduced adverse selection, market makers will respond with tighter spreads, implying a reduction in toxic order flow.

Although all other control variables are significant in the one-minute frequency model estimation, the explanatory power of the regression is small with the *Adjusted R^2* being only about 0.12%, again owing to the short horizon over which the model is estimated. This is underscored by the larger *Adjusted R^2* value for the one-hour frequency estimation at 2.84%

4.4. Predicting short-horizon returns using the unexpected (information-driven) component of trading volume

According to Fama (1970), (developed) financial markets are largely informationally efficient over a daily horizon. Chordia et al. (2008) argue that although markets are quite efficient over a long-horizon, there are inefficiencies in markets at shorter horizons because traders need time to act on new information. Motivated by this, Chordia et al. (2008) examine the predictability of short-term returns from past order imbalance and find that, indeed, markets are inefficient over short periods. In their study, Chordia et al. (2008) employ order imbalance as an explanatory variable because order imbalance signals private information, due to its capturing of buying and selling pressure. They show that short horizon returns predictability is smaller when markets experience periods of relative liquidity. The elimination of short horizon predictability is driven by the information-driven component of the order flow rather than increased order flow as a whole. Thus, we expect our estimated information-driven component of trading volume to be negatively correlated with short-horizon returns. This is because informed trading eliminates arbitrage opportunities. In addition to eliminating short horizon return predictability, informed trading decreases price volatility as long as there is no case of excessive aggressiveness in financial markets. Therefore, the risk premium demanded by the traders should decrease with the volume of information-motivated traders in the market (see Wang, 1993). This analysis serves as a further test of the empirical relevance of the state space modelling approach for estimating liquidity and informed trading components of trading volume. The estimated regression model is as follows:

$$R_{it} = \alpha + \beta_1 \sigma_{it-1}^{p} + \beta_2 Espread_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5 \sigma_{it-1}^{2s} + \varepsilon_{i,t} \qquad (10)$$

where R_{it} is the midpoint return for stock *i* at time *t*; all of the other variables are as previously defined. All variables are computed over a one-second frequency. It could be insightful to estimate the model over a lower frequency, such as the one-minute interval, as well. The reason for this is that the trading volume in our sample appears to be mainly driven by HFTs, given the sample period and market we focus on (see Brogaard et al., 2014). Thus, if HFTs are responsible for driving a substantial proportion of the informed trading volume, the predictability of return should be greatly diminished over a one-minute interval, since a one-minute interval cannot be considered a short-horizon for an HFT-driven market. Thus, we estimate the following regression at a one-minute frequency; the only difference to Equation (10) is the addition ofMT_{it}, which can only be validly computed at a minimum frequency of about one minute:

$$R_{it} = \alpha + \beta_1 \sigma^{p}_{it-1} + \ \beta_2 Espread_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5 \sigma^{2s}_{it-1} + \beta_6 MT_{it-1} + \varepsilon_{i,t} \quad (11)$$

σ^{2s}_{it-1} is the most important variable in both Equations (10) and (11) regression; we expect to see a significant and inverse relationship between informed trading and future short-horizon return for Equation (10), estimated at the one-second frequency. In Equation (11), we expect that both MT_{it-1} and σ^{2s}_{it-1} should be insignificant at the one-minute interval because of the superfast trading systems of HFTs trading in S&P 500 stocks.

INSERT TABLE 9 ABOUT HERE

Table 9 presents the estimated coefficients for Equation (10). All of the coefficients, including the unexpected component variable, are statistically significant at the 0.01 level. The unexpected component estimate is negative; this suggests that an increase in the level of informed trading eliminates/reduces return predictability/arbitrage. Thus, the unexpected component of trading volume, as obtained using the state space model approach, signals private information similar to the order imbalance metrics developed by Chordia et al. (2008). The *Adjusted R^2* is 0.06%. As already discussed, the low *Adjusted R^2* is linked to the estimation frequency of the regression model, which is one second in this case.

We next estimate a similar regression model (Equation 11) over a longer time frequency of one-minute; the results are presented in the final column of Table 9. As predicted, the unexpected component is not statistically significant, owing to the lack of return predictability over a time period stretching into a minute. However, the *Adjusted* R^2 coefficient at 0.09% is larger than for the one-second frequency estimation in Equation (10). The lack of statistical significance for the unexpected component in the one-minute frequency regression model is due to the prevalence of HFT activity in the data we use, and the ability of HFTs to eliminate arbitrage opportunities very quickly, leading to the elimination of return predictability at low frequencies. We also include the order imbalance metric used by Chordia et al. (2008) in the regression model and, in contrast to the results presented by Chordia et al. (2008), the metric is not significant here. This shows that while one-second stock return is predictable from lagged metrics that signal private information, one-minute stock returns are not predictable in financial markets dominated by HFTs.

A key finding here is that although the lag of the unexpected component predicts one-second stock returns, one-minute stock returns are not predictable using either the unexpected component or the order imbalance metric based on Chordia et al. (2008) model. Thus, the latter part of the findings is not consistent with the results presented by Chordia et al. (2008), who show that even five-minute stock returns can be predicted from past order imbalance. The inconsistency here is linked to the data period employed by both studies. While Chordia et al. (2008) employ a dataset covering the years 1993 to 2002, when HFTs were not the main drivers of trading in financial markets, we employ a much more recent dataset from 2016 to 2017. For example, based on an analysis of data, which predates ours by seven years, Brogaard et al. (2014) show that at least fifty percent of New York's trading volume is driven by HFTs. It implies that the speed of price adjustment through the incorporation of new information has become much lower. Specifically, HFTs do not need a full minute to absorb and act on new information. Furthermore, Brogaard et al. (2014) show that HFTs are more active in large stocks. As our sample consists of the most active and largest stocks in U.S. financial markets, we expect that HFTs are the dominant traders in our sample period. Thus, the definition of short-horizon has shifted since the period investigated by Chordia et al. (2008); the one or five-minute (as in the case of Chordia et al., 2008) horizons cannot be considered as short-horizons for the purpose of predicting short-horizon returns. The negative relationship between the unexpected component and the one-second short-horizon return documented above is due to a decrease in the risk premium demanded by the traders when informed trading reduces volatility in the absence of excessive aggressiveness in the market.

