FLOW TOXICITY AND VOLATILITY
IN A HIGH FREQUENCY WORLD

David Easley
dae3@cornell.edu

Marcos M. López de Prado
marcos.lopezdeprado@tudor.com

Maureen O’Hara
mo19@cornell.edu

April 28, 2011

ABSTRACT
Order flow is regarded as toxic when it adversely selects market makers, who may be unaware that they are providing liquidity at a loss. Flow toxicity can be expressed by the Probability of Informed Trading (PIN). We present a new procedure to estimate the Probability of Informed Trading based on volume imbalance and trade intensity (the VPIN* informed trading metric). An important advantage of the VPIN metric over previous estimation procedures is that it does not require the intermediate estimation of non-observable parameters describing the order flow or the application of numerical methods. An additional advantage is that VPIN is updated in volume-time, rather than clock-time, which improves its predictive power in a high frequency world. Monte Carlo experiments show that our estimate is accurate for all theoretically possible combinations of parameters. Finally, we show that the VPIN metric predicts short-term, toxicity-induced volatility.

Keywords: Flash crash, liquidity, flow toxicity, volume imbalance, trade intensity, market microstructure, probability of informed trading, VPIN.

JEL codes: C02, D52, D53, G14.

1 We would like to thank Robert Engle, Andrew Karolyi, John Campbell, Luis Viceira, Torben Andersen, Oleg Bondarenko, and seminar participants at Cornell University, Harvard University, the QFF seminar at the Chicago Mercantile Exchange, the University of Piraeus, Complutense University, and the CFTC for helpful comments.

2 Scarborough Professor and Donald C. Opatrny Chair Department of Economics, Cornell University.

3 Head of High Frequency Futures, Tudor Investment Corp., and Postdoctoral Fellow of RCC at Harvard University.

4 Purcell Professor of Finance, Johnson Graduate School of Management, Cornell University.

* VPIN is a trademark of Tudor Investment Corp. The authors have applied for a patent on VPIN and have a financial interest in it.
High frequency (HF) trading firms account for over 70% of the volume in U.S. equity markets and are fast approaching 50% of the volume in futures markets. These HF firms typically act as market makers, providing liquidity to position-takers by placing passive orders at various levels of the electronic order book. A passive order is defined as an order that does not cross the market, thus the originator has no direct control on the timing of its execution. HF market makers generally do not make directional bets, but rather strive to earn razor thin margins on large numbers of trades. Their ability to do so depends on limiting their position risk, which is greatly affected by their ability to control adverse selection in the execution of their passive orders.

Practitioners usually refer to adverse selection as the “natural tendency for passive orders to fill quickly when they should fill slowly and fill slowly (or not at all) when they should fill quickly” (Jeria and Sofianos, [2008]). This intuitive formulation is consistent with market microstructure models (see, for example, Glosten and Milgrom [1985], Kyle [1985], and Easley and O’Hara [1987, 1992b]), whereby informed traders take advantage of uninformed traders. Order flow is regarded as toxic when it adversely selects market makers who may be unaware that they are providing liquidity at a loss.

This paper develops a new framework for measuring order flow toxicity in a high frequency world. As discussed by Easley, Engle, O’Hara and Wu [2008], a fundamental insight of the microstructure literature is that the order arrival process is informative for subsequent price moves in general, and flow’s toxicity in particular. Extracting this information from order flow in a high frequency framework, however, is complicated by the very nature of trading in this new market structure. We argue that in the high frequency world trade-time, as captured by volume, is now a more relevant metric than clock-time. Information is also a different concept, relating now to an underlying event that induces unbalanced trade over a relatively short horizon. Information events can arise for a variety of reasons, some related to asset returns, but others reflecting more systemic or portfolio-based effects. Our particular application is to futures contracts, where information is more likely to be related to systemic factors, or to variables reflecting hedging or other portfolio considerations.

This paper presents a new procedure to estimate flow toxicity directly and analytically, based on a process subordinated to volume arrival, which we name Volume-Synchronized Probability of Informed Trading, or the VPIN informed trading metric. The original PIN estimation approach (see Easley, Kiefer, O’Hara and Paperman [1996]) entailed maximum likelihood estimation of unobservable parameters fitted on a mixture of three Poisson distributions of daily buy and sell orders on stocks. That static approach was extended by the Easley, Engle, O’Hara and Wu [2008] GARCH specification, which models a time-varying arrival rate of informed and uninformed traders. The approached based on the VPIN metric developed in this paper does not require the intermediate numerical estimation of non-observable parameters, and is updated in stochastic time which is calibrated to have an equal volume of trade in each time interval. Thus, our methodology overcomes the difficulties of estimating PIN models in high volume markets, and provides an analytically tractable way to measure the toxicity of order flow using high frequency data.
We provide empirical evidence on the statistical properties of the VPIN metric. We show how volume bucketing (time intervals selected so that each has an equal volume of trade) reduces the impact of volatility clustering in the sample. Because large price moves are associated with large volumes, sampling by volume is a proxy for sampling by volatility. The result is a time series of observations that follows a distribution that is closer to normal and is less heteroskedastic than it would be if it were sampled uniformly in clock-time. We also demonstrate via Monte Carlo simulations that our approach provides accurate estimates of VPIN for all possible combinations of parameters. We illustrate the usefulness of the VPIN metric by estimating it for futures contracts on stock indices, currencies, commodities and rates. We also demonstrate that VPIN has important linkages with future volatility. Because toxicity is harmful for liquidity providers, high levels of VPIN should presage high volatility. We show that VPIN predicts short-term toxicity-induced volatility, particularly as it relates to large price moves.

