The Effect of NYSE American's Latency Delay on Informed Trading

By
Jeremy Morris
B.A. University of Victoria 2020

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF ARTS
in the Department of Economics

Jeremy Morris, 2023
University of Victoria

I acknowledge and respect the lək̓ʷəŋən peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

The Effect of NYSE American's Latency Delay on Informed Trading

By

Jeremy Morris

B.A. University of Victoria 2020

Supervisory Committee

Dr. Ke Xu, Supervisor

Department of Economics

Dr. Kenneth Stewart, Supervisory Committee

Department of Economics

Abstract

Insider trading plays a significant role in financial markets and how markets respond to new information. Informed high-frequency traders pose a major risk to liquidity providers in financial markets due to adverse selection, which can result in market failure. To mitigate this risk, some exchanges have implemented speed bumps which delay trades. Using trade and quote (TAQ) data of 45 stocks on the NYSE American and the NASDAQ from May 2017 to August 2017, I identify the impact of a trading delay of 350 microseconds on the probability of informed trading for the NYSE American using difference-in-differences estimation. I find a statistically significant decline in the probability of informed trading after the implementation of the speed bump on the NYSE American stock exchange.

Contents

Introduction

Asymmetric information is a common feature of financial markets. In equities markets the transaction price of a firm's stock is determined by the discounted future profits of the firm. Since the stock's transaction price reflects the forecast of the firm's profitability, any information on the future profitability of a firm will have an impact on the transaction price today. Therefore, any market participant with information on the future prospects of the firm can benefit from trading on that information. Due to the competitiveness and speed of financial markets, this pricing occurs quickly. Other participants, the market makers, face risk when trading with these informed traders since they consistently lose value on these trades. In effect, market makers are exchanging assets at inaccurate prices, either over-evaluating or under-evaluating the asset. This poses a risk to market liquidity if market makers leave the market when a large amount of informed trading occurs, or when they suspect they are trading with informed traders (Easley & O'Hara, 1992; Easley et al., 2012; Grossman & Stiglitz, 1980; Kyle, 1985). This can result in events such as the Flash Crash, where a liquidity crisis caused a financial panic for thirty minutes in the E-mini S&P 500 futures market (Easley et al., 2012).

In this environment, there are three major categories of traders: informed, uninformed, and market makers (Grossman & Stiglitz, 1980; Kyle, 1985). Market makers are traders who take on the opposing side of an initiated trade, taking the bid-ask spread as compensation. Informed traders and uninformed traders, distinguished by their degree of information about financial markets, initiate trades with market makers. For market makers, informed traders are risky to trade with since informed trades will come up favourably to the informed traders at the expense of market makers

(Kyle, 1985). Market makers are unable to identify which traders have portfolios that are consistent with insider trading (Grossman & Stiglitz, 1980; Kyle, 1985). Therefore, when market makers suspect that there are too many informed traders in the market, they will exit the market, creating a liquidity crisis (Easley & O'Hara, 1992). In a speech at the 11th Annual SIFMA Market Structure Conference Berman (2010) discussed the role of high-frequency trading firms in the Flash Crash:

> It would therefore seem, at least on May 6, that HFTs took as much liquidity from the equity markets as they provided. Furthermore, data show that in the 15 minutes leading up [sic] the Flash Crash at 2:45, HFT activity increased proportionately with the market, as HFT algorithms transitioned from being aggressively neutral to being net aggressive sellers. This transition was roughly in line with the market as a whole, though HFT activity after 2:45 dropped off considerably. So while it does not seem that HFTs directly caused a wave of selling, HFTs did ride that wave down as prices declined.

One of the more common types of informed traders is high-frequency traders, and it is this group that will be affected by a latency delay or speed bump. By slowing down trades, these high-frequency traders are not able to implement their informed trading strategies as effectively (Aldauf & Mollner, 2020). By implementing a speed bump market makers are, in theory, less exposed to informed trading making low-frequency traders more secure (Brolley & Cimon, 2020). Following this rationale some equity markets including the IEX, NYSE American, CHX, and TX Alpha have implemented speed bumps (Chen et al., 2017). The NYSE American's implementation of a latency delay is the subject of this paper and the findings of this thesis have implications for all exchanges with latency delays.

If latency delays can reduce market uncertainty and lower the toxicity of the order flow, flash crashes could be less likely to occur. This provides the motivation to study the effect of the NSYE American's speed bump on order flow toxicity. In addition, order flow toxicity impacts the cost of trading and volatility providing further motivation for the study of order flow toxicity in financial markets (Yildiz et al., 2020). Small changes in trading costs and realized volatility due to order

flow toxicity have significant economic effects due to the large volume of trading on financial markets. In 2017 NYSE American implemented one such speed bump of 350 microseconds, or $^{7}/_{20,000}$ of a second, to all trades. This delay is small, but in the world of high-frequency trading algorithms, represents a substantial delay. Latency created by trades travelling between Chicago and New York City financial markets was 4 ms or $^{1}/_{250}$ of a second (Kuhle, 2023; Laughlin et al., 2014). For the sake of comparison, the speed bump at the NYSE American was roughly 9% of the time it takes for a trade to travel from Chicago to New Jersey, where the New York metropolitan area based exchanges are located. For low-frequency traders such a delay is inconsequential. Yet, even this very small delay can stop New York-based high-frequency traders from front-running trades from other New York-based traders. Jones (2018) found that NASDAQ listed securities on the NYSE had an average latency of 17 microseconds. For New York based traders the NYSE American's delay of 350 microseconds is therefore quite substantial. NYSE American would on November 18, 2019, ended the use of the speed bump, citing a larger bid-ask-spread, the difference between the selling and buying prices, and declining market share. This represents a trade-off for exchanges between market efficiency and averting informed trading (Liu & Xu, 2023).

Collecting data on informed trading is difficult. Companies and traders have to file their holdings with the Security Exchange Commission but this isn't done in real-time and therefore is not useful for measuring intraday activity. Instead, using the volume-synchronized probability of informed trading (VPIN), created by Easley et al. (2012), I can estimate the probability of informed trading for a given stock using the transaction price and volume intraday data. This technique has been used to measure liquidity-induced volatility in many markets, especially in the area of flash crashes (Easley et al., 2012; Kitamura, 2017).

Previous work on latency delays has shown clear implications for my analysis, with theoretical results suggesting that a latency delay will lower informed trading. Aldauf and Mollner (2020) provides the underlying mechanism for why a latency delay would lower the number of high-frequency traders. They suggest that high-frequency traders can exploit changes in the midpoint of the bid/ask spread by trading with low-frequency traders before they can update their orders.

Brolley and Cimon (2020) find that informed traders will migrate away from exchanges with speed bumps to conventional exchanges. Cartea and Penalva (2013) find that high-frequency trading increases order imbalances. Both of these results suggest that the VPIN metric should fall after the NYSE American implemented a speed bump. In contrast, Aoyagi (2019) suggests that the impact of a speed bump's implementation on informed trading depends on the initial number of informed traders on the given exchange. Chakrabarty et al. (2020) found that market quality improved on the IEX due to the latency delay on the exchange. In a study of the Taiwan Futures Exchange Chang and Chou (2022) find that high-frequency traders provide liquidity during stable market conditions but become liquidity demanders during turbulent market conditions. In a working paper, Liu and Xu (2023) find that the NYSE American's speed bump reduced informed trading and lowered price discovery. My analysis offers support to this result using an alternative methodology.

Using three months of financial market data I select 45 of the most frequently traded stocks on the NYSE American for the entire sample length as my treatment group. The speed bump was implemented in the middle of the sample period such that there is sufficient data for the pre-treatment and post-treatment periods to estimate the effect of the speed bump. Using those same 45 stocks I create a control group using data from the NASDAQ. This three-month window represents a large data set due to the high volume of trades that are completed every day on these exchanges. I measure the effect of the implementation of the speed bump on the probability of informed trading using a difference-in-differences quasi-experimental identification strategy. My findings suggest a statistically significant decline in the probability of informed trading after the implementation of the speed bump. The 3.6% decline in informed trading due to the latency delay is a substantial effect given that the latency delay is 350 microseconds. My results suggest that such speed bumps are effective in deterring informed trading in equities markets. As such latency delays can be used in equity markets to lower order flow toxicity.