4.5.Does fast trading drive the elimination of return predictability?[12]

Further to the above findings, we address the role of HFTs in the elimination of return predictability. In comparison with non-HFTs, HFTs could be viewed as being informed, simply on the basis that they trade with either private or public information (e.g. the sudden arrest of a firm's CEO for fraudulent activities) at a faster pace than non-HFTs. This is what is referred to as latency arbitrage; it involves the exploitation of a trading time disparity between fast and slow traders, when that trade is executed solely because of a latency advantage. Ibikunle (2018) argues that this speed advantage is tantamount to an information advantage when traders trade at different speeds, since the end result remains the same – a set of traders exploit information (whether private or public) ahead of a different set of traders. Thus, exchanges with infrastructures that especially accommodate HFTs tend to display efficient prices ahead of others when instruments are simultaneously traded across those exchanges. This is the case with the analysis of price leadership in the London equity market conducted by Ibikunle (2018). Chaboud et al. (2014) and Brogaard et al. (2014) also show that HFTs enhance informational efficiency by speeding up price discovery and eliminating arbitrage opportunities. This property is consistent with what the classical informed trader in the market microstructure literature does with her trading activity.

In order to capture the transitory nature of informed trading volumes as encapsulated by HFT activity, we design a test to capture transitory informed trading in the market when arbitrageurs observe that instruments' prices have deviated from their underlying values. It is

[12] We thank an anonymous referee for suggesting this analysis.

critical to note that while HFTs could be considered informed in comparison with non-HFTs, not all HFTs employ arbitrage strategies. Menkveld (2013) and Hagströmer and Nordén (2013) show that the majority of HFTs (about 80%) tend to apply market making strategies. Furthermore, in a market dominated by algorithmic traders (ATs) the speed advantage will not consistently confer appreciable advantages over the also fast competition. Thus, our test is designed to capture the changes in HFT volumes attributable to informed HFT activity.

For this test, we employ the transactions dataset for 120 NASDAQ and NYSE stocks provided to us by NASDAQ. The data disaggregates transactions into HFT and non-HFT transactions for the year 2009. Employing the dataset, we re-estimate Equations (10) and (11) with one additional variable, $D_{HFT,t-1} * \sigma^{2s}_{it-1}$:

$$R_{it} = \alpha + \beta_1\sigma^{p}_{it-1} + \beta_2 Illiq_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5\sigma^{2s}_{it-1} + \beta_6 D_{HFT,t-1} * \sigma^{2s}_{it-1} + \varepsilon_{i,t} \qquad (12)$$

$$R_{it} = \alpha + \beta_1\sigma^{p}_{it-1} + \beta_2 Illiq_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5\sigma^{2s}_{it-1} + \beta_6 MT_{it-1} + \beta_7 D_{HFT,t-1} * \sigma^{2s}_{it-1} + \varepsilon_{i,t} \qquad (13)$$

$D_{HFT,t-1} * \sigma^{2s}_{it-1}$ is obtained by interacting a new variable, *D_{HFT}*, with the unexpected component variable. D_{HFT} is a dummy equalling one during periods of high HFT activity. $Illiq_{it}$ is a proxy for illiquidity and corresponds to the Amihud (2002) illiquidity ratio in both Equation (12) and Equation (13) when the NASDAQ-provided data is employed (results presented in Panel A of Table 10) and the Effective spread when the TRTH data is employed (results presented in Panel B of Table 10). In order to determine intervals of high HFT activity, we compute the proportion of HFT trades to non-HFT trades using the designations (HFT/non-HFT) for the transactions in the NASDAQ data. A one-second or one-minute interval is designated as an interval of high HFT activity if the proportion of HFT trades for that interval is one standard deviation higher than the mean for the surrounding -60, +60 corresponding intervals. We employ only one-second and one-minute frequencies because the existing literature (see as an example, Chordia et al., 2008) shows that short horizon predictability is eliminated within minutes. The NASDAQ data, as pointed out by Brogaard et al. (2014), does not identify all HFTs. Hence, for robustness, we employ an alternative measure of HFT activity in our analysis; this is the widely deployed proxy based on the ratio of messages to the number of transactions (see as examples, Boehmer et al., 2015; Malceniece et al., 2018). As in Equations (10) and (11), Equations (12) and (13) are estimated at one-second and one-minute frequencies respectively. If the interaction variable's coefficient is negative and statistically significant, it implies that an unexpected (transitory) rise in HFT activity is informed and reduces return predictability. This conclusion will be especially strengthened if the unexpected component is not statistically significant in Equations (12) and (13), since it would mean that the reduction in return predictability is primarily driven by unexpected HFT volumes. A result of this nature would be in line with the assumptions underlying our state space modelling approach. Informed trading volume is transitory and only arises to exploit deviations in the price of an instrument from its fundamental value. If the interaction variable in Equations (12) and (13), which captures periods of HFT spurts, is negative and statistically significant, it would show that HFT activity above the (expected) mean indicates transitory informed trading volumes.

INSERT TABLE 10 ABOUT HERE

We present the results based on the two approaches for estimating D_{HFT} in Table 10; Panel A shows the results using the NASDAQ-defined HFT/non-HFT transactions, while Panel B shows the results using the ratio of messages to transactions HFT proxy. Contrary to the results in Table 9, although it remains negative, the unexpected component coefficients for the one-second frequency estimation in both panels are not statistically significant. However, when the unexpected component variable is interacted with an HFT dummy, it is highly statistically significant, while retaining its negative sign. This implies that the reduction in the return predictably observed in the earlier analysis is driven by informed HFT activity. Consistent with the assumption underlying our state space modelling approach, unexpected trading volume, i.e. an increase in HFT volume above the mean, aids the speedy incorporation of information into instruments' prices and leads to the elimination of arbitrage opportunities.

4.6. The evolution of the unexpected (information-driven component) around earnings announcements

In this section, we conduct a final test of the information relevance of our state space model-based informed trading proxy. Specifically, we examine the behaviour of the informed trading proxy (unexpected/transitory component of trading volume) before and after earnings announcements. We focus on earnings announcements as information events for two reasons. Firstly, there is an overwhelming set of evidence on information leakage prior to these events (see as an example Christophe et al., 2004); this implies that earnings announcements provide an ideal basis for testing our estimated proxy for informed trading. Secondly, testing the behaviour of informed trading proxies by using earnings announcements is a well-established and widely accepted approach in the market microstructure literature (see as examples Benos and Jochec, 2007; Easley et al., 2008). To the extent that the unexpected component (proxy for informed trading) is successful at estimating the existence of asymmetric information, we expect that the unexpected component estimates for the days preceding earnings announcements would be significantly higher than the unexpected component estimates for the days after earnings announcements. This is due to the effects of information leakage prior to earnings announcements. We obtain earnings announcement dates from CompuStat. During our sample period, there is a total of 397 earnings announcements for the stocks in our sample. We compute the cross sectional averages of the unexpected component for 21 working days before and after each earnings announcement day, and test the null that the unexpected component before and after earnings announcements are equal. The null is tested using a two-sample t-test and pairwise Wilcoxon-Mann-Whitney U test.[13]

This section presents and discusses the results on the analysis of the evolution of the unexpected component around earnings announcements. Consistent with the literature (see as an example Christophe et al., 2004), the unexpected component should be higher in the lead up to earnings announcements than following such announcements, due to information leakage.