Estimates of the toxicity of order flow have a number of immediate applications. Market makers, for example, can use the VPIN metric as a real-time risk management tool. In other research (see Easley, López de Prado, and O’Hara [2010]), we presented evidence that order flow as captured by the VPIN metric was becoming increasingly toxic in the hours before the May 6th 2010 ‘flash crash’, and that this toxicity contributed to the withdrawal of many liquidity providers from the market. Tracking the VPIN metric would allow market makers to control their risk and potentially remain active in volatile markets. Regulators could use VPIN to monitor the “quality” of liquidity provision, and pro-actively restrict trading or impose market controls. In a high frequency world, effective regulation needs to be done on an ex-ante basis, anticipating problems before, and not after, they lead to market breakdowns. Monitoring VPIN metric levels can signal when liquidity provision is at risk and allow for market halts or other regulatory actions to forestall crashes. Traders can also use measures based on the VPIN metric in designing algorithms to control execution risks. Microstructure models have long noted (see, for example, Admati and Pfleiderer [1988]) that intra-day seasonalities can reflect the varying participation rates of informed and uninformed traders. Designing algorithms to delay or accelerate trading depending on the VPIN metric may reduce the risk of executing orders when it is disadvantageous to do so.

Our analysis of order toxicity and its effects in high frequency markets is related to a variety of recent research. There is now a developing body of work looking at high frequency trading in a multiplicity of markets. Hendershott and Ryan [2009] present evidence on HF trading on the Deutsche Borse; Brodegaard [2010] and Hasbrouck and Saar (HS) [2010] analyze the role and strategies of high frequency traders in U.S. equity markets. Kirilenko, Kyle, Samadi, and Tuzun [2010] extensively characterize the behavior of HF traders and other market participants in the S&P futures market. There is also a developing literature looking at the more normative effects of computerized or HF trading on liquidity. Hendershott, Jones and Menkveld [2011] analyze the empirical relationship between algorithmic trading and liquidity, finding

---

5 See Clark [1973] or Ané and Geman [2000] for a primer on this vast field of research: Subordinated stochastic processes.

6 Among others, Tauchen and Pitts [1983], DeGennaro and Shrieves [1995], Jones et al. [1994].

7 See Kirilenko, Kyle, Samadi, and Tuzun [2010] for empirical evidence on market maker behavior during the flash crash.
that algorithmic trading improves liquidity for large stocks. Chaboud, Chiquoine, Hjalmarsson, and Vega [2009] provide a similar analysis of the effects of computerized trading in foreign exchange. These analyses complement recent theoretical research looking at the relation between liquidity and market fragility (see Brunnermeir and Pedersen [2009] and Huang and Wang [2011]).

Methodologically, our paper is related to research by Engle and Lange [2001] and Deuskar and Johnson [2011]. Engle and Lang proposed a market depth measure, VNET, which is calculated using order imbalance measured over price change increments. Our analysis here is calculated over volume increments (or buckets), but both their analysis and ours depart from standard time-based approaches to analyze the effects of asymmetric information in dynamic market environments. Deuskar and Johnson also analyze order flow imbalance in futures markets. These authors estimate the flow-driven component of systematic risk and its dynamic properties. Our focus is not on asset pricing issues, but their finding that flow-driven risk accounts for over half of the risk in the market portfolio underscores our argument that order flow imbalance (toxicity) has important effects on market behaviour and performance.

This paper is organized as follows. Section 1 discusses the theoretical framework and shows how PIN impacts the bid-ask spread. Section 2 presents our procedure for estimating the VPIN metric. Section 3 evaluates the accuracy of the VPIN metric via Monte Carlo simulations. Section 4 estimates the VPIN metric for contracts representative of various asset classes: Equity indices, currencies, commodities and rates. Section 5 discusses predictive properties of the VPIN metric for volatility. VPIN forecasts toxicity-induced volatility, as illustrated in Section 6. Section 7 summarizes our findings. A Technical Appendix presents the pseudocode for computing the VPIN metric, and its Monte Carlo accuracy.

1. THE MODEL

In this section we describe the basic model that allows us to use data about order flow to infer the toxicity of the orders. We begin with a standard microstructure model in which we derive our measure of flow toxicity, PIN, and we then show how to modify PIN to achieve a better fit with current markets. Readers conversant with the standard PIN approach can proceed directly to Section 2.