Literature Review

Public information is incorporated into the transaction price of a stock such that the market reaches equilibrium, quickly eliminating arbitrage opportunities. Such a situation is described as the semi-strong efficient market hypothesis since prices reflect all public information. If a party has private information that has not yet been incorporated into the equilibrium price they can still benefit by trading even at the equilibrium price since they have information which hasn't been incorporated into the stock's price. By doing so the equilibrium price of the stock will reflect all information (Fama, 1970). There are three major types of traders: informed, noise, and market makers which define the model used in this analysis. In the Kyle model, risk-neutral informed traders create market imbalances which can be detected by market makers who adjust prices upward in the presence of increased demand. Informed traders then must optimize their trading strategy such that they maximize their profits by holding additional shares while also increasing the transaction price through their trading activity (Kyle, 1985). In the Grossman-Stiglitz model, the informed trader is risk-averse and submits a menu of prices and trade quantities, such that their marginal utility of owning additional shares is offset by the additional variances of their portfolio (Grossman & Stiglitz, 1980). Building on the Kyle model, Easley and O'Hara (1992) use maximum likelihood estimation to estimate the probability of informed trading (PIN). In their specification, information arrives daily and is incorporated into prices by the end of each trading day. This model detects informed trading using order imbalances caused by informed trading. Easley et al. (2002) find that PIN is positively related to asset returns. This insight is the foundation for my analysis. I use an updated form of this methodology found in Easley et al. (2012). In this new methodology

the volume-synchronized probability of informed trading is estimated using the order imbalance for each minute of each trading day in the sample. Abada and Yagüe (2012), uses data from the Spanish Stock Exchange of 15 Spanish companies data and found that VPIN proxies for adverse selection risk in a comparable way to PIN when the VPIN parameters are chosen to mimic the original implementation of PIN estimators. As such, Abada and Yagüe (2012) find that VPIN, despite a very different implementation, can effectively mimic PIN estimates given parameters that are consistent with PIN. This result indicates that the theoretical link between PIN and VPIN is empirically consistent.

Using the VPIN methodology Easley et al. (2012) found that VPIN is positively associated with periods of market instability and uncertainty. Using the sample between January 1, 2008, and August 1, 2011, they found that the VPIN estimates for the E-mini S&P 500 reached its peak during the Flash Crash on May 6, 2010. However, Andersen and Bondarenko (2014b) dispute whether VPIN was predictive of the Flash Crash. They find that the VPIN metric rose prior to the Flash Crash and continued to rise afterwards providing an ex-post measurement rather than a predictive one. They also find that VPIN is highly correlated with trading intensity and volatility but isn't strongly predictive of future volatility and is sensitive to the classification technique used to classify trades.

In response to Andersen and Bondarenko (2014b), Easley et al. (2014) show that the cumulative distribution function of VPIN was above 99% prior to the Flash Crash, showing that the VPIN estimates were at unusually high levels before, and after, the Flash Crash. In addition, Easley et al. (2014) also states that VPIN, by construction, should rise during a period of intense trade imbalances, such as the Flash Crash, and that VPIN isn't designed to predict future volatility. Instead, VPIN is measuring order flow toxicity, which can have a short-run impact on volatility as market makers leave the market. Andersen and Bondarenko (2014a) counter this response, suggesting that a CDF of 99% isn't high enough, suggesting instead the use of a daily maximum. Using such a maximum they note that VPIN was high before the Flash Crash. Citing Chakrabarty et al. (2015) they found that bulk classification underperforms the Lee and Ready algorithm developed by Lee

and Ready (1991). Andersen and Bondarenko (2014a) also suggest that controlling for realized volatility is necessary for measuring the predictive power of VPIN on future volatility. Easley et al. (2014) agree, stating that VPIN captures information in volume, not price, and measures the impact of toxicity-related volume information on volatility.

Easley et al. (2012) found that the S&P 500 VPIN rose during the Fukushima nuclear disaster in the wake of the Tōhoku earthquake. A major difference from the Flash Crash was that the market reacted during the night session, from 6 p.m. to 11 p.m. EST. This suggests that informed trading was occurring in periods of low trade intensity. In addition, they found that VPIN is resilient to "fat finger error" price movements, when human or computer error results in a sudden rise or fall in prices which quickly return to their previous values, but which aren't related to informed trading. Easley et al. (2012) show one such case on June 8, 2011, in the New York Mercantile Exchange natural gas market. The VPIN metric wasn't high before this mistake, which is expected since in this case the increase in volatility was due to error and not order flow toxicity. In the forex market, Kitamura (2017) found that VPIN was able to predict the August 24 2015 ¥/$ forex flash crash, when the exchange plummeted from ¥119.15 to ¥116.15 within four minutes before rising to ¥118.9 at closing. In the energy futures market, Bjursell et al. (2017) find evidence that VPIN estimates increased significantly before the inventory announcements period and peaked afterwards.

Geopolitical instability can also impact financial markets and lead to increases in informed trading. This large-scale military operation provided an effective area for insider trading due to the secrecy involved and the possibility of economic sanctions on Russian companies in the event of war. Using the Russian RTS equity index and 161 individual Russian equities Silva and Volkova (2018) find that VPIN spiked in the days prior to the Russian invasion of Crimea. The authors found that the maximum CDF of the previous day's VPIN predicted a decline in the next day's return. They also found that VPIN was more predictive of future returns for Russian firms with government ownership between 0% and 90%, suggesting that firms with moderate government ownership were more influenced by informed trading, with the high and low government ownership interaction

terms being insignificant. The authors suggest that firms with over 90% ownership might not be affected by sanctions, since those firms would, or would be expected to, be the main targets of government support.

While VPIN's relationship with liquidity in short-run liquidity crises is investigated quite thoroughly in the above papers, its long-run relationship with liquidity outside of crises is investigated by Yildiz et al. (2020). Even outside of these rapid drops, liquidity is a major concern for traders on a regular basis. Yildiz et al. (2020) use the relative quoted and effective spreads to measure market liquidity. To measure revenue to liquidity providers they use the relative realized spread and to measure the gross loss to liquidity demanders due to adverse selection they use the price impact of the trade. In addition, Yildiz et al. (2020) investigate VPIN in the equity market, rather than in futures markets and due to structural differences between the two markets the relationship between VPIN, liquidity, and volatility will be different. The authors suggest that this is partially due to index futures having lower adverse selection costs, lower execution costs, and lower short selling costs. This suggests that the relationships found in futures markets may differ compared to those found in this analysis. Yildiz et al. (2020) suggest that these additional costs will affect the arrival of information. Yildiz et al. (2020) using S&P 500 data from 2015 found that changes in VPIN influence the relative effective spread, relative realized spread, and the adverse selection component of spread in a given volume bucket. Yildiz et al. (2020) also find VPIN is positively related to market volatility. They also find that low-volume stocks have higher VPIN values in line with the results found in Easley et al. (1996). Following this approach I use the relative quoted spread, relative effective spread, relative realized spread, and the adverse selection component of the spread as controls in my analysis. The authors also find that trade size is positively correlated with VPIN, in line with Easley and O'Hara (1992) who found that trading size is related to adverse selection. The authors find that order imbalances, sell volume, and buy volume have the highest correlations with VPIN. As such I include these metrics as control variables in my analysis along with the bid and offer volumes.

The largest empirical test of VPIN's ability to predict liquidity-based volatility of one hundred

highly liquid futures contracts over a period of five and a half years amounting to over 3 billion trades was done by Wu et al. (2013). Using a US National Laboratories supercomputer, Wu et al. (2013) were able to test whether VPIN predicts volatility, based on the Maximum Intermediate Return (MIR), an instantaneous measurement of volatility. Wu et al. (2013) use the 99% CDF of an asset's VPIN metric to define periods of high order flow toxicity, or a VPIN event, and then measure the following values of the MIR. They then use the Kolmogorov-Smirnov test to compare the resulting MIR sequences to a random event series to see if they are of the same probability distribution. The authors find that the VPIN events are extremely different from a series of random events, having different distributions. The authors test 1,600 possible parameter values for VPIN and MIR. The values the authors consider are: the nominal price of a bar π, the buckets per day β, the support window σ, the event duration η, the parameter for BVC ν, and the threshold for CDF of VPIN τ. Of the combinations that were tested, the best-performing combinations had a false positive rate of 7%. The authors find that VPIN performs well, being a strong predictor of liquidity-induced volatility in this very large and representative data set.

In order to understand how the speed bump on the NYSE American might affect the VPIN estimates, we should review the literature on the effects of speed bumps. Aldauf and Mollner (2020) create a theoretical model of uniform delays found on the IEX and NYSE American featuring pegged orders which are market orders that have been queued behind visible orders, and so are invisible, and older pegged orders at the midpoint between the best offer and best bid. One way that high-frequency traders can benefit is by sniping old peg orders when the midpoint changes. A latency delay allows the exchange's matching engine to view changes in the National Best Bid and Offer database, the aggregated public price quotes, and reprice pegged orders before new orders from fast traders arrive on the market. This protects low-frequency traders from high-frequency traders sniping their orders. Aldauf and Mollner (2020) found that a lower ratio of high-frequency traders compared to market makers results in prices that deviate less from fundamental value, albeit with higher costs of queuing. This is the mechanism which explains why high-frequency traders would be deterred from using an exchange with a speed bump.