INSERT TABLE 11 ABOUT HERE

Table 11 presents the cross-sectional averages of the unexpected component prior to and after earnings announcements. We compute the cross sectional means of the unexpected components over two different event windows: [-21, -1], the pre-event window, and [+1, +21], the post-event window. As expected, the cross sectional mean of the value of the unexpected component before earnings announcements is 1.8% higher than its value after earnings announcements, and both two-sample t-tests and pairwise Wilcoxon-Mann-Whitney U tests show that the difference between these two periods is statistically significant at the <0.001 level.

Since our proxies for informed and liquidity trading are derived from trading volume, there is a mechanical correlation between the proxies and trading volume. This implies that the variation in trading volume before and after earnings announcements is the sole driver of the evolution of the unexpected component around the earnings announcements. Therefore, in order to eliminate the possibility that our proxy for informed trading works because of the correlation between the unobservable (unexpected and expected) components and the observable (trading volume) variable in state space representation, we examine the behaviour of trading volume before and after earnings announcements as well. As seen in Table 11, in contrast to the evolution of the unexpected component, trading volume is statistically significantly higher after the earnings announcement days than prior to those days. This is

[13] We also employ a 15-day event window for this analysis, i.e. seven days before and after each earnings announcement. The results are consistent with the ones we report in Table 11.

unsurprising; when investors have time to evaluate the information contained in earnings announcements, trading activity increases because they act on it. However, at that time the information will no longer have profit value or drive price because informed traders would have exploited it in the run up to the official announcements. In summary, the results in this section show that our state space-based informed trading proxy captures information prior to earnings announcements, which is in line with the information leakage literature's findings. Thus, the results bolster the empirical relevance of our state space approach to estimating informed and liquidity trading proxies from trading volume.

5. Conclusion

In this paper, we develop a state space model for decomposing trading volume into liquidity-driven (expected) and information-driven (unexpected) components. There are two central assumptions underlying the specification of the state space approach we use. Firstly, we argue that the expected component from the model is mainly driven by liquidity-seeking order flow, and secondly, that the unexpected component as motivated is primarily driven by information-motivated order flow. In addition to providing a robust set of arguments grounded in the literature to back up our claims, we further develop a set of univariate analysis and multivariate regression models to formally test these arguments. Firstly, we find that the unexpected component obtained from the state space model is significantly correlated with volatility, liquidity and toxicity, even after controlling for volume (and in the case of volatility and toxicity, we also control for liquidity in addition to volume), whereas the expected component is not significantly related to them once volume and liquidity are controlled for. These results are consistent with the theoretical models presented in Kyle (1985) and Glosten and Milgrom (1985); the consistency therefore implies that the expected and unexpected components can be viewed as encapsulating the liquidity- and information-motivated trades in our sample, respectively. The findings can also be linked to informed traders not using market (aggressive) orders during normal trading periods, when there are no upheavals or extreme liquidity constraints in the market, as predicted by Kaniel and Liu (2006) and Menkveld (2013).

Furthermore, we demonstrate that the unexpected component is a significant predictor of short-horizon returns. This again shows that the unexpected component signals private information, which is due to its capturing information-motivated trading volume. The estimated and statistically significant negative relationship between the lag unexpected component of trading volume (informed trading) and one-second short-horizon return is linked to a reduction in the risk premium demanded by the traders, given that increased informed trading is linked with a reduction in price volatility during the normal trading period, i.e. in the absence of excessive aggressiveness in trading. However, in contrast to Chordia et al. (2008), we find that one-minute returns cannot be predicted using either the unexpected component metric or the order imbalance, as employed by Chordia et al. (2008) for a five-minute return. This implies that in today's high frequency trading environment, arbitrage opportunities are eliminated at a much faster rate than in the early 2000s period examined by the latter study. We show that this sharp decline in the window for return predictability is driven by informed HFT activity.

Finally, we show that, in line with expectations based on the information leakage literature, on average the unexpected component before earnings announcements is statistically and significantly higher than its value after earnings announcements.

References

Admati AR, Pfleiderer P. A Theory of Intraday Patterns: Volume and Price Variability. The Review of Financial Studies 1988; 1; 3-40

Amihud Y. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 2002; 5; 31-56

Avramov D, Chordia T, Goyal A. The Impact of Trades on Daily Volatility. The Review of Financial Studies 2006; 19; 1241-1277

Benos E, Jochec M. 2007. Testing the PIN variable. University of Illinois: Illinois; 2007.

Bessembinder H, Seguin PJ. Price Volatility, Trading Volume, and Market Depth: Evidence from Futures Markets. Journal of Financial and Quantitative Analysis 1993; 28; 21-39

Boehmer E, Fong KYL, Wu J. 2015. International evidence on algorithmic trading. American Finance Association (AFA) Annual Meeting. San Diego; 2015.

Brogaard J, Hendershott T, Riordan R. High-Frequency Trading and Price Discovery. The Review of Financial Studies 2014; 27; 2267-2306

Brunnermeier MK. Asset Pricing Under Asymmetric Information: Bubbles, Crashes, Technical Analysis, and Herding. Oxford University Press, Oxford; 2001.

Chaboud AP, Chiquoine B, Hjalmarsson E, Vega C. Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market. The Journal of Finance 2014; 69; 2045-2084

Chakrabarty B, Moulton P, Shkilko A. Evaluating Trade Classification Algorithms: Bulk Volume Classification versus the Tick Rule and the Lee-Ready Algorithm. Journal of Financial Markets 2015; 25; 52-79

Chordia T, Roll R, Subrahmanyam A. Market Liquidity and Trading Activity. Journal of Finance 2001; 56; 501-530

Chordia T, Roll R, Subrahmanyam A. Order imbalance, liquidity, and market returns. Journal of Financial Economics 2002; 65; 111-130

Chordia T, Roll R, Subrahmanyam A. Evidence on the speed of convergence to market efficiency. Journal of Financial Economics 2005; 76; 271-292

Chordia T, Roll R, Subrahmanyam A. Liquidity and market efficieny. Journal of Financial Economics 2008; 87; 249-268

Christophe SE, Ferri MG, Angel JJ. Short-Selling Prior to Earnings Announcements. The Journal of Finance 2004; 59; 1845-1875

Clark PK. A subordinated stochastic process model with finite variance for speculative prices. Econometrica 1973; 41; 135-155

Collin-Dufresne P, Fos V. Do Prices Reveal the Presence of Informed Trading? The Journal of Finance 2015; 70; 1555-1582

Collin-Dufresne P, Fos V. Insider Trading, Stochastic Liquidity, and Equilibrium prices. Econometrica 2016; 84; 1441-1475

Copeland TE. A Model of Asset Trading Under the Assumption of Sequential Information Arrival. The Journal of Finance 1976; 31; 1149-1168

Copeland TE. A Probability Model of Asset Trading. Journal of Financial and Quantitative Analysis 1977; 12; 563-578

Cornell B. The Relationship between Volume and Price Variability in Futures Markets. Journal of Futures Markets 1981; 1; 303-316

Daigler RT, Wiley MK. The Impact of Trader Type on the Futures Volatility-Volume Relation. The Journal of Finance 1999; 54; 2297-2316

Dufour A, Engle RF. Time and the impact of a trade. The Journal of Finance 2000; 55; 2467–2498

Durbin J, Koopman S. Time Series Analysis by State Space Models. Oxford University Press, Oxford, UK; 2012.