In a series of papers, Easley et. al. demonstrate how a microstructure model can be estimated for individual stocks using trade data to determine the probability of information-based trading, PIN. Microstructure models view trading as a game between liquidity providers and traders (position takers) that is repeated over trading periods $i=1,...,I$. At the beginning of each period, nature chooses whether an information event occurs. These events occur independently with probability $\alpha$. If the information is good news, then informed traders know that by the end of the trading period the asset will be worth $S_i$; and, if the information is bad news, that it will be worth $\bar{S}_i$, with $\bar{S}_i > S_i$. Good news occurs with probability $(1-\delta)$ and bad news occurs with the remaining probability, $\delta$. After an information event occurs or does not occur, trading for period $i$ begins with traders arriving according to Poisson processes throughout the trading period. During periods with an information event, orders from informed traders arrive at rate $\mu$. These informed traders buy if they have seen good
news, and sell if they have seen bad news. Every period orders from uninformed buyers and uninformed sellers each arrive at rate $\varepsilon$.

The structural model allows us to relate observable market outcomes (i.e. buys and sells) to the unobservable information and order processes that underlie trading. The previous literature focuses on estimating the parameters determining these processes via maximum likelihood. Intuitively, the model interprets the normal level of buys and sells in a stock as uninformed trade, and it uses that data to identify the rate of uninformed order flow, $\varepsilon$. Abnormal buy or sell volume is interpreted as information-based trade, and it is used to identify $\mu$. The number of periods in which there is abnormal buy or sell volume is used to identify $\alpha$ and $\delta$.

A liquidity provider uses his knowledge of these parameters to determine the price at which he is willing to go long, the Bid, and the price at which he is willing to go short, the Ask. These prices differ, and so there is a Bid-Ask Spread, because the liquidity provider does not know whether the counterparty to his trade is informed or not. This spread is the difference in the expected value of the asset conditional on someone buying from the liquidity provider and the expected value of the asset conditional on someone selling to the liquidity provider. These conditional expectations differ because of the adverse selection problem induced by the possible presence of better informed traders.

As the trading period progresses, liquidity providers observe trades and are modeled as if they use Bayes rule to update their beliefs about the toxicity of the order flow, which in our model is described by the parameter estimates. Let $P(t) = (P_n(t), P_b(t), P_g(t))$ be a liquidity provider’s belief about the events "no news" (n), "bad news" (b), and "good news" (g) at time t. His belief at time 0 is $P(0) = (1-\alpha, \alpha\delta, \alpha(1-\delta))$.

To determine the Bid or Ask at time $t$, the liquidity provider updates his beliefs conditional on the arrival of an order of the relevant type. The time $t$ expected value of the asset, conditional on the history of trade prior to time $t$, is

$$E[S_t | t] = P_n(t)S_t^* + P_b(t)S_t + P_g(t)S_t,$$

where $S_t^* = \delta S_t + (1-\delta)\overline{S}$ is the prior expected value of the asset.

The Bid is the expected value of the asset conditional on someone wanting to sell the asset to a liquidity provider. So it is

$$B(t) = E[S_t | t] - \frac{\mu P_n(t)}{\varepsilon + \mu P_n(t)} [E[S_t | t] - \overline{S}] .$$

---

8 The literature has considered various more complex models of the arrival process, but to illustrate our ideas we stay with the simplest model. The simple model has an advantage over more complex models in that it yields a simple expression for the probability of information-based trade which is also easy to compute. In spite of its simplicity and its obvious abstraction from the reality of the trading process this expression has proven useful in a variety of settings.
Similarly, the Ask is the expected value of the asset conditional on someone wanting to buy the asset from a liquidity provider. So it is

\[ A(t) = E[S_t | t] + \frac{\mu P_x(t)}{\epsilon + \mu P_x(t)} [S_t - E[S_t | t]] \]

These equations demonstrate the explicit role played by arrivals of informed and uninformed traders in affecting quotes. If there are no informed traders (\( \mu = 0 \)), then trade carries no information, and so the Bid and Ask are both equal to the prior expected value of the asset. Alternatively, if there are no uninformed traders (\( \epsilon = 0 \)), then the Bid and Ask are at the minimum and maximum prices, respectively. At these prices no informed traders will trade either, and the market, in effect, shuts down. Generally, both informed and uninformed traders will be in the market, and so the Bid is below \( E[S_t | t] \) and the Ask is above \( E[S_t | t] \).