Brolley and Cimon (2020) found that the implementation of a speed bump will drive informed traders away from exchanges with the increased latency and that uninformed traders will flock to the delayed exchange. This directly implies that the VPIN metric should decline as the numerator, the number of informed trades will decrease as a proportion of overall trades. The effect on the overall market might be less positive. In Brolley and Cimon (2020) the effect of the speed bump is to move uninformed trading to the exchange with the speed bump improving liquidity on that exchange. This results in a decline in liquidity on the exchange without a speed bump. Aoyagi (2019) found a mixed effect on price discovery, with the result depending on the original quantity of informed traders on the exchange before the speed bump is introduced. When there is an initially high level of informed trading, after the implementation of a speed bump Aoyagi (2019) predicts lower price discovery, however, when there are few initial informed traders the implementation of a speed bump may improve price discovery. The tactic of fading, whereby market makers can exploit the increase in latency by changing their prices before trades so that the trades are executed at a price worse for the trader and better for the market maker is a concern on exchanges with speed bumps (Aoyagi, 2019). A distinction is that Aoyagi (2019) uses a model with a random latency rather than the uniform one that was implemented on the NYSE American which is the subject of my analysis. According to Aoyagi (2019) the effect of the speed bump is a decline in the number of informed traders on the exchange. This would be consistent with a negative effect of the speed bump on the volume-synchronized probability of informed trading.

The role of stale orders in high-frequency trading is also a result found in Hoffmann (2014), expanding on the Foucault (1999) game-theoretic model with the addition of traders having differing speeds. Since limit orders are fixed over time, unlike market orders, they can become stale when new information arrives on the market. Foucault (1999) creates a model that explains the mix of limit orders and market orders used in financial markets and finds that volatility is the main driver between the choice of limit and market orders, the proportion of limit orders being positively related to volatility and the size of the spread. Hoffmann (2014) extends this framework with fast traders being able to adjust their limit orders if the next trader is a slow trader but not if the next

trader is a fast trader. This lowers the inefficiency compared to the model in Foucault (1999) since limit orders are less likely to be picked off. Slow traders respond by making limit orders that are unlikely to be executed. With a speed bump the risk of adverse selection is decreased for fast traders but increases for slow traders.

Cartea and Penalva (2013) provide a model in which high-frequency traders act as an intermediary between market makers and liquidity traders and find that high-frequency traders exacerbate the price impact of liquidity trades and generate a temporary order imbalance, which means that the VPIN estimate, based on the order imbalance, would rise. Hu (2018) studies the effect of a speed bump on the IEX and finds that liquidity measures improve after the introduction of the speed bump. Chakrabarty et al. (2020) found that back-running strategies declined on the IEX and that market quality improved. Chen et al. (2017) found that asymmetric and randomized speed bump implemented on the TX Alpha decreased traded volume and led to an increase in the quoted spread.

Using VPIN to measure the impact of a latency delay on informed trading has not been done yet in the literature. However, the literature strongly indicates that the effect of such a delay on the probability of informed trading would be negative. In this analysis I find that this is indeed the case for the NYSE American. I find that there is a decline in the probability of informed trading after the latency delay was implemented on the NYSE American.

Data Description

For my analysis I use a three-month subset, June to August 2017, of the NYSE Exchange Proprietary Market Daily Trade and Quote (TAQ) database. Since the NYSE American stock exchange implemented the speed bump of 350 microseconds on July 24, 2017, I have 37 trading days of data prior to the speed bump's implementation and 29 trading days after its implementation. The TAQ data set contains intraday transaction data for all North American exchanges. Each trade completed in this three-month period on every exchange is recorded. This provides high-frequency data for this three-month period at the nanosecond frequency. Due to this high frequency and the large volumes of financial trades that are completed each day this three-month period represents a very large data set. The trade data set records all trades in North American exchanges with trade price, trade volume, the exchange the trade was traded on, the time the trade was completed, and trade indicators for each individual trade. The National Best Bid and Offer (NBBO) records information on the highest bid, lowest ask, the sizes of the bid and ask, and the recorded time.

Table: 1

NYSE American lists primarily small and medium-cap stocks. Prior to the implementation of the speed bump, NYSE American had only 240 stocks listed on the exchange, however, after the implementation of the speed there was a substantial increase in the number of listed companies, with the NSYE American listing over 4,500 companies. Furthermore, after the implementation of the speed bump, there is a large increase in trading volume (Liu & Xu, 2023). This indicates a substantial interest on the part of market participants for time-delayed trading.

The sample excludes stocks that were not traded on the NYSE American exchange prior to the implementation of the speed bump on July 24th. Of these, I select the 45 most actively traded stocks by volume on the NYSE American that are also listed on the NASDAQ. For each of these 45 stocks, I select all trades and quotes data from both the NASDAQ and the NYSE American exchanges.

To clean the data I follow the methods from Hendershott and Moulton (2011). I filter out all trades with the following criteria: non-positive price or volume, trades with special trade conditions, and trades that are 150% larger or 50% smaller than the previously listed trade. For the quote data set I filter out any quotes with non-positive prices or volumes, bid prices greater than the asking price, where the quote is greater than 25% of the quote midpoint, or when the asking price is more than 150% of the bid price.

I apply the Lee and Ready algorithm developed by Lee and Ready (1991) to estimate whether a trade was buyer or seller-initiated. The Lee and Ready algorithm determines this by using the midpoint, the average of the bid and ask prices, and compares each trade's price to the midpoint. When the trade price is higher than the midpoint it is classified as a buyer-initiated trade. If the trade price is below the midpoint it is considered a seller-initiated trade. When a trade price is equal to the midpoint, the trade is classified by the tick rule, comparing the current transaction price to the previous transaction price. If the transaction price has increased it is classified as a buyer-initiated and if the price has fallen it is classified as a seller-initiated trade. The Trade and NBBO data sets are not synchronous, so in order to compare the trade to the quote midpoint poses a technical problem. The NBBO quote data can be updated at a different time than the contemporaneous trade is recorded. Typically, trades are recorded after the quotes have been updated with a lag of a few seconds. Lee and Ready (1991) propose using a five-second delay between the trade and quote data since the quote is updated faster than the trade data. The Lee and Ready algorithm is not perfect and there exists potential for misclassification.

For the NASDAQ data set, there exists a substantial difference between the pre-treatment and post-treatment periods. The VPIN average for the NASDAQ in the pre-treatment period is 0.36

and in the post-treatment period is 0.38.

Table: 2

For the NYSE American, there is also a difference between the pre-treatment and post-treatment periods. Prior to the implementation of the speed bump, the NYSE American had an average VPIN of 0.32 and in the post-treatment period, VPIN rose to 0.38. The realized spread also declined, showing a decrease in revenue to liquidity providers after the implementation of the speed bump.

Table: 3

The differences in VPIN values between the NASDAQ and the NYSE American are also statistically significant, with the NASDAQ having a higher average VPIN value than the NYSE American. The NASDAQ is also reported to be more liquid than the NYSE American, which given its larger size is expected. The realized spread, revenue to liquidity providers, is higher on the NASDAQ and conversely the loss to liquidity demanders, measured by the impact spread, is lower. Volatility is slightly higher on the NASDAQ compared to the NYSE American.

Table: 4

Using the Lee and Ready algorithm classification, I calculate the liquidity measures using the R package developed by Boudt et al. (2022). This calculates a wide variety of liquidity measures, and in line with Yildiz et al. (2020) I use the relative quoted spread, the relative effective spread, the relative realized spread, and the relative price impact. To estimate the volume-synchronized probability of informed trading, I use the R package developed by Ghachem and Ersan (2022). This provides the VPIN estimates and the metrics used to calculate it, such as aggregated buy and sell volumes and the order imbalance.