Easley D, De Prado M, O'Hara M. The microstructure of the "flash crash": flow toxicity, liquidity crashes, and the probability of informed trading. Journal of Portfolio Management 2011; 37; 118-129

Easley D, De Prado M, O'Hara M. Flow Toxicity and Liquidity in a High-frequency World. The Review of Financial Studies 2012; 25; 1457-1493

Easley D, Engle RF, O'Hara M, Wu L. Time-Varying Arrival Rates of Informed and Uninformed Trades. Journal of Financial Econometrics 2008; 6; 171-207

Easley D, Kiefer N, O'Hara M, Paperman J. Liquidity, Information, and Infrequently Traded Stocks. The Journal of Finance 1996; 51; 1405-1436

Easley D, Kiefer NM, O'Hara M. One Day in the Life of a Very Common Stock. The Review of Financial Studies 1997; 10; 805-835

Easley D, O'Hara M. Price, trade size, and information in securities markets. Journal of Financial Economics 1987; 19; 69-90

Easley D, O'Hara M. Time and the process of security price adjustment. The Journal of Finance 1992; 47; 577-606

Engle RF. The econometrics of ultra-high frequency data. Econometrica 2000; 68; 1–22

Engle RF, Russell JR. Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 1998; 66; 1127-1162

Epps TW, Epps ML. The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis. Econometrica 1976; 44; 305-321

Fama E. Efficient capital markets: a review of theory and empirical work. The Journal of Finance 1970; 25; 383-417

Glosten L, Milgrom P. Bid, ask, and transaction prices in a specialist market with heterogeneously informed agents. Journal of Financial Economics 1985; 14; 71-100

Hagströmer B, Nordén L. The diversity of high-frequency traders. Journal of Financial Markets 2013; 16; 741-770

Harris L. Cross-security Tests of the Mixture of Distributions Hypothesis. Journal of Financial and Quantitative Analysis 1986; 21; 39-46

Harris L. Transactions Data Tests of the Mixture of Distributions Hypothesis. Journal of Financial and Quantitative Analysis 1987; 22; 127-141

Harris M, Raviv A. Differences of opinion make a horse race. The Review of Financial Studies 1993; 6; 473-506

Hasbrouck J. Measuring the Information Content of Stock Trades. The Journal of Finance 1991; 46; 179-207

Hellwig MF. On the Aggregation of Information in Competitive Markets. Journal of Economic Theory 1980; 22; 477-498

Hendershott T, Menkveld AJ. Price pressures. Journal of Financial Economics 2014; 114; 405-423

Huberman G, Stanzl W. Optimal Liquidity Trading. Review of Finance 2005; 9; 165-200

Ibikunle G. Opening and closing price efficiency: Do financial markets need the call auction? Journal of International Financial Markets, Institutions & Money 2015; 34; 208-227

Ibikunle G. Trading places: Price leadership and the competition for order flow. Journal of Empirical Finance 2018; 49; 178-200

Jennings RH, Starks LT, Fellingham JC. An Equilibrium Model of Asset Trading with Sequential Information Arrival. The Journal of Finance 1981; 36; 143-161

Justiniano A, Primiceri GE. The Time-Varying Volatility of Macroeconomic Fluctuations. American Economic Review 2008; 98; 604-641

Kaniel R, Liu H. So What Orders Do Informed Traders Use? The Journal of Business 2006; 79; 1867-1913

Karpoff JM. The Relation Between Price Changes and Trading Volume: A Survey. Journal of Financial and Quantitative Analysis 1987; 22; 109-126

Kyle AS. Continuous Auctions and Insider Trading. Econometrica 1985; 53; 1315-1335

Lamoureux CG, Lastrapes WD. Heteroskedasticity in Stock Return Data: Volume versus GARCH effects. The Journal of Finance 1990; 45; 221-229

Lee C, Ready M. Inferring Trade Direction from Intraday Data. The Journal of Finance 1991; 46; 733-746

Malceniece L, Malcenieks K, Putniņš TJ. High Frequency Trading and Co-Movement in Financial Markets. Journal of Financial Economics 2018; Forthcoming.

McCarthy J, Najand M. State Space Modeling of Price and Volume Dependence: Evidence from Currency Futures Journal of Futures Markets 1993; 13; 335-344

Menkveld AJ. High frequency trading and the new market makers. Journal of Financial Markets 2013; 16; 712-741

Menkveld AJ, Koopman SJ, Lucas A. Modeling Around-the-Clock Price Discovery for Cross-Listed Stocks Using State Space Methods. Journal of Business & Economic Statistics 2007; 25; 213-225

Milgrom PR, Stokey N. Information, trade and common knowledge. Journal of Economic Theory 1982; 26; 17-27

Morris S. Trade with Heterogeneous Prior Beliefs and Asymmetric Information. Econometrica 1994; 62; 1327-1347

Pacurar M. Autoregressive conditional duration models in finance: A survey of the theoretical and empirical literature. Journal of Economic Surveys 2008; 22; 711-751

Russell JR, Engle RF. A discrete-state continous-time model of financial transactions prices and times: the autoregressive conditional multinomialautoregressive conditional duration model. Journal of Business and Economic Statistics 2005; 23; 166–180.

Rzayev K, Ibikunle G. 2017. Order aggressiveness and flash crashes. 2017.

Schwert GW. Why Does Stock Market Volatility Change Over Time? The Journal of Finance 1989; 44; 1115-1153

Shalen CT. Volume, volatility, and the dispersion of beliefs. The Review of Financial Studies 1993; 6; 405-434

Smirlock M, Starks L. 1984. A Transactions Approach to Testing Information Arrival Models. Washington University: Washington; 1984.