The Bid-Ask Spread at time \( t \) is denoted by \( \Sigma(t) = A(t) - B(t) \). Calculation shows that this spread is

\[ \Sigma(t) = \frac{\mu P_x(t)}{\epsilon + \mu P_x(t)} [S_t - E[S_t | t]] + \frac{\mu P_x(t)}{\epsilon + \mu P_x(t)} [E[S_t | t] - S_t] \]

The first term in the spread equation is the probability that a buy is an information-based trade times the expected loss to an informed buyer, and the second is a symmetric term for sells. The spread for the initial quotes in the period, \( \Sigma_i \), has a particularly simple form in the natural case in which good and bad events are equally likely. That is, if \( \delta = 1 - \delta \) then

\[ \Sigma = \frac{\alpha \mu}{\alpha \mu + 2 \epsilon} [S_t - S_i] \]

The key component of this model is the probability that an initial order is from an informed trader; the probability of information-based trade is known as PIN. It is straightforward to show that the probability that the opening trade in a period is information-based is given by

\[ PIN = \frac{\alpha \mu}{\alpha \mu + 2 \epsilon} \]

where \( \alpha \mu + 2 \epsilon \) is the arrival rate for all orders and \( \alpha \mu \) is the arrival rate for information-based orders. PIN is thus a measure of the fraction of orders that arise from informed traders relative to the overall order flow, and the spread equation shows that it is the key determinant of spreads.

These equations illustrate the fact that liquidity providers need to correctly estimate their PIN in order to identify the optimal levels at which to enter the market. An unanticipated increase in PIN will result in losses to those liquidity providers who don’t adjust their prices.

2. VPIN and the ESTIMATION OF PARAMETERS
The standard approach to computing the PIN model uses maximum likelihood to estimate the unobservable parameters \((\alpha, \mu, \delta, \epsilon)\) driving the stochastic process of trades and then derives PIN from these parameter estimates. In this section, we propose an alternative approach which provides a direct analytic estimation of PIN that does not require the intermediate numerical estimation of non-observable parameters. It also has the virtue of being updated in stochastic time, matching the speed of arrival of new information to the marketplace. This volume-based approach, which we term VPIN, provides a simple metric for measuring order toxicity in a high frequency environment. First, we begin with a discussion of the role of information and time in high frequency trading.

2.1. THE NATURE OF INFORMATION AND TIME
Information in the standard sequential trade model is generally viewed as data that is informative about the future value of the asset. Thus, in an equity market setting it is natural to view information as being about the future payoffs of the asset, where this information may relate to the prospects of the company, the market for its products, etc. In an efficient market, the value of the asset should converge to its full information value as informed traders seek to profit on their information by trading. Because market makers can be long or short the stock, future movements in the value of the asset affect their profitability and so they attempt to infer any underlying new information from the patterns of trade. It is their updated beliefs that are impounded into their bid and ask prices.

In a high frequency world, market makers face the same basic problem, although the horizon under which they operate changes things in interesting ways. A high frequency market maker who anticipates holding the stock for minutes is affected by information that influences its value over that interval. This information may be related to underlying asset fundamentals, but it may also reflect factors related to the nature of trading in the overall market or to the specifics of liquidity demand over a particular interval. For example, in a futures contract, information that induces increased hedging demand for a contract will generally influence the futures price, and so is relevant for a market maker. This broader definition of information means that information events may occur frequently during the day, and they may have varying importance for the magnitude of future price movements. Nonetheless, their existence is still signaled by the nature and timing of trades.

The most important aspect of high frequency modeling is that trades are not equally spaced in terms of time. Trades arrive at irregular frequency, and some trades are more important than others as they reveal differing amounts of information. For example, as Figure 1 shows, trading in E-mini S&P 500 futures (the blue curve and scale on the left side of the graph) and EUR/USD futures (the red curve and scale on the right side of the graph) exhibit a different intraday seasonality. The arrival of new information to the marketplace triggers waves of decisions that translate into volume bursts. Information relevant to different products arrives at different times, thus generating distinct intraday volume seasonalties.
Easley and O’Hara [1992b] developed the idea that the time between trades was correlated with the existence of new information, providing a basis for looking at trade time instead of clock time. In this study, rather than modeling clock-time, we take the different approach of working in volume-time. It seems sensible that the more relevant a piece of news is, the more volume it attracts. By drawing a sample every occasion the market exchanges a constant amount of volume, we attempt to mimic the arrival to the market of news of comparable relevance. If a particular piece of news generates twice as much volume as another piece of news, we will draw twice as many observations, thus doubling its weight in the sample.

3.2. VOLUME BUCKETING

In the example above, if we draw one E-mini S&P 500 futures sample every 200,000 traded contracts, we will draw on average about 9 samples per day. On very active days, we will draw a large multiple of 9, while inactive days will contribute fewer data points. Once again note that, since the EC1 Currency contract trades about 1/10 of E-mini S&P 500 futures’ daily volume on average, targeting 9 draws per day will require reducing the volume-distance between two observations to about 20,000 contracts. Because of their differing intra-day patterns of trade, by the time we draw our first E-mini S&P 500 futures observation of the day, we are about to draw our fourth observation of the day for EC1 Currency.