Methodology

4.1 Theoretical Foundation

The foundational model of Kyle (1985) provides the underlying theory of informed trading behavior. Kyle (1985) describes a model of large risk-neutral informed traders, optimizing the impact of their trades on market makers. In the Kyle model, there are three types of traders: informed, noise, and market makers. In the Kyle model informed traders have private information on the asset's value while noise traders are trading randomly. Informed traders are risk-neutral and optimize on profit maximization for a given quantity $\tilde{x}$. Informed and noise traders submit their market orders to the market maker, who is unable to discern what orders are from noise traders, $\tilde{u}$, and informed traders, $\tilde{x}$. The ex-post liquidation value of the asset, $\tilde{v}$, is normally distributed with a mean of p_0 and a variance of Σ_0. The quantity traded by noise traders, $\tilde{u}$ is normally distributed with a mean of zero and a variance of σ_u^2,

$$\tilde{v} \sim N(p_0, \Sigma_0) \tag{4.1}$$

$$\tilde{u} \sim N(0, \sigma_u^2) \tag{4.2}$$

The market maker adjusts the price based on aggregate demanded quantities for the asset, $\tilde{x} + \tilde{u}$, so that the market clears using a pricing rule $P(\tilde{x}, \tilde{u})$. The profit to the informed trader, $\tilde{\pi}$, is the difference between the ex-post liquidation value of the asset, $\tilde{v}$, and the market clearing transaction

price, $\tilde{p}$, times the quantity traded by the informed trader, $\tilde{x}$.

$$\tilde{\pi} = (\tilde{v} - \tilde{p})\tilde{x} \tag{4.3}$$

$$\tilde{p} = P(\tilde{x}, \tilde{u}) \tag{4.4}$$

The informed trader faces a trade-off on the size of the market order, since the larger the market order the larger the effect on the price but the greater the profit. The optimization results in that the informed trader will trade until the profit on the trade is equal to the price impact of the trade. Defining the constants λ and β, where the inverse of lambda is a measurement of the quantity of trades, the order flow, needed to induce a price change of one dollar and β is twice that quantity.

$$\beta = \left(\frac{\sigma_{\mu}^2}{\Sigma_0}\right)^{\frac{1}{2}} \tag{4.5}$$

$$\lambda = 2\left(\frac{\sigma_{\mu}^2}{\Sigma_0}\right)^{\frac{-1}{2}} \tag{4.6}$$

The equilibrium of the Kyle model is defined for linear functions X and P, with the following unique equilibrium for a single auction.

$$X(\tilde{v}) = \beta(\tilde{v} - p_0) \tag{4.7}$$

$$P(\tilde{u} + \tilde{x}) = p_0 + \lambda(\tilde{u} + \tilde{x}) \tag{4.8}$$

Generalizing to a sequential auction, the insider trader's quantity is described as a function of the previously observed prices and the ex-post evaluation of the stock. Market makers set a market clearing price based on previous order flows and profits from auctions, $n, ..., N$, which are determined by the sum of the difference in price times the quantity traded, for $n = 1, ..., N$.

$$\tilde{x_n} = X_n(\tilde{p}_1, ..., \tilde{p}_{n-1}, \tilde{v}) \tag{4.9}$$

$$\tilde{p_n} = P_n(\tilde{u}_1 + \tilde{x}_1, ..., \tilde{u}_{n-1} + \tilde{x}_{n-1}) \tag{4.10}$$

$$\tilde{\pi}_n = \sum_{i=n}^{N} (\tilde{v}_i - \tilde{p}_i)\tilde{x} \tag{4.11}$$

These equations define the trading model in a sequential auction. For a sequential auction equilibrium through the trading day there exists a unique linear equilibrium, which is a recursive linear equation, with the following solutions.

$$\tilde{p}_n = \tilde{p}_{n-1} + \lambda(\Delta\tilde{x}_n + \Delta\tilde{u}_n) \tag{4.12}$$

$$\tilde{x}_n = \tilde{x}_{n-1} + \beta_n(\tilde{v} + \tilde{p}_{n-1})\Delta t_n \tag{4.13}$$

Since the volume of noise trades lessens the price impact, informed traders will be more aggressive in markets with a substantial amount of noise trading. In addition, Kyle (1985) separates the informativeness of trading into trades that are informative on stock prices and trades that are not based on prices, such as liquidity trading.

4.2 Probability of Informed Trading Estimators

The original framework for the probability of informed trading models was developed by Easley et al. (1996). Building on Kyle (1985), informed traders are risk-neutral and profit-maximizing. Information can arrive daily, with a probability α determined exogenously. The new information can either positively impact prices or negatively impact stock prices. The probability of the new information being negative is δ and the probability it is positive is therefore 1-δ. PIN is estimated based on a decision tree, where events are assigned probabilities. After the completion of a day of trading, the new information which arrived at the beginning of the day is fully incorporated. The arrival rate for uninformed trades, both buy and sell trades, is ε daily orders which are both independent Poisson processes. Informed traders orders, μ, arrive based on the information that arrived in the market at the beginning of the trading day, δ. This too, follows an independent

Poisson process. Therefore, when good information arrives, a high signal, the buy orders per minute is μ + ε, while in the presence of bad information, a low signal, the sell orders are μ + ε. On non-event days, the buy and sell orders are simply ε since there is no informed trading activity on those days. In this model, therefore, the market imbalance μ, indicates the presence of informed trading.

Figure: 1

The market maker is Bayesian, who uses the arrival of trades, the order flow, to estimate the probability that informed trading is occurring. Theta is the parameter vector which defines the likelihood function.

$$\Theta = (\alpha, \delta, \mu, \varepsilon_B, \varepsilon_S) \tag{4.14}$$

Using the notation from Abada and Yagüe (2012): L is the likelihood function, B is the number of buy trades, S is the number of sell trades, α is the probability of information arriving, ε_s is the sell order from noise traders, ε_b is the buy order from noise traders, μ is the informed traders' order, e is Euler's number, and δ is the probability of negative information.

$$\begin{aligned} L((B,S)|\Theta) =& (1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \\ &+ \alpha\delta e^{\varepsilon_b}\frac{e_b^B}{B!}e^{-\varepsilon_s-\mu}\frac{(\varepsilon_s+\mu)^S}{S!} \\ &+ \alpha(1-\delta)e^{-\varepsilon-\mu}\frac{(\varepsilon_b+\mu)^B}{B!}e^{-\varepsilon_s}\frac{{\varepsilon_s}^S}{S!} \end{aligned} \tag{4.15}$$

Since these processes are assumed to be independent across days, the likelihood function for the data set,

$$M = [(B_1, S_1), ..., (B_J, S_J)], \tag{4.16}$$

is simply the product of the daily likelihoods.

$$L(M|\Theta) = \prod_{j=1}^{J} L(\Theta|B_j, S_j) \tag{4.17}$$

Using this maximum likelihood estimation of the parameters vector Θ, Easley et al. (1996) compute the probability of informed trading estimator. PIN is then a composite variable defined by the parameters: α, μ, ε_b, and ε_s.

$$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} \tag{4.18}$$

The numerator is the rate of arrival of informed traders and the denominator is the rate of all orders. PIN is therefore the estimated ratio of informed trade orders to the number of total orders. Using this model, Easley et al. (1996), differing from Kyle (1985), that active stocks have a lower probability of informed trading estimates than medium and low-activity stocks. They suggest that this is due to the greater importance of private information for low-activity stocks. While this model sufficed in the 1990s, the computerization of finance and the resulting rise of high-frequency trading has made this model impractical to estimate due to data sizes. The dramatic increase in trading speed has meant that a daily value is no longer adequate for modern financial markets. In addition, the rise of high-frequency algorithmic trading has not only changed the speed of trading but also how information is processed. Computers now trade far faster than any human being can react to and at such speeds trading information being incorporated over an entire day is no longer realistic (Abada & Yagüe, 2012; Easley et al., 2012). The increase in the number of trades makes the PIN estimation using a maximum likelihood function impractical (O'Hara, 2015).

These changes in the market microstructure of financial markets in the new millennium motivated the creation of VPIN, which is designed to capture these changes. The link between PIN and VPIN is found in Easley et al. (2008) updating the PIN methodology for this new environment. They find that the number of informed orders can be approximated by the order imbalance, the difference between the total sell and buy orders.

$$E(V^{Sell} - V^{Buy}) \approx \alpha\mu \tag{4.19}$$

And that the total number of orders is exactly the total volume bought and sold.

$$E(V^{Sell} + V^{Buy}) = \alpha\mu + 2\varepsilon \tag{4.20}$$

The second component of the VPIN metric is the concept of volume, or trade, time. High-frequency algorithms operate on the volume of trades they process and their order, rather than the timing of trades (Easley et al., 2012). Instead of updating the VPIN metric every hour, instead, it is updated once a given volume has been traded in the market.