Sun Y, Ibikunle G. Informed trading and the price impact of block trades: A high frequency trading analysis. International Review of Financial Analysis 2016; 0; 114-129

Suominen M. Trading Volume and Information Revelation in Stock Markets. the Journal of Financial and Quantitative Analysis 2001; 36; 545-565

Van Ness BF, Van Ness RA, Yildiz S. The role of HFTs in order flow toxicity and stock price variance, and predicting changes in HFTs' liquidity provisions. Journal of Economics and Finance 2016; 41; 739-762

Wang J. A Model of Intertemporal Asset Prices Under Asymmetric Information. The Review of Economic Studies 1993; 60; 249-282

Zhang MY, Russell JR, Tsay RS. A nonlinear autoregressive conditional duration model with applications to financial transaction data. Journal of Econometrics 2001; 104; 179–207

Table 1. Summary of trading activity

The table presents trading summary statistics for the most active 100 S&P 500 stocks from October 1, 2016 through to September 30, 2017. The Lee and Ready (1991) algorithm is used to classify trades as buyer- and seller-initiated.

Buyer-initiated (000,000s)	Seller-initiated (000,000s)	Total trades (000,000s)
106.89	109.48	216.37

Buyer-initiated (00,000,000s)	Seller-initiated (00,000,000s)	Total trading volume (00,000,000s)
347.71	375.61	723.32

Buyer-initiated	Seller-initiated	Average trade sizes
325.30	343.09	334.30

Buyer-initiated ($'0,000,000,000)	Seller-initiated ($'0,000,000,000)	Total USD volume ($'0,000,000,000)
156.70	171.66	328.36

Table 2. High frequency trading activity

The table presents the trading summary statistics for 120 randomly selected NASDAQ and NYSE-listed stocks traded for all dates in 2009. HH indicates a trade based on a HFT demanding liquidity and a HFT supplying liquidity. HN implies that a HFT demands liquidity and a non-HFT supplies liquidity, while NH is the opposite. NN refers to trades where both counterparties are non-HFTs. We compute HFT volume as the sum of HH, HN and NH.

Type						
HH (0,000,000s)	HN (0,000,000s)	NH (0,000,000s)	NN (0,000,000s)	HFT (0,000,000s)	Non-HFT (0,000,000s)	HFT %
Sell Side						
400.27	505.53	691.26	640.87	1597.07	640.87	71.3%
Buy Side						
398.04	512.48	689.20	642.37	1599.73	642.37	71.3%
Total						
798.32	1018.02	1380.46	1283.24	3196.8	1283.24	71.3%

Table 3. Summary statistics for variables

The table presents the descriptive statistics for variables of interest. *Espread* is the effective spread, computed as twice the absolute value of the difference between the last execution price for each interval and the midpoint of the prevailing bid and ask prices. *Rspread* is the relative spread, and is obtained by dividing the difference between the best ask and bid prices for each interval by the midpoint of both prices. *Qspread* is the quoted spread, and is simply the difference between the best ask and bid prices for each interval. *BSI* is the absolute difference between buyer- and seller-initiated traders, $|\Delta p|$ is absolute value of price change, *R* is the one-second midpoint return, σ^p is the standard deviation of mid-price returns, and *MT* is the proxy for market toxicity, calculated as the absolute value of the difference between the numbers of buy and sell trades divided by the sum of the numbers of buy and sell trades. One-second frequency is used for all variables, except *MT*. *MT* is computed by using one-minute frequency. The sample contains the most active 100 S&P 500 stocks traded between October 1, 2016 through to September 30, 2017 on NYSE and NASDAQ.

Variables	**Mean**	**Median**	**Standard deviation**
Espread	0.00906	0.01000	0.04625
Rspread	0.00039	0.00028	0.00090
Qspread	0.01863	0.01000	0.05640
BSI	1584.05	424.00	35771
$\lvert\Delta p\rvert$	0.00918	0.00900	0.06707
R	$-0.412\text{x}10^{-6}$	0.00	0.00139
σ^p	$0.92\text{x}10^{-4}$	$0.59\text{x}10^{-4}$	0.00091
MT	0.54067	0.50375	0.34194

Table 4. State Space Estimates

The table contains mean cross-sectional estimates of unexpected (information-driven) and expected (liquidity-driven) components of trading volume for the most active 100 S&P 500 stocks trading between October 1, 2016 and September 30, 2017. Stocks are divided into quartiles according to their level of trading activity; trading activity is based on trading volume. Quartile 1 contains the least active companies, while Quartile 4 contains the most active stocks. The estimates are based on the following state space model for decomposing trading volume:

$$v_{it} = m_{it} + s_{it} \; ; m_{it} = m_{it-1} + u_{it}$$

where $v_{it} = ln(TVolume_{it})$, $TVolume_{it}$ corresponds to trading volume of stock i at time t , $m_{it}$is a non-stationary expected component of stock *i* at time t, s_{it} is a stationary unexpected component for stock *i* at time *t* and u_{it} is an idiosyncratic disturbance error. σ^{2_s} and σ^{2_u} are the variance estimates of the unexpected and expected components of trading volume respectively, estimated by maximum likelihood (constructed using the Kalman filter). Estimations are presented for one-second, one-minute, and one-hour frequencies.

	Stock quartiles			
Variable	Least active	2	3	Most active
		One-second frequency		
σ^{2_s}	1.02	1.24	1.37	1.51
σ^{2_u}	0.46	0.49	0.53	0.78
		One-minute frequency		
σ^{2_s}	1.21	1.36	1.63	1.88
σ^{2_u}	0.49	0.55	0.72	0.85
		One-hour frequency		
σ^{2_s}	1.34	1.65	1.77	1.96
σ^{2_u}	0.51	0.59	0.76	0.97

Table 5. Correlation matrix for variables

The table plots the correlation matrix of the variables employed in this study's models. *Espread* is the effective spread, computed as twice the absolute value of the difference between the last execution price for each interval and the midpoint of the prevailing bid and ask prices. *Rspread* is the relative spread, and is obtained by dividing the difference between the best ask and bid prices for each interval by the midpoint of both prices. *Qspread* is the quoted spread, and is simply the difference between the best ask and bid prices for each interval. *TV* is the natural logarithm of trading volume, *BSI* is the absolute difference between buyer- and seller-initiated traders, $|\Delta p|$ is absolute value of price change, σ^p is the standard deviation of mid-price returns, and σ^{2s} and σ^{2u} are the state space model-estimated proxies for informed and liquidity trading volumes respectively. The sample contains the most active 100 S&P 500 stocks traded between October 1, 2016 and September 30, 2017 on NYSE and NASDAQ.