To implement this volume dependent sampling, we group sequential trades into equal volume buckets of an exogenously defined size $V$. A volume bucket is a collection of trades that, combined, add up to a volume $V$. If the last trade needed to complete a

---

9 A variety of authors have further developed this notion of time as an important characteristic of trading. Of particular importance, Engle [1996] and Engle and Russell [2005] develop the role of time in a new class of autoregressive-conditional duration (ACD) models.
bucket is for a size greater than needed, the excess size is given to the next bucket. We let $\tau=1,\ldots,n$ be the index of equal volume buckets. A detailed algorithm for this volume sorting process is presented in the Appendix. Sampling by volume buckets allows us to divide the trading session into trading periods over which trade imbalances have a meaningful economic impact on the liquidity providers.

2.3. BUY VOLUME AND SELL VOLUME CLASSIFICATION

An issue we have not yet addressed is how to distinguish buy volume and sell volume. Recall that signed volume is necessary because of its potential correlation with order toxicity. While the overall level of volume signals the possible presence of new information, the direction of the volume signals its composition. Thus, a preponderance of buy (sell) volume would suggest toxicity arising from the presence of good (bad) information.

Microstructure research has generally relied on tick-based algorithms to sign trades. Trade classification, however, has always been problematic. One problem is that reporting conventions in markets could treat orders differently depending upon whether they were buys or sells. The NYSE, for example, would report only one trade if a large sell block crossed against multiple buy orders on the book, but would report multiple trades if it were a large buy block crossing against multiple sell orders. Similarly, splitting large orders into multiple small orders meant that trades occurring in short intervals were not in fact independent observations. Aggregating trades on the same side of the market over short intervals into a single observation was the convention empirical researchers used to deal with these problems.

A second difficulty is that signing trades also requires relating the trade price to the prevailing quote. Traders taking the market maker’s bid (ask) were presumed to be sellers (buyers), and trades falling in between were signed using a tick-based algorithm. The Lee-Ready [1991] algorithm also suggested using a 5 second delay between the reported quote and trade price to reflect the fact that the mechanism reporting quotes to the tape was not the same as the trade-reporting mechanism. Even in the simpler world of specialist trading, trade classification errors were substantial.

In a high frequency setting, trade classification is much more difficult. In the futures markets we investigate, there is no specialist, and liquidity arises from an electronic order book containing limit orders placed by a variety of traders. In this electronic market, an informed trader could hit the current quote or could submit a limit order that improves the current quote or even one that enters the book behind the current quote. If an order subsequently hits this informed trader’s quote, the informed trader will be taking the opposite side of the trade from the one that would be inferred from a tick rule. Additionally, order splitting is the norm, cancellations of quotes and orders are rampant, and the sheer volume of trades is overwhelming. Using S&P e-mini futures data from May 2010, we found that an average day featured 2,650,391 quote changes due to order additions or cancellations, and 789,676.2 quote changes due to

---

10 The nature of informed and uninformed trading in an electronic limit order market has been examined by numerous authors, see for example, Foucault, Kadan and Kandel [2005], Bloomfield, O’Hara, and Saar [2005], and Foucault, Kadan and Kandel [2009].

11 Over the last three years of E-mini S&P500 futures data, for example, in an average 10 minute period there are were 2200 trades encompassing 21,000 contracts traded, with their respective standard deviations of 4000 trades and 3000 contracts.
trades. Because the Best-Bid-or-Offer (BBO) changes several times per trade, many contracts exchanged at the same price in fact occurred against the bid and the offer. In a high frequency world, applying standard tick-based algorithms over individual transactions is futile.

In our analysis, we aggregate trades over short time intervals and then use the price change between the beginning and end of the interval to sign the aggregate volume in the time interval. Aggregation mitigates the effects of order splitting, and using the price change gives a less noisy metric to determine the tick change. We calculate volumes using one-minute time bars, although our analysis appears to work equally well with other time aggregations. This methodology will misclassify some volume. Our goal is not to correctly classify all trades (a hopeless exercise in any case), but rather to develop an indicator of trade imbalance that is useful for creating a measure of toxicity. We show later in the paper that our aggregation procedure dominates the use of raw transaction data in signing volume for purposes of estimating order toxicity.

2.4. VOLUME-SYNCHRONIZED PROBABILITY OF INFORMED TRADING (THE VPIN INFORMED TRADING METRIC)

The standard PIN model looks only at the number of buy and sell trades to infer information about the underlying information structure; there is no explicit role for volume. In the high frequency markets we analyze, the number of trades is problematic. Going back to the theoretical foundation for PIN, what we really want is information about trading intentions that arise from informed or uninformed traders. The link between these trading intentions and transactions data is very noisy as trading intentions may be split into many pieces to minimize market impact, one order may produce many trade executions, and information-based trades may be done in various order forms. For these reasons, we treat each reported trade as if were an aggregation of trades of unit size. So, for example, a trade for five contracts at some price \( p \) is treated as if it were five trades of one contract each at price \( p \). Note that this convention explicitly puts trade intensity into the analysis. Next, we turn to how we link these trades to the theory.