4.3 Volume-synchronized Probability of Informed Trading

The variable of interest for my analysis is the volume-synchronized probability of informed trading (VPIN) developed by Easley et al. (2012). Using the R package by Ghachem and Ersan (2022) I estimate the VPIN for each stock in the sample. The first step in the calculation of the VPIN metric is to define standardized volumes of trade, the volume bucket size (VBS), where i is a day in the sample, j is a firm, N is the total number of days in the sample, and V is the trade volume for firm j on day i.

$$VBS_j = \frac{\frac{1}{N}\sum_{i=1}^{N} V_{i,j}}{50} \tag{4.21}$$

The VBS is the average daily trade volume for a given stock divided by fifty. This results in each trading day being divided into, on average, 50 volume buckets. The VBS is unique for each firm in the sample. These segments are sequential and indicated by τ. Trades in bucket τ are aggregated by the minute and added to the total volume traded in τ until the volume traded reaches the VBS. Any excess volume in the last added trade is added to the next volume bucket. For each volume bucket, τ, the order imbalance is calculated. Easley et al. (2012) developed a technique for determining the number of buyer-initiated and seller-initiated trades probabilistically in every minute which is used for calculating VPIN, this process is bulk classification and uses the standard normal distribution, Z, evaluated at the change in the transaction price, ΔP, and the standard deviation of the change

in the transaction price, $\sigma_{\Delta P}$. The buy volume is the VBS times the value of the standard normal distribution of the standardized change in the transaction price in the bucket τ for the firm j. The sell volume is the complementary of the standard normal distribution for the buy side times the volume bucket size.

$$V_{\tau,j}^{Buy} = VBS_j \times Z\left(\frac{\Delta P_{\tau,j}}{\sigma_{\Delta P_{\tau,j}}}\right) \tag{4.22}$$

$$V_{\tau,j}^{Sell} = VBS_j \times \left(1 - Z\left(\frac{\Delta P_{\tau,j}}{\sigma_{\Delta P_{\tau,j}}}\right)\right) \tag{4.23}$$

$$OI_{\tau,j} = \mid V_{\tau,j}^{Sell} - V_{\tau,j}^{Buy} \mid \tag{4.24}$$

The order imbalance is the absolute value of the volume sold and the volume bought in each volume bucket. With these order imbalances (OI) we can calculate the VPIN metric.

$$VPIN_{\tau,j} = \frac{\sum_{\tau=1}^{50} OI_{\tau,j}}{50 \times VBS_j} \tag{4.25}$$

VPIN is therefore a moving average of the order imbalance over the volume bucket sample length, in this case as in Easley et al. (2012), the sample length is 50 buckets. This results in VPIN being updated on average 50 times a day. Longer sample lengths are possible, to estimate any given sample length η we can use the following equations.

$$VBS_j = \frac{\frac{1}{N}\sum_{i=1}^{N} V_{i,j}}{50} \tag{4.26}$$

$$VPIN_{\tau,j} = \frac{\sum_{\tau=1}^{\eta} OI_{\tau,j}}{\eta \times VBS_j} \tag{4.27}$$

Abada and Yagüe (2012) find that the sample length parameter changes exactly what the VPIN metric is measuring. For instance, a sample length of 50 is equivalent to a daily average, while a sample length of 250 is a five-day average. Abada and Yagüe (2012) note that it is possible

to change the time bar length, the sample length, and the number of buckets. Changing these parameters will change the interpretation of the VPIN metric. In this analysis the sample length is set to 50, the VPIN is therefore approximately a rolling daily average, with a rolling window of 50 volume buckets. The maximum possible value for the VPIN estimate is one and the lowest possible estimate is zero.

4.4 Liquidity Measures

Next, I create measures of liquidity using the methods from Hendershott et al. (2011) using the R package developed by Boudt et al. (2022). Following the methodology from Bessembinder (2003) and Boehmer (2005), I estimate a wide variety of liquidity measurements as controls, in addition to the measurements from the VPIN estimation estimated using the package developed by Ghachem and Ersan (2022). To measure market liquidity I use the effective and quoted spreads. The quoted spread is the bid and ask spread divided by the mid-quote which is the average of the highest ask and lowest bid, $(Bid_t + Ask_t)/2$. The effective spread is twice the ratio of the difference between the trade price and the mid-quote, where D is the direction of the trade from the Lee and Ready algorithm, which is equal to 1 for a buyer-initiated trade and -1 for a seller-initiated trade. These measures use matched trades and quotes so that the time, t, is the time of the matched trade and quote, not the time of the quote data.

$$RelativeQuotedSpread_t = \frac{Bid_t - Ask_t}{MidQuote_t} \times 100 \tag{4.28}$$

$$RelativeEffectiveSpread_t = 2D_t \times \frac{TradePrice_t - MidQuote_t}{MidQuote_t} \times 100 \tag{4.29}$$

The realized spread measures the cost of immediacy, estimating revenue to the liquidity providers over a 300-second, five-minute, window following the methodology in Chang and Chou (2022). This estimates the revenue to the liquidity providers or the cost of immediacy for the liquidity

demander.

$$RelativeRealizedSpread_t = 2D_t \times \frac{TradePrice_t - MidQuote_{t+300}}{MidQuote_t} \times 100 \tag{4.30}$$

The final measurement of liquidity used is the price impact of a trade. The price impact is half the difference between the effective and realized spreads and measures the loss to liquidity demanders.

$$Relative\ Price\ Impact_t = \frac{RelativeEffectiveSpread_t - RelativeRealizedSpread_t}{2 \times MidQuote_t} \tag{4.31}$$

This is in line with the methodology of Yildiz et al. (2020), who aggregate liquidity measures to the VPIN volume bucket. In addition, I include the volumes bought and sold, the order imbalance, and the bid and offer volumes. These measures are then aggregated into the VPIN buckets by a time-weighted average to control for liquidity effects for firm j and volume bucket τ.

$$\begin{aligned} Liquidity_{\tau,j} &= RelativeQuotedSpread_{\tau,j} + RelativeEffectiveSpread_{\tau,j} \\ &\quad + RelativeRealizedSpread_{\tau,j} + RelativePriceImpact_{\tau,j} \\ &\quad + SellVolume_{\tau,j} + BuyVolume_{\tau,j} + OI_{\tau,j} \\ &\quad + BidVolume_{\tau,j} + OfferVolume_{\tau,j} \end{aligned} \tag{4.32}$$

4.5 Volatility

To measure the volatility I use the daily realized volatility of returns and the daily realized variance of the returns of the firm j and the VPIN bucket τ using the R package developed by Boudt et al. (2022).

$$Volatility_{\tau,j} = RealizedVolatility_{\tau,j} + RealizedVariance_{\tau,j} \tag{4.33}$$

This gives a quadratic relationship for volatility, allowing for a non-linear effect of volatility on the volume-synchronized informed trading estimates. These values, as daily values for each firm, are not aggregated and are merged with the VPIN estimates on the date when each volume bucket, τ,

begins.

4.6 Fixed Effects

Trading is highly irregularly spaced with large increases in trading at the beginning and the end of the trading day and reduced trading activity during the lunch break (Balaban et al., 2018; Easley et al., 2012). To control for this intra-day seasonality I use hourly fixed effects, δ_τ, as controls following the methodology from Yildiz et al. (2020). In addition, trading varies based on the day of the week, with activity on Friday being affected by traders closing for the weekend and activity on Monday incorporating new information that was announced during the weekend (Balaban et al., 2018; Kim & Ryu, 2022). This Monday effect decreases returns and increases volatility (Balaban et al., 2018; Kim & Ryu, 2022). To control for this weekly seasonality I use day of the week fixed effects, η_τ. These temporal dummy variables will control for the effect of the hour and the weekday on VPIN. Following Yildiz et al. (2020) I used firm-fixed effects to treat time-invariant differences between firms in their VPIN values. The firm-fixed effects, ϕ_j, adjust the mean for each stock, capturing some of the firm-specific non-observables. These fixed effects are used in every regression. For my two-way fixed effects specification I use a time trend indicator variable for each day of the sample, π_τ. The difference-in-differences specification uses only the hourly, firm, and day-of-week fixed effects.