	$Qspread$	$Rspread$	$Espread$	TV	BSI	σ^{2s}	σ^{2u}	$\|\Delta p\|$	σ^p
$Qspread$	1								
$Rspread$	0.79909	1							
$Espread$	0.90722	0.72476	1						
TV	-0.04144	-0.06261	-0.01673	1					
BSI	0.00133	0.01265	0.00284	0.11090	1				
σ^{2s}	0.00000	0.00013	0.00007	0.00326	0.44342	1			
σ^{2u}	0.00000	-0.00001	0.00005	0.00021	-0.00001	-0.00000	1		
$\|\Delta p\|$	0.08621	0.05052	0.06789	0.01904	0.01101	0.00004	0.00008	1	
σ^p	0.12381	0.16384	0.11483	0.01700	0.01050	0.00004	0.00001	0.42911	1

Table 6. Predictive power of lagged expected and unexpected components of trading volume on market volatility

The predictive power of one-second expected and unexpected components of trading volume is estimated using the following models:

$$|\Delta p_{it}| = \alpha + \beta_1 Espread_{it-1} + \beta_2 TV_{it-1} + \beta_3 BSI_{it-1} + \beta_4 \sigma^{2_s}_{it-1} + \beta_5 \sigma^{2_u}_{it-1} + \varepsilon_{i,t}$$

$$\sigma^p_{it} = \alpha + \beta_1 \sigma^p_{it-1} + \beta_2 Espread_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5 \sigma^{2_s}_{it-1} + \beta_6 \sigma^{2_u}_{it-1} + \varepsilon_{i,t}$$

where $|\Delta p_{it}|$ is the absolute value of price change, $Espread_{it-1}$ is the effective spread, computed as twice the absolute value of the difference between the last execution price for each interval and the midpoint of the prevailing bid and ask prices. $\sigma^p_{it-1}$is the standard deviation of stock returns, TV_{it-1} is the natural logarithm of trading volume, BSI_{it-1} is the absolute difference between buyer- and seller-initiated traders, $\sigma^{2_s}_{it-1}$and $\sigma^{2_u}_{it-1}$ are the state space model-based proxies (estimated using Kalman filter constructed maximum likelihood) for informed and uninformed trading. The sample contains the most active 100 S&P 500 stocks traded between October 1, 2016 and September 30, 2017 on NYSE and NASDAQ. ***, ** and * correspond to statistical significance at the 0.01, 0.05 and 0.10 levels, respectively.

Panel A

	Dependent Variable: $\|\Delta p_{it}\|$		
	One-second frequency	One-minute frequency	One-hour frequency
$Intercept$	$0.847\text{x}10^{-2***}$ (681.52)	$0.198\text{x}10^{-1***}$ (138.29)	$0.110\text{x}10^{-1***}$ (26.05)
$Espread_{it-1}$	$0.742\text{x}10^{-1***}$ (280.32)	$0.457\text{x}10^{-1***}$ (76.42)	$0.287\text{x}10^{-1***}$ (15.34)
TV_{it-1}	$0.967\text{x}10^{-3***}$ (14.98)	$0.410\text{x}10^{-2***}$ (6.45)	$0.177\text{x}10^{-2***}$ (4.07)
BSI_{it-1}	$0.100\text{x}10^{-6***}$ (152.25)	$0.134\text{x}10^{-6***}$ (130.21)	$0.758\text{x}10^{-6***}$ (13.66)
$\sigma^{2_s}_{it-1}$	$-0.334\text{x}10^{-4***}$ (-12.89)	$-0.496\text{x}10^{-3***}$ (-7.95)	$-0.330\text{x}10^{-2***}$ (-4.87)
$\sigma^{2_u}_{it-1}$	$0.842\text{x}10^{-5}$ (0.15)	$-0.211\text{x}10^{-4}$ (-0.07)	$-0.863\text{x}10^{-4}$ (-0.03)
Sample size *(n)*	29959938	8880028	204354
Adjusted R²	0.40 %	0.86 %	3.17 %

Panel B

	Dependent Variable: σ^p_{it}		
	One-second frequency	One-minute frequency	One-hour frequency
$Intercept$	$0.741\text{x}10^{-4***}$ (426.41)	$0.740\text{x}10^{-4***}$ (256.55)	$0.789\text{x}10^{-4***}$ (14.75)
σ^p_{it-1}	$0.133\text{x}10^{-1***}$ (86.44)	$0.377\text{x}10^{-1***}$ (50.32)	0.6191^{***} (86.09)
$Espread_{it-1}$	$0.147\text{x}10^{-2***}$ (394.72)	$0.158\text{x}10^{-2***}$ (59.77)	$0.414\text{x}10^{-4}$ (1.50)
TV_{it-1}	$0.839\text{x}10^{-5***}$ (9.30)	$0.865\text{x}10^{-5***}$ (8.83)	$0.115\text{x}10^{-4***}$ (5.55)
BSI_{it-1}	$0.221\text{x}10^{-8***}$ (265.98)	$0.222\text{x}10^{-8***}$ (135.22)	$0.204\text{x}10^{-8***}$ (29.48)
$\sigma^{2_s}_{it-1}$	$-0.721\text{x}10^{-6***}$ (-19.92)	$-0.725\text{x}10^{-5***}$ (-16.76)	$-0.901\text{x}10^{-3***}$ (-12.71)
$\sigma^{2_u}_{it-1}$	$0.761\text{x}10^{-7}$ (0.01)	$0.687\text{x}10^{-6}$ (0.03)	$-0.575\text{x}10^{-3}$ (-0.01)
Sample size *(n)*	29959938	8880028	204354
Adjusted R²	0.92%	1.71%	5.27%

Table 7. Predictive power of lagged expected and unexpected components of trading volume on market liquidity

The predictive power of one-second expected and unexpected components of trading is estimated using the following model:

$$Spread_{it} = \alpha + \beta_1\sigma^{p}_{it-1} + \beta_2 TV_{it-1} + \beta_3 BSI_{it-1} + \beta_4\sigma^{2s}_{it-1} + \beta_5\sigma^{2u}_{it-1} + \varepsilon_{i,t}$$

where $Spread_{it}$ corresponds to one of effective, quoted and relative spreads respectively. Effective spread is computed as twice the absolute value of the difference between the last execution price for each interval and the midpoint of the prevailing bid and ask prices. Relative spread is obtained by dividing the difference between the best ask and bid prices for each interval by the midpoint of both prices. Quoted spread is simply the difference between the best ask and bid prices for each interval. σ^{p}_{it-1} is the standard deviation of stock returns, TV_{it-1} is the natural logarithm of trading volume, *BSI* is the absolute difference between buyer- and seller-initiated transactions, and $\sigma^{2s}_{it-1}$and σ^{2u}_{it-1} are the state space model-based proxies (estimated using Kalman filter constructed maximum likelihood) for informed and uninformed trading. The sample contains the most active 100 S&P 500 stocks traded between October 1, 2016 and September 30, 2017 on NYSE and NASDAQ. ***, ** and * correspond to statistical significance at the 0.01, 0.05 and 0.10 levels, respectively.