We know from Easley, Engle, O’Hara and Wu [2008] that the expected trade imbalance is \( E[|V_{\tau}^S - V_{\tau}^B|] \approx \alpha \mu \) and that the expected arrival rate of total trades is:

\[
\frac{1}{n} \sum_{\tau=1}^{n} (V_{\tau}^B + V_{\tau}^S) = V = \alpha (1 - \delta)(\xi + \mu + \epsilon) + \alpha \delta (\mu + \xi + \epsilon) + (1 - \alpha)(\epsilon + \xi) = \alpha \mu + 2 \epsilon
\]

Volume bucketing allows us to estimate this specification very simply. In particular, recall that we divide the trading day into equal-sized volume buckets throughout the day. That means that \( V_{\tau}^B + V_{\tau}^S \) is constant, and it is equal to \( V \), for all \( \tau \).

From the values computed above, we can derive the Volume-Synchronized Probability of Informed Trading, the VPIN informed trading metric, as

\[
VPIN = \frac{\alpha \mu}{\alpha \mu + 2 \epsilon} = \frac{\alpha \mu}{V} \approx \frac{\sum_{\tau=1}^{n} |V_{\tau}^S - V_{\tau}^B|}{nV}
\]
Estimating the VPIN metric requires choices of \( V \), the amount of volume that will be in every bucket, and \( n \), the number of buckets used to approximate the expected trade imbalance and intensity. As an initial specification, we focus on \( V \) equal to one-fiftieth of the average daily volume. If we then chose \( n = 50 \), we will calculate the VPIN metric over 50 buckets, which on a day of average volume would correspond to finding a daily VPIN. Our results are robust to a wide range of choices of \( V \) and \( n \) as we discuss in Section 6.

The VPIN metric is updated after each volume bucket. Thus, when bucket 51 is filled, we drop bucket 1 and calculate the new VPIN based on buckets 2 – 51. We update the VPIN metric in volume-time for two reasons. First, we want the speed at which we update VPIN to mimic the speed at which information arrives to the marketplace. We use volume as a proxy for the arrival of information to accomplish this goal. Second, we would like each update to be based on a comparable amount of information. Volume can be very imbalanced during segments of the trading session with low participation, and in such low-volume segments it seems unlikely that there is new information. So updating the VPIN metric in clock-time would lead to updates based on heterogeneous amounts of information.

As an example, consider the trading of the E-mini futures on May 6th 2010. Volume on this date (remembered for the ‘flash crash’ that took place) was extremely high, so our procedure produces 137 estimations of the VPIN metric, compared to the average 50 daily estimations. Because our sample length \( (n) \) is also 50, the time range used for some estimations of the VPIN metric on May 6th 2010 was only a few hours, compared to the average 24 hours.

![Figure 2 – Time gap between first and last data point used to estimate each new VPIN metric update in the E-mini futures](image)
Figure 2 illustrates the way time ranges become “elastic”, contingent on the *trade intensity* (a proxy for speed of information arrival). At 9:30am, the data used to compute VPIN covered almost an entire day. But as the New York Stock Exchange opened on May 6th 2010, our algorithm updated the VPIN metric more frequently and based its estimates on a shorter interval of clock-time. By 12:17pm, VPIN was being computed looking back only one-half of a day. Note how reducing the time period covered by the sample did not lead to noisier estimations. On the contrary, the VPIN metric kept changing following a continuous trend. The reason is that time ranges do not contain comparable amounts of information. Instead, it is volume ranges that produce comparable amounts of information per update.12

GARCH specifications provide an alternative way to deal with the volatility clustering typical of high frequency data sampled in clock-time. Working in volume-time reduces the impact of volatility clustering, since we produce estimates based on samples of equal volume. Because large price moves are associated with large volumes, sampling by volume can be viewed a proxy for sampling by volatility. The result is a collection of observations that follows a distribution that is closer to normal and is less heteroskedastic than it would be if we sampled uniformly in clock-time. Thus, working in volume-time can be viewed as a simple alternative to employing a GARCH specification.

To see how this transformation allows a partial recovery of normality, we consider an E-mini S&P 500 futures price sample from January 1st 2008 to October 22nd 2010. We draw an average of 50 observations a day, equally spaced by time in the first case (chronological time), and equally spaced by volume in the second (volume time). Next, we compute first order differences and standardize each sample. Both samples are negatively skewed and have fat tails, but the volume-time sample is much closer to normal, exhibits less serial correlation, and is less heteroskedastic13. This becomes more obvious as the sampling frequency increases (compare tables for 50 and 100 draws per day). Table 1 provides the resulting statistics and Figure 3 provides a graphical illustration of the normalized price changes.

---

12 At a daily level the relationship between volume and price change (which is a proxy for information arrival) has been explored by many authors including Clark [1973], Tauchen and Pitts [1983] and Easley and O’Hara [1992b].