4.7 Difference-in-Differences

Using a difference-in-differences approach I have three variables of interest: the treated group (Treatment), the period after the treatment (Time), and the difference-in-differences estimator (DID).

$$\begin{aligned} VPIN_{\tau,j} &= \beta_1 DID_{\tau,j} + \beta_2 Time_{\tau,j} + \beta_3 Treatment_{\tau,j} + \beta_4 Liquidity_{\tau,j} \\ &\quad + \beta_5 Volatility_{\tau,j} + \phi_j + \eta_\tau + \delta_\tau + \hat{\varepsilon}_{\tau,j} \end{aligned} \tag{4.34}$$

The exchange variable, Treatment, is a dummy variable measuring the difference in the conditional mean of the VPIN estimates between the control group, the NASDAQ, and the treatment group, the NYSE American. The NYSE American is indicated with a value of 1 and the NASDAQ is indicated with a value of 0. Time is a dummy variable equal to 0 before the speed bump was implemented on the NYSE American and 1 afterwards. The Time indicator measures the difference in the conditional mean of the VPIN estimates before and after the implementation of the speed bump for both the treatment and the control. The effect of the speed bump itself is the difference-in-differences estimator which is equal to 1 for the NYSE American after the exchange implemented the speed bump on July 24 and zero otherwise, measuring the difference in the conditional mean for the treated group after the treatment. The two-way fixed effects model includes the daily indicator variable.

$$\begin{aligned} VPIN_{\tau,j} &= \beta_1 DID_{\tau,j} + \beta_2 Time_{\tau,j} + \beta_3 Treatment_{\tau,j} + \beta_4 Liquidity_{\tau,j} \\ &\quad + \beta_5 Volatility_{\tau,j} + \phi_j + \eta_\tau + \delta_\tau + \pi_\tau + \hat{\varepsilon}_{\tau,j} \end{aligned} \tag{4.35}$$

Empirical Results

The alternative hypothesis is that informed trading, as estimated by the volume-synchronized probability of informed trading, will decline as informed traders leave and uninformed traders flock to the NYSE American. The null hypothesis of this analysis is that the implementation of the speed bump will have no effect on informed trading. My results show that the NYSE American's speed bump was responsible for a decline in informed trading on that exchange. The difference-in-differences estimate shows a 3.6% decline in the probability of informed trading due to the implementation of the latency delay. The two-way fixed effects specification estimates a 3.1% decline in the probability of informed trading due to the latency delay's implementation. The effect is statistically significant with a p-value below 0.01 with both heteroskedasticity robust and firm-clustered standard errors in both specifications. The Treatment coefficients can be interpreted as the difference in the probability of informed trading between the exchanges, all else equal. The Time coefficients can be interpreted as the difference in informed trading on both exchanges after the speed bump was implemented on the NYSE American, all else equal. The difference-in-differences coefficient is the causal effect of the speed bump on the volume-synchronized probability of informed trading.

In the first regression, I apply a two-way fixed effects model with heteroskedasticity-robust standard errors. For this specification the estimated effect of the speed bump on the probability of informed trading is a decline of 3.1%. For this specification, the Time effect is not statistically significant. The treated group, Treatment, has a lower average VPIN value which implies that the NYSE American had a higher conditional mean of informed trading compared to the control group,

the NASDAQ. In the second regression I implement a firm-fixed effects model without daily fixed effects using heteroskedasticity-robust standard errors. In the firm-fixed effects specification the estimated effect of the speed bump is -3.7%, and remains highly significant. In this specification, the Time variable is statistically significant and the Treatment has a lower coefficient but remains statistically significant.

Table: 5

I repeat these specifications with clustered standard errors with firm-level clusters. For the third specification, neither the Treatment nor the Time effects are statistically significant, however, the difference-in-differences estimator remains statistically significant for both the two-way fixed effects and difference-in-differences specifications. In the two-way fixed effects specification with clustered standard errors, the p-value for the difference-in-differences estimator rises to the 0.001 significance level but remains highly statistically significant.

These results show that the speed bump implemented on the NYSE American exchange had a negative effect on informed trading. In all specifications, the effect is statistically and economically significant. As can be seen in Table 2, VPIN estimates for the NYSE American were 31.5% prior to the speed bump's implementation and 36.8% after the implementation. In Table 5 the effect of the speed bump is -3.1%, representing a substantial decline in informed trading after the implementation of the speed bump on the NYSE American. The difference-in-differences estimator can be interpreted as the causal effect of the speed bump on informed trading, all else equal.

This result is in line with the expected effect of a speed bump and shows that even a small latency delay can have substantial effects on informed trading. This confirms the alternative hypothesis. The effect of this speed bump on lowering order flow toxicity suggests that similar speed bumps may lower the likelihood of liquidity crises, like the Flash Crash, from occurring on exchanges with these speed bumps. It also suggests that speed bumps effectively deter informed trading on these exchanges, however, this result doesn't prove that speed bumps deter informed trading for the financial system as a whole.

Conclusion

In financial markets, there are three different types of traders: informed, uninformed, and market makers (Grossman & Stiglitz, 1980; Kyle, 1985). Informed traders have private information on stock values which has not yet been incorporated into the market price. In order to profit from this private information informed traders must trade with market makers. Market makers are traders who take on the opposing side of an initiated trade taking the bid-ask spread as compensation. As such, market makers who are exposed to informed traders can face unexpected losses on trades with informed traders (Easley & O'Hara, 1992). Since North American financial markets are anonymous, market makers are unable to distinguish between trades by uninformed traders and informed traders. As such, market makers will pull out of markets where there is substantial information risk leading to a liquidity crisis which can cause substantial issues for the functioning of the financial system by creating high volatility periods (Easley & O'Hara, 1992; Easley et al., 2012; Yildiz et al., 2020).

In order to detect informed trading, researchers have used market imbalances. Easley and O'Hara (1992) began this strand of research by creating the probability of informed trading estimator, however, the rise of high-speed trading and the resulting changes in financial markets made their methodology impractical and slow (Easley et al., 2012). In response, Easley et al. (2012) developed an intraday measure, the volume-synchronized probability of informed trading (VPIN) estimator with an updated methodology for a high-frequency setting. VPIN is correlated with market liquidity, trading costs, and volatility of returns (Yildiz et al., 2020). Given the extremely high volumes of trade in financial markets, small changes in the VPIN can have economically significant

effects.

An innovation in exchange technology designed to mediate the effect of high-frequency traders is the latency delay. Aldauf and Mollner (2020) suggest that high-frequency traders can exploit changes in the midpoint of the bid/ask spread by trading with low-frequency traders before orders can be updated. Cartea and Penalva (2013) find that high-frequency traders increase order imbalances. Brolley and Cimon (2020) find that informed traders will move away from exchanges with latency delays. Both of these results suggest that the VPIN metric should decline after the NYSE American implemented a speed bump. In contrast, Aoyagi (2019) suggests that a latency delay may increase or decrease the number of informed traders, depending on the initial number of informed traders on the exchange. In a study of the IEX Chakrabarty et al. (2020) found that back-running strategies declined and market quality improved due to the latency delay on the IEX. In an working paper, Liu and Xu (2023) find that the implementation of a speed bump on the NYSE American reduced informed trading and lowered price discovery. My analysis offers support to this result using the alternative methodology of volume-synchronized informed trading estimates.

On July 24th the NYSE American implemented a latency delay of 350 microseconds. Using a data set from the TAQ database, I selected 45 of the most actively traded securities from the NYSE American, the treatment group, and the NASDAQ, the control group, which were traded on the NYSE American for the entire sample period. Using this data set I estimate the VPIN metric for each stock on both exchanges along with liquidity and volatility measures. Using a difference-in-differences estimation strategy I am able to estimate the causal effect of the speed bump on informed trading on the NYSE American. I find a statistically significant negative causal effect in informed trading on the NYSE American due to the introduction of the latency delay. This result remains statistically significant using both heteroskedasticity robust and firm-level clustered standard errors.

This result strongly suggests that latency delays are an effective strategy to limit the amount of informed trading in equity markets. By introducing speed bumps exchanges can lower order flow toxicity. The introduction of the speed bump on the NYSE American caused VPIN values to

fall indicating that the speed bumps reduce liquidity-induced volatility on these exchanges. This suggests that speed bumps may be able to decrease the severity and or frequency of events like the Flash Crash. Future research avenues would include the effects of latency delays on informed trading for the financial system as a whole, the effect of latency delays on flash crash severity and frequency, the use of other estimation strategies of informed trading, and the use of synthetic controls to measure the effect of speed bumps on informed trading.

Appendix

Table 1: Exchange Summary Statistics[1]

	NASDAQ					NYSE American				
	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max
VPIN	0.37	0.18	0.35	0.00	1.00	0.34	0.21	0.30	0.00	1.00
Offer	6.72	12.53	2.36	0.16	88.43	6.74	13.16	2.12	0.17	73.75
Bid	6.70	12.53	2.34	0.16	87.94	6.72	13.12	2.11	0.17	73.38
Relative Quoted Spread (bps)	1.05	2.09	0.47	−17.30	165.65	0.85	0.98	0.56	−0.49	19.60
Relative Effective Spread (bps)	0.65	1.75	0.24	−13.35	176.86	0.66	1.11	0.37	−15.19	24.21
Relative Realized Spread (bps)	−1.22	9.45	−0.10	−80.97	231.28	−0.34	8.79	−0.02	−194.02	121.78
Relative Price Impact (bps)	0.96	4.84	0.16	−48.71	41.56	0.48	4.51	0.14	−58.85	99.56
Realized Volatility	0.06	0.06	0.03	0.00	0.44	0.06	0.06	0.04	0.00	0.48
Order Imbalance	1397.76	3533.00	440.77	0.00	46030.90	661.08	969.48	232.80	0.00	5632.88
Sell Volume	2000.37	4198.96	902.84	0.00	46030.90	988.84	900.48	683.46	0.00	5632.88
Buy Volume	1941.59	3904.27	905.83	0.00	46030.90	970.69	871.95	678.35	0.00	5632.88

[1] Summary statistics for both the NYSE American and the NASDAQ for entire sample period.