Panel A

	Dependent Variable: $RSpread_{it}$		
	One-second frequency	One-minute frequency	One-hour frequency
$Intercept$	0.385x10^{-3***}	0.435x10^{-3***}	0.497x10^{-3***}
	(240.73)	(181.54)	(47.93)
σ^{p}_{it-1}	0.608x10^{-1***}	0.199x10^{-3***}	0.203x10^{-1***}
	(241.98)	(57.72)	(71.57)
TV_{it-1}	0.677x10^{-5}	0.927x10^{-4}	-0.492x10^{-4}
	(0.80)	(1.41)	(-0.12)
BSI_{it-1}	0.252x10^{-8***}	0.246x10^{-8***}	0.825x10^{-8***}
	(387.31)	(231.72)	(61.22)
σ^{2s}_{it-1}	-0.902x10^{-5***}	-0.979x10^{-4***}	-0.349x10^{-4***}
	(-26.48)	(-15.25)	(-12.23)
σ^{2u}_{it-1}	-0.352x10^{-6}	-0.258x10^{-4}	-0.615x10^{-4}
	(-0.05)	(-0.08)	(-0.08)
Sample size *(n)*	29959938	8880028	204354
Adjusted R^2	1.09%	1.45%	14.01%

Panel B

	Dependent Variable: $QSpread_{it}$		
	One-second frequency	One-minute frequency	One-hour frequency
$Intercept$	0.182x10^{-1***}	0.179x10^{-1***}	0.230x10^{-1***}
	(181.80)	(93.45)	(31.74)
σ^{p}_{it-1}	2.767***	2.24***	128***
	(230.62)	(101.94)	(55.54)
TV_{it-1}	-0.918x10^{-3*}	-0.145x10^{-3}	-0.385x10^{-2}
	(-1.74)	(-0.35)	(-1.37)
BSI_{it-1}	0.921x10^{-7***}	0.968x10^{-7***}	0.352x10^{-7***}
	(182.69)	(143.04)	(37.42)
σ^{2s}_{it-1}	-0.329x10^{-4***}	-0.386x10^{-4***}	-0.155x10^{-3***}
	(-15.50)	(-9.44)	(-6.22)
σ^{2u}_{it-1}	0.116x10^{-6}	0.310x10^{-6}	-0.284x10^{-5}
	(0.03)	(0.02)	(-0.05)
Sample size *(n)*	29959938	8880028	204354
Adjusted R^2	0.49%	1.08%	11.15%

Panel C

	Dependent Variable: $ESpread_{it}$		
	One-second frequency	One-minute frequency	One-hour frequency
$Intercept$	0.874x10^{-2***}	0.882x10^{-2***}	0.981x10^{-2***}

	(107.46)	(67.52)	(15.54)
σ_{it-1}^{p}	2.009***	13.34***	107.66***
	(127.41)	(71.03)	(14.92)
TV_{it-1}	-0.160x10^{-3***}	-0.102x10^{-3}	-0.197x10^{-2}
	(-3.74)	(-0.28)	(-0.80)
BSI_{it-1}	0.604x10^{-7***}	0.634x10^{-7***}	0.240x10^{-6***}
	(118.35)	(89.80)	(29.34)
$\sigma_{it-1}^{2_s}$	-0.216x10^{-4***}	-0.254x10^{-4***}	-0.107x10^{-3***}
	(-12.55)	(-11.23)	(-12.80)
$\sigma_{it-1}^{2_u}$	-0.758x10^{-4}	-0.186x10^{-4}	-0.208x10^{-4}
	(-0.20)	(-0.11)	(-0.04)
Sample size *(n)*	29959938	8880028	204354
Adjusted R^2	0.37%	1.09%	10.18%

Table 8. Predictive power of lagged expected and unexpected components of trading volume on market toxicity

The predictive power of one-minute expected and unexpected components of trading volume is estimated using the following model:

$$MT_{it} = \alpha + \beta_1 Espread_{it-1} + \beta_2 TV_{it-1} + \beta_3 BSI_{it-1} + \beta_4 \sigma_{it-1}^{2s} + \beta_5 \sigma_{it-1}^{2u} + \varepsilon_{i,t}$$

where MT_{it} is a proxy for market toxicity, which is computed as the absolute value of the difference between the numbers of buy and sell trades over a one-minute interval, divided by the total number of trades for that interval. $Espread_{it-1}$ is the effective spread, computed as twice the absolute value of the difference between the last execution price for each interval and the midpoint of the prevailing bid and ask prices. TV_{it-1} is the natural logarithm of trading volume, BSI_{it-1} is the absolute difference between buyer- and seller-initiated transactions, and σ_{it-1}^{2s} and σ_{it-1}^{2u} are the state space model-based proxies (estimated using Kalman filter constructed maximum likelihood) for informed and uninformed trading. The sample contains the most active 100 S&P 500 stocks traded between October 1, 2016 and September 30, 2017 on NYSE and NASDAQ. ***, ** and * correspond to statistical significance at the 0.01, 0.05 and 0.10 levels, respectively.

	Dependent Variable: MT_{it}	
	One-minute frequency	One-hour frequency
$Intercept$	0.539^{***} (465.44)	0.598^{***} (125.92)
$ESpread_{it-1}$	$0.767 \text{x} 10^{-1***}$ (57.21)	$0.821 \text{x} 10^{-1***}$ (40.30)
TV_{it-1}	$0.128 \text{x} 10^{-3***}$ (8.96)	$0.440 \text{x} 10^{-2}$ (1.61)
BSI_{it-1}	$0.153 \text{x} 10^{-6***}$ (66.13)	$0.354 \text{x} 10^{-6***}$ (65.56)
σ_{it-1}^{2s}	$-0.578 \text{x} 10^{-2***}$ (-4.14)	$-0.234 \text{x} 10^{-2***}$ (-22.76)
σ_{it-1}^{2u}	$-0.396 \text{x} 10^{-3}$ (-0.57)	$-0.670 \text{x} 10^{-3}$ (-0.29)
Sample size *(n)*	8880028	204354
Adjusted R^2	0.12%	2.84%

Table 9. Predictive power of lagged unexpected component of trading volume on short horizon stock returns

The predictive power of one-second/minute expected and unexpected components of trading volume is estimated using the following model:

$$R_{it} = \alpha + \beta_1 \sigma_{it-1}^{p} + \beta_2 Espread_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5 \sigma_{it-1}^{2_s} + \beta_6 MT_{it-1} + \varepsilon_{i,t}$$

where $R_{it}$is the midpoint one-minute return, $\sigma_{it-1}^{p}$is the standard deviation of stock returns, $Espread_{it-1}$ is the effective spread, computed as twice the absolute value of the difference between the last execution price for each interval and the midpoint of the prevailing bid and ask prices. TV_{it-1} is the natural logarithm of trading volume, BSI_{it-1} is the absolute difference between buyer- and seller-initiated trades. MT_{it-1} is a proxy for market toxicity, which is computed as the absolute value of the difference between the numbers of buy and sell trades over a one-minute interval, divided by the total number of trades for that interval, and $\sigma_{it-1}^{2_s}$ is the state space model-based proxy (estimated using Kalman filter constructed maximum likelihood) for informed trading. The sample contains the most active 100 S&P 500 stocks traded between October 1, 2016 and September 30, 2017 on NYSE and NASDAQ. ***, ** and * correspond to statistical significance at the 0.01, 0.05 and 0.10 levels, respectively.