13 \[ L - B^* = T(T + 2) \sum_{i=1}^{10} \rho_{t,t-i}^2 / T - i \], where \( \rho_{t,t-i} \) is the sample autocorrelation at lag \( i \). Both samples have the same number of observations \( T \). For clarity, we have not factored these statistics: *White* is the \( R^2 \) of regressing the squared series against all cross-products of the first 10-lagged series, and 

\[ J - B^* = 1 \left[ s^2 + k^2 / 4 \right] \], where \( s \) is skewness and \( k \) is excess kurtosis.
3. ACCURACY OF THE VPIN METRIC ESTIMATE
3.1. THE APPROXIMATION

The VPIN metric is an approximation of the underlying PIN model. Before turning to its estimation, it is useful to determine the statistical properties of this approximation. Suppose that buys and sells are generated by the trade model in Section 2 using the parameters \((\alpha, \delta, \mu, \varepsilon) = (2/5, 2/5, 10, 3)\). The distribution of the number of buys and sells that this process generates is shown in Figure 4. The probability is concentrated in three areas: No Event, Event Uptick and Event Downtick. The model predicts that events will trigger bursts of trading and that a large number of sells is more likely than a large number of buys in this parameterization \(\delta < 1/2\).

This distribution of trades results from the mixture of probabilities in the model described in Section 2. Assuming that order arrivals follow a Poisson process with parameters \((\mu, \varepsilon)\), we can derive a general mixed distribution by simply combining the various outcomes with their event probabilities \((\alpha, \delta)\):

<table>
<thead>
<tr>
<th>Stats (50)</th>
<th>Chrono time</th>
<th>Volume time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>StDev</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Skew</td>
<td>-0.0788</td>
<td>-0.2451</td>
</tr>
<tr>
<td>Kurt</td>
<td>31.7060</td>
<td>15.8957</td>
</tr>
<tr>
<td>Min</td>
<td>-21.8589</td>
<td>-20.6117</td>
</tr>
<tr>
<td>Max</td>
<td>19.3092</td>
<td>13.8079</td>
</tr>
<tr>
<td>L-B*</td>
<td>34.4551</td>
<td>22.7802</td>
</tr>
<tr>
<td>White*</td>
<td>0.0971</td>
<td>0.0548</td>
</tr>
<tr>
<td>J-B*</td>
<td>34.3359</td>
<td>6.9392</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stats (100)</th>
<th>Chrono time</th>
<th>Volume time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>StDev</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Skew</td>
<td>-0.1606</td>
<td>-0.4808</td>
</tr>
<tr>
<td>Kurt</td>
<td>44.6755</td>
<td>23.8651</td>
</tr>
<tr>
<td>Min</td>
<td>-28.3796</td>
<td>-29.2058</td>
</tr>
<tr>
<td>Max</td>
<td>24.6700</td>
<td>15.5882</td>
</tr>
<tr>
<td>L-B*</td>
<td>115.3207</td>
<td>36.1189</td>
</tr>
<tr>
<td>White*</td>
<td>0.0873</td>
<td>0.0370</td>
</tr>
<tr>
<td>J-B*</td>
<td>72.3729</td>
<td>18.1782</td>
</tr>
</tbody>
</table>

Table 1 – Results of sampling by Chronological time versus Volume time

Figure 3 – Statistics and PDFs on the E-mini S&P 500 futures, sampled by regular time intervals and regular volume intervals
\[ P(V^B, V^S) = (1 - \alpha)P(V^B, \varepsilon)P(V^S, \varepsilon) + \alpha[\delta P(V^B, \varepsilon)P(V^S, \mu + \varepsilon) + (1 - \delta)P(V^B, \mu + \varepsilon)P(V^S, \varepsilon)] \]

Figure 4 – The bivariate distribution of volume for buys and sells

PIN is a statistic based on these same parameters, which summarizes the information contained in this distribution into a probability of a quote being hit by an informed trader. In this example, it has an exact value of

\[ PIN = \frac{\alpha \mu}{\alpha \mu + 2\varepsilon} = \frac{1}{6}. \]

Of course, in reality we don’t know the values of these parameters, and so we don’t know PIN or its volume-synchronized version VPIN. Instead, we derive the VPIN metric from the observed trade intensity and imbalance between buys and sells. In particular, we approximate the arrival rates which allows us to derive an approximation of the probability of volume-synchronized, information-based trade of

\[ VPIN = \frac{\alpha \mu}{\alpha \mu + 2\varepsilon} \approx \frac{\sum_{t=1}^{n} |V^S_t - V^B_t|}{nV}. \]

We discuss the accuracy of this approximation in the next section.

3.2. MONTE CARLO VALIDATION

Although three parameters appear in the formulation of the VPIN metric, only two need to be simulated. This observation follows from the facts that \( \varepsilon = \frac{V - \alpha \mu}{2} \) and that in our methodology \( V = \frac{1}{n} \sum_{t=1}^{n} (V^B_t + V^S_t) \) is constant. Therefore we can measure the accuracy of our VPIN metric estimate by running a bivariate Monte Carlo on only the parameters \((\alpha, \mu)\).
We generate nodes of \((\alpha, \mu)\) scanning the entire range of possible parameters, \(\alpha \in [0,1]\) and \(\mu = [0, V]\), with \(V = 100\), and \(n = 50\) (where \(n\) is the number of volume buckets used to approximate the expected trade imbalance and intensity). The arrival rate of uninformed trade \(\varepsilon\) is then determined at each node, using the aforementioned expression. Ten equidistant partitions are investigated per dimension (\(11^2\) nodes), carrying out 100,000 simulations of flow arrival per node (\(11^2 \times 10^5\) simulations in total). This yields a Monte Carlo accuracy of the order of \(10^{-4}\).