Table 2: NASDAQ Pre/Post-Treatment Summary Statistics[2]

	Post-Treatment (N=61205)		Pre-Treatment (N=70278)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
VPIN	0.3616	0.1778	0.3755	0.1740	0.0139***	0.0010
Relative Quoted Spread (bps)	0.9883	1.7861	1.1121	2.3269	0.1237***	0.0114
Relative Effective Spread (bps)	0.5900	1.5665	0.6983	1.8980	0.1083***	0.0096
Relative Realized Spread (bps)	-0.9785	7.5927	-1.4295	10.7986	-0.4510***	0.0510
Relative Price Impact (bps)	0.8627	3.7580	1.0387	5.6107	0.1761***	0.0261
Realized Volatility	0.0537	0.0522	0.0568	0.0622	0.0031***	0.0003
Offer	7.3844	13.9457	6.1336	11.1253	-1.2508***	0.0703
Bid	7.3676	13.9414	6.1170	11.1222	-1.2506***	0.0703
Order Imbalance	1407.6645	3442.1994	1389.1320	3610.2216	-18.5324	19.4692
Sell Volume	2016.0542	4132.9476	1986.7139	4255.5929	-29.3403	23.1684
Buy Volume	1985.3265	3932.4324	1903.4917	3879.2099	-81.8348***	21.6052
	N	Pct.	N	Pct.		
NASDAQ	61205	100.0	70278	100.0		

[2] Balance table for the NASDAQ exchange, pre-treatment variables are compared to post-treatment. Significance levels: 0.05, 0.01, and 0.001.

Table 3: NYSE American Pre/Post-Treatment Summary Statistics[3]

	Post-Treatment (N=52631)		Pre-Treatment (N=67431)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
VPIN	0.3150	0.1838	0.3678	0.2182	0.0528***	0.0012
Relative Quoted Spread (bps)	0.9553	1.0319	0.7690	0.9226	-0.1863***	0.0057
Relative Effective Spread (bps)	0.7230	1.1838	0.6155	1.0443	-0.1076***	0.0065
Relative Realized Spread (bps)	0.2026	8.4325	-0.7572	9.0300	-0.9598***	0.0506
Relative Price Impact (bps)	0.2555	4.4147	0.6631	4.5762	0.4076***	0.0261
Realized Volatility	0.0609	0.0529	0.0597	0.0693	-0.0012***	0.0004
Offer	5.4881	10.6751	7.7206	14.7390	2.2325***	0.0734
Bid	5.4538	10.5830	7.7030	14.7295	2.2492***	0.0731
Order Imbalance	601.2380	890.9533	707.7914	1024.1873	106.5534***	5.5352
Sell Volume	971.8420	835.4438	1002.1121	947.9419	30.2701***	5.1563
Buy Volume	926.1119	816.6217	1005.4843	911.2990	79.3724***	4.9986
	N	Pct.	N	Pct.		
NYSE American	52631	100.0	67431	100.0		

[3] Balance table for the NYSE American exchange, pre-treatment variables are compared to post-treatment. Significance levels: 0.05, 0.01, and 0.001.

Table 4: NYSE American/NASDAQ Balance Table[4]

	NASDAQ (N=131483)		NYSE American (N=120062)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
VPIN	0.3690	0.1759	0.3446	0.2055	-0.0244***	0.0008
Relative Quoted Spread (bps)	1.0545	2.0935	0.8507	0.9764	-0.2038***	0.0064
Relative Effective Spread (bps)	0.6479	1.7523	0.6626	1.1089	0.0147*	0.0058
Relative Realized Spread (bps)	-1.2195	9.4453	-0.3365	8.7860	0.8831***	0.0364
Relative Price Impact (bps)	0.9568	4.8381	0.4845	4.5106	-0.4723***	0.0186
Realized Volatility	0.0553	0.0578	0.0602	0.0627	0.0049***	0.0002
Offer	6.7158	12.5330	6.7419	13.1601	0.0261	0.0514
Bid	6.6992	12.5293	6.7170	13.1222	0.0179	0.0513
Order Imbalance	1397.7588	3533.0007	661.0821	969.4810	-736.6768***	10.1371
Sell Volume	2000.3717	4198.9571	988.8428	900.4803	-1011.5290***	11.8680
Buy Volume	1941.5855	3904.2736	970.6902	871.9498	-970.8954***	11.0574

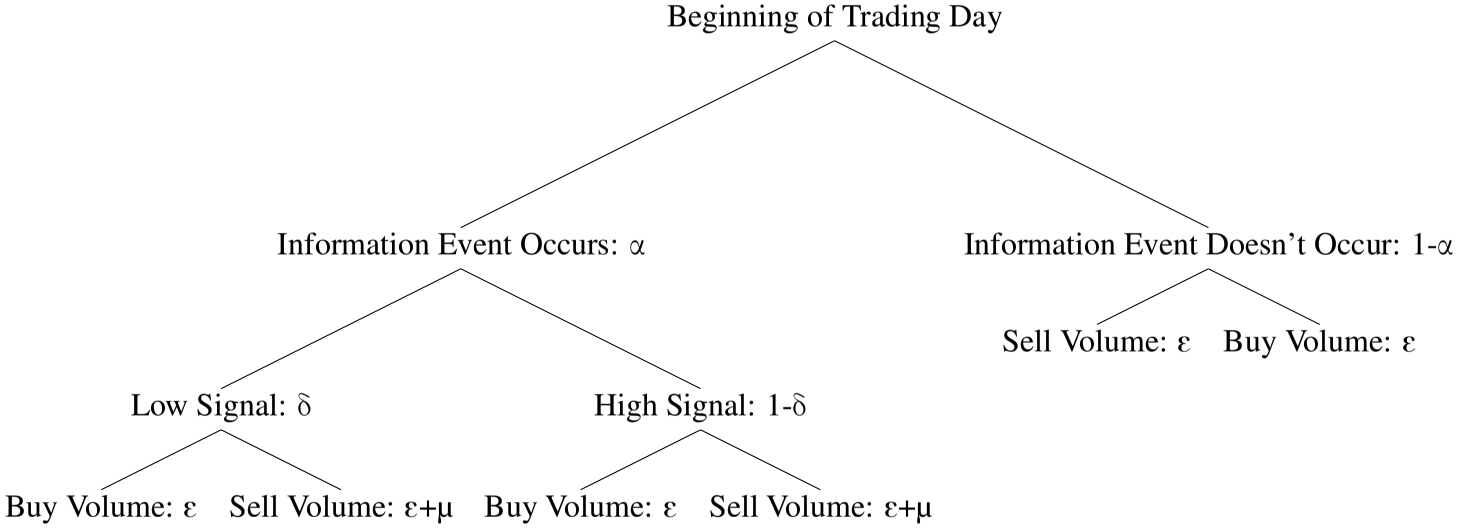
[4] Balance table for the NYSE American and the NASDAQ. Significance levels: 0.05, 0.01, and 0.001.

Table 5: Difference-in-Differences Estimation[5]

	Time & Firm Fixed Effects	Firm Fixed Effects	Time & Firm Fixed Effects	Firm Fixed Effects
Time	−0.003	0.020***	−0.003	0.020
	0.006 (0.568)	0.001 (<0.001)	0.035 (0.925)	0.014 (0.165)
Treatment	−0.021***	−0.016***	−0.021	−0.016
	0.001 (<0.001)	0.001 (<0.001)	0.013 (0.101)	0.013 (0.226)
DID	−0.031***	−0.036***	−0.031**	−0.036***
	0.001 (<0.001)	0.001 (<0.001)	0.010 (0.001)	0.011 (<0.001)
Num. Obs.	251 545	251 545	251 545	251 545
R2	0.307	0.283	0.307	0.283
R2 Adj.	0.306	0.283	0.306	0.283
Std. Errors	Robust	Robust	Clustered by Firm	Clustered by Firm

[5] Difference-in-differences estimation, with time and firm-level fixed effects. DID is the difference-in-differences estimator. Time is the treatment period indicator variable. Treatment is the treatment indicator. Robust indicates heteroskedastic robust standard errors and Clustered by Firm indicates firm-level clustered standard errors. Time & Firm Fixed Effects indicates two-way fixed effects model including firm, hourly, day of week, and daily fixed effects. Firm Fixed Effects indicates firm, hourly, and day-of-week fixed effects. Significance levels: 0.05, 0.01, and 0.001. Standard errors reported below coefficients. P-values are reported below coefficients in parentheses.