	Dependent Variable: R_{it}	
	One-second frequency	One-minute frequency
$Intercept$	-0.535x10^{-5***} (-20.14)	-0.774x10^{-4***} (-8.19)
σ_{it-1}^{p}	0.524x10^{-3**} (2.22)	-0.106x10^{-4} (-1.30)
$ESpread_{it-1}$	0.440x10^{-3***} (77.59)	0.871x10^{-3***} (59.59)
TV_{it-1}	0.108x10^{-6***} (7.89)	0.838x10^{-5***} (7.75)
BSI_{it-1}	0.540x10^{-9***} (51.12)	0.123x10^{-8***} (47.22)
$\sigma_{it-1}^{2_s}$	-0.153x10^{-6***} (-27.71)	-0.417x10^{-4} (-1.52)
MT_{it}		0.265x10^{-5} (0.69)
Sample size *(n)*	29959938	8880028
Adjusted R^2	0.06%	0.09%

Table 10. Predictive power of lagged unexpected component of trading volume and lagged unexpected component of trading volume interacted with a dummy variable for high frequency trading on short horizon stock returns

The predictive power of one-second/minute expected and unexpected components of trading volume, as well as unexpected component, interacted with a dummy variable for HFT is estimated using the following model:

$$R_{it} = \alpha + \beta_1 \sigma^p_{it-1} + \beta_2 Illiq_{it-1} + \beta_3 TV_{it-1} + \beta_4 BSI_{it-1} + \beta_5 \sigma^{2_s}_{it-1} + \beta_6 MT_{it-1} + \beta_7 D_{HFT,t-1} * \sigma^{2_s}_{it-1} + \varepsilon_{i,t}$$

where $R_{it}$is the midpoint one-minute return, $\sigma^p_{it-1}$is the standard deviation of stock returns, $Illiq_{it-1}$ is the Amihud illiquidity proxy in Panel A and the effective spread in Panel B. The Amihud illiquidity proxy is computed as absolute return divided by trading volume for each interval and the effective spread is computed as twice the absolute value of the difference between the last execution price for each interval and the midpoint of the prevailing bid and ask prices. TV_{it-1} is the natural logarithm of trading volume, BSI_{it-1} is the absolute difference between buyer- and seller-initiated trades. MT_{it-1} is a proxy for market toxicity, which is computed as the absolute value of the difference between the numbers of buy and sell trades over a one-minute interval, divided by the total number of trades for that interval, and $\sigma^{2_s}_{it-1}$ is the state space model-based proxy (estimated using Kalman filter constructed maximum likelihood) for informed trading. D_{HFT} is a dummy equalling one during periods of high HFT activity. A one-second or one-minute interval is designated as an interval of high HFT activity if HFT trades for that interval is one standard deviation higher than the mean for the surrounding -60, +60 corresponding intervals. The sample for Panel A contains 120 NASDAQ and NYSE stocks traded for all dates in 2009. The sample for Panel B contains the most active 100 S&P 500 stocks traded between October 1, 2016 and September 30, 2017 on NYSE and NASDAQ. ***, ** and * correspond to statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A

	Dependent Variable: R_{it}	
	One-second frequency	One-minute frequency
$Intercept$	-0.311x10^{-3***} (-6.76)	0.423x10^{-2**} (2.06)
σ^p_{it-1}	0.630*** (8.86)	0.049x10^{-1} (1.64)
$Amihud_{it-1}$	-3.668*** (-4.60)	-1.405 (-0.25)
TV_{it-1}	0.240x10^{-4***} (3.96)	0.385x10^{-3**} (2.52)
BSI_{it-1}	0.01x10^{-9*} (1.74)	0.01x10^{-6*} (1.85)
$\sigma^{2_s}_{it-1}$	-0.097x10^{-6} (-1.58)	0.556x10^{-4} (1.49)
$D_{HFT,t-1} * \sigma^{2_s}_{it-1}$	-0.346x10^{-4***} (-3.38)	-0.170x10^{-3} (-1.51)
MT_{it-1}		-0.153x10^{-2} (-1.32)
Adjusted R^2	0.09%	0.25%

Panel B

	Dependent Variable: R_{it}	
	One-second frequency	One-minute frequency
$Intercept$	-0.258x10^{-5***} (-3.19)	-0.107x10^{-4*} (-1.85)

σ^{p}_{it-1}	0.242x10^{-2**} (2.45)	-0.650x10^{-4} (-1.02)
$ESpread_{it-1}$	0.287x10^{-3***} (17.44)	0.566x10^{-3***} (2.99)
TV_{it-1}	0.234x10^{-6**} (2.07)	0.217x10^{-5***} (11.17)
BSI_{it-1}	0.264x10^{-10} (0.87)	0.492x10^{-8***} (13.49)
σ^{2s}_{it-1}	-0.284x10^{-8} (-0.01)	-0.267x10^{-7} (-1.08)
$D_{HFT,t-1} * \sigma^{2s}_{it-1}$	-0.264x10^{-5***} (-3.51)	-0.128x10^{-10} (-1.45)
MT_{it-1}		0.114x10^{-5} (0.29)
Adjusted R^{2}	0.04%	0.11%

Table 11. The behaviour of unexpected component around earnings announcements

The table displays the cross-sectional mean and statistical tests of differences of the unexpected component and trading volume between the periods of the before and after earnings announcements. The statistical tests conducted are two-sample t-tests and pairwise Wilcoxon-Mann-Whitney U tests. Unexpected component is the state space model-based proxy (estimated using Kalman filter constructed maximum likelihood) for informed trading and trading volume is the natural logarithm of trading volume. The sample contains the most active 100 S&P 500 stocks traded between October 1, 2016 and September 30, 2017 on NYSE and NASDAQ. ***, ** and * correspond to statistical significance at the 0.01, 0.05, and 0.10 levels, respectively

	Variables	
Periods	Unexpected component	Trading Volume
Before earnings announcements	1.2538	6.5194
After earnings announcements	1.2315	6.5363
Difference	0.0222	-0.0169
p value for t-test	<0.001***	0.0123**
P value for Wilcoxon-Mann-Whitney U	<0.001***	0.0077***