### Table 2 – Monte Carlo results for \(n=50\)

Table 2 shows that there appears to exist a negligible bias, on the order of \(10^{-3}\), towards underestimating the true value of the VPIN metric. Figures 5 and 6 illustrate the accuracy of approximation graphically. These results motivate our use of \(n=50\) in our estimation of VPIN.

\[14\] See Appendix 2 for additional details.
3.3. THE STABILITY OF VPIN ESTIMATES

The choice of how to classify trades has a more important effect on the estimate of VPIN than does the approximation discussed above. In particular, because VPIN involves looking at trade imbalance and intensity, aggregating over time bars would be expected to reduce the noise in this variable as well as to rescale it. In Easley, Lopez de Prado, and O’Hara [2011a], we argued that this necessitates looking at the
relative levels of VPIN as captured by their cumulative distribution function rather than at absolute levels of VPIN.

This point can be illustrated by looking at the behavior of VPIN on a given day for different trade classification algorithms. The “flash crash” on May 6 is of particular importance in futures (and equity) markets and so we illustrate the behavior of VPIN on this day using three alternative trade classification schemes. Figure 7(a) shows VPIN calculated using one-minute time bars; Figure 7(b) uses 10-second bars; and Figure 7(c) uses individual transactions.

![Figure 7(a) – VPIN estimated with One-Minute Time Bars](image)

![Figure 7(b) – VPIN Estimated with Ten-Second Time Bars](image)
These three graphs illustrate the effects of aggregation on the absolute levels of VPIN; as expected, it declines significantly when the size of the time bar is reduced. They also illustrate that on the day of the flash crash VPIN reached unusually high levels as measured by the estimated VPIN relative to its CDF (the dashed line in each graph). The one-minute and ten-second time bars produce qualitatively similar stories about the relative level of VPIN. In both cases, it rises before the crash and stays high throughout the rest of the day. The trade-by-trade classification results are different. Here the estimated VPIN increases before the crash, although not as dramatically as it does with time bars, but it falls to unusually low levels immediately after the crash and it remains low for the rest of the day. This inertia, however, is over a period in which the market rose by more than 4%. It seems highly unlikely that volume was in fact balanced over this period. We suspect that this is more a result of trade misclassification than anything else, and so we do not use trade-by-trade classification. Instead, we report results for one-minute time bars while noting that the relative results with ten-second time bars are nearly identical.

3.4. TOXICITY AND TRADING INTENSITY
As discussed in Section 2.1, volume buckets divide the trading session into trading periods over which trade imbalances have a meaningful economic impact on the liquidity providers. That a bucket (or trading period) is completed over a short period of time is indicative that critical or unexpected news has arrived into the market. For example, Figure 2 showed that the time range of buckets added to the sample was decreasing as the moment of the ‘flash crash’ was approaching on May 6th, 2010.

Each bucket has the same volume by construction, but the number of time bars that are used to fill a bucket varies. The fewer the number of time bars covered in a bucket, the greater is the trading intensity. Note that in our model of the trade process
an increase in the trade intensity signals to market makers an increased posterior likelihood of an information event having occurred. Thus, the higher the trade intensity (or equivalently, the fewer time bars needed to fill a bucket) the higher will be VPIN. For instance, consider the (admittedly unrealistic) case that one bucket is (unexpectedly) completely filled by a single time-bar, moving the E-mini S&P500’s price down 20 ticks from 1,000 to 995; our algorithm will label all of that volume as “sell”. Even if the next time-bar is several upticks for the same volume, market makers are not unaffected. Their estimates of the toxicity of the order flow will be increased and the spread will widen. More practically, the asymmetric payoff that market makers face may force them to quickly take losses before the price turns around (Easley, López de Prado and O’Hara [2011b]). In fact, this scenario is extremely toxic from a market maker’s perspective, as each unexpected offsetting time-bar may not allow them to recover losses, but rather compounds them.

4. ESTIMATING THE VPIN METRIC ON FUTURES
Having established the robustness of our estimation procedures for the VPIN metric, we now investigate its properties for a variety of contracts. The sample consists of 1-minute bars (time stamp, traded price, traded volume) for five futures contracts: The E-mini S&P 500 (CME); T-Note (CBOT); EUR/USD (CME); Brent (ICE); Silver (COMEX) futures. Our sample period is January 1st 2008 to October 30th 2010, using at each point in time the expiration with highest daily volume, rolled forward. This data is divided into an average of 50 equal volume buckets per day (V). Parameters are then estimated on