Figure 1: Tree Diagram of the PIN Trading Process



Bibliography

Abada, D., & Yagüe, J. (2012). From PIN to VPIN: An introduction to order flow toxicity. *Spanish Review of Financial Economics*, *10*(2), 74–83. https://doi.org/10.1016/j.srfe.2012.10.002

Aldauf, M., & Mollner, J. (2020). High-frequency trading and market performance. *Journal of Finance (New York)*, *75*(3). https://doi.org/10.1111/jofi.12882

Andersen, T. G., & Bondarenko, O. (2014a). Reflecting on the VPIN dispute. *Journal of Financial Markets*, *17*, 53–64. https://doi.org/10.1016/j.finmar.2013.08.002

Andersen, T. G., & Bondarenko, O. (2014b). VPIN and the Flash Crash. *Journal of Financial Markets*, *17*, 1–46. https://doi.org/10.1016/j.finmar.2013.05.005

Aoyagi, J. (2019). Strategic speed choice by high-frequency traders under speed bumps. *ISER DP No. 1050*. https://ssrn.com/abstract=333738

Balaban, E., Ozgen, T., & Karidis, S. (2018). Intraday and interday distribution of stock returns and their asymmetric conditional volatility: Firm-level evidence. *Physica A: Statistical Mechanics and its Applications*, *503*, 905–915. https://doi.org/10.1016/j.physa.2018.02.116

Berman, G. E. (2010). Speech by SEC staff: Market participants and the May 6 Flash Crash [Transcript]. https://www.sec.gov/news/speech/2010/spch101310geb.htm

Bessembinder, H. (2003). Trade execution costs and market quality after decimalization. *Journal of Financial and Quantitative Analysis*, *38*(4), 747–777. http://www.jstor.org/stable/4126742

Bjursell, J., Wang, G. H. K., & Zheng, H. (2017). VPIN, jump dynamics and inventory announcements in energy futures markets. *Journal of Futures Markets*, *37*(6), 542–577. https://doi.org/10.1002/fut.21839

Boehmer, E. (2005). Dimensions of execution quality: Recent evidence for US equity markets. *Journal of Financial Economics*, *78*(3), 553–582. https://doi.org/10.1016/j.jfineco.2004.11.002

Boudt, K., Kleen, O., & Sjørup, E. (2022). Analyzing intraday financial data in R: The highfrequency package. *Journal of Statistical Software*, *104*(8), 1–36. https://doi.org/10.18637/jss.v104.i08

Brolley, M., & Cimon, D. A. (2020). Order flow segmentation, liquidity and price discovery: The role of latency delays. *Journal of Financial and Quantitative Analysis*, *55*(8). https://doi.org/10.1017/S002210901900067X

Cartea, Á., & Penalva, J. (2013). Where is the value in high frequency trading? *Quarterly Journal of Finance*, *2*, 1–46. https://ssrn.com/abstract=1855555

Chakrabarty, B., Huang, J., & Jain, P. K. (2020). Effects of a speed bump on market quality and exchange competition. *SSRN*. https://ssrn.com/abstract=3280645

Chakrabarty, B., Pascual, R., & Shkilko, A. (2015). Evaluating trade classification algorithms: Bulk volume classification versus the tick rule and the Lee-Ready algorithm. *Journal of Financial Markets (Amsterdam, Netherlands)*, *15*, 52–79. https://doi.org/10.1016/j.finmar.2015.06.001

Chang, Y.-K., & Chou, R. K. (2022). Algorithmic trading and market quality: Evidence from the Taiwan Index Futures Market. *Journal of Futures Markets*, *42*(10), 1837–1855. https://doi.org/10.1002/fut.22362

Chen, H., Foley, S., Goldstein, M. A., & Ruf, T. (2017). The value of a millisecond: Harnessing information in fast, fragmented markets. https://doi.org/10.2139/ssrn.2860359

Easley, D., de Prado, M. L., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, *25*(5), 1457–1493. https://doi.org/10.1093/rfs/hhs053

Easley, D., de Prado, M. M. L., & O'Hara, M. (2014). VPIN and the Flash Crash: A rejoinder. *Journal of Futures Markets (Amsterdam, Netherlands)*, *17*, 47–52. https://doi.org/10.1016/j.finmar.2013.06.007

Easley, D., Engle, R. F., O'Hara, M., & Wu, L. (2008). Time-varying arrival rates of informed and uninformed trades. *Journal of Financial Econometrics*, *6*(2), 171–207. https://doi.org/10.1093/jjfinec/nbn003

Easley, D., Hvidkjaer, S., & O'Hara, M. (2002). Is information risk a determinant of asset returns? *Journal of Finance*, *57*, 2185–2221. https://doi.org/10.1111/1540-6261.00493

Easley, D., Kiefer, N. M., O'Hara, M., & Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. *Journal of Finance*, *51*(4), 1405–1436. https://doi.org/10.2307/2329399

Easley, D., & O'Hara, M. (1992). Adverse selection and large trade volume: The implications for market efficiency. *Journal of Financial and Quantitative Analysis*, *27*(2), 185–208. https://doi.org/10.2307/2331367

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*(2), 383–417. https://doi.org//10.2307/2325486

Foucault, T. (1999). Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets*, *2*(2), 99–134. https://doi.org/10.1016/S1386-4181(98)00012-3

Ghachem, M., & Ersan, O. (2022). PINstimation: An R package for estimating models of probability of informed trading. https://ssrn.com/abstract=4117946

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, *70*(3), 393–408.

Hendershott, T. J., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *Journal of Finance (New York)*, *66*(1), 1–33. https://doi.org/10.1111/j.1540-6261.2010.01624.x

Hendershott, T. J., & Moulton, P. C. (2011). Automation, speed, and stock market quality: The NYSE's hybrid. *Journal of Financial Markets (Amsterdam, Netherlands)*, *14*(4), 568–604. https://doi.org/https://doi.org/10.1016/j.finmar.2011.02.003

Hoffmann, P. (2014). A dynamic limit order market with fast and slow traders. *Journal of Financial Economics*, *113*(1). https://doi.org/10.1016/j.jfineco.2014.04.002

Hu, E. (2018). Intentional access delays, market quality, and price discovery: Evidence from IEX becoming an exchange. *U.S. Security Exchange Commission*. https://ssrn.com/abstract=3195001

Jones, C. M. (2018). Understanding the market for U.S. equity market data. *U.S. Security Exchange Commission*. https://www.sec.gov/comments/4-729/4729-4545881-176154.pdf

Kim, K., & Ryu, D. (2022). Sentiment changes and the Monday effect. *Finance Research Letters*, *47*, 102709–. https://doi.org/10.1016/j.frl.2022.102709

Kitamura, Y. (2017). Predicting a flash crash in the yen/dollar foreign exchange market. *Applied Eeconomics Letters*, *24*(14), 987–990. https://doi.org/10.1080/13504851.2016.1245831

Kuhle, W. (2023). Latency arbitrage and the synchronized placement of orders. *Financial Innovation (Heidelberg)*, *9*(1), 99–18. https://doi.org/https://doi.org/10.1186/s40854-023-00491-5

Kyle, A. S. (1985). Continuous auctions and insider trading. *Annals of Mathematics*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

Laughlin, G., Aguirre, A., & Grundfest, J. (2014). Information transmission between financial markets in Chicago and New York. *Financial Review (Buffalo, N.Y.)*, *49*(2), 283–312. https://doi.org/10.1111/fire.12036

Lee, C. M. C., & Ready, M. J. (1991). Inferring trade direction from intraday data. *Journal of Finance*, *46*(2), 733–746. https://doi.org/10.2307/2328845

Liu, B., & Xu, K. (2023). Speed bump and stock market quality: evidence from NYSE American. *Available at SSRN:* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4593778

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, *116*(2), 257–270. https://doi.org/10.1016/j.jfineco.2015.01.003

Silva, F. B. G., & Volkova, E. (2018). Can VPIN forecast geopolitical events? evidence from the 2014 Crimean Crisis. *Annals of Finance*, *14*(1), 125–141. https://doi.org/10.1007/s10436-017-0314-z

Wu, K., Wes, E. B., Gu, M., Leinweber, D., & Rubel, O. (2013). A big data approach to analyzing market volatility. https://escholarship.org/uc/item/6rx2w70r

Yildiz, S., Van Ness, B., & Van Ness, R. (2020). VPIN, liquidity, and return volatility in the U.S. equity markets. *Global Finance Journal*, *45*, 100479. https://doi.org/10.1016/j.gfj.2019.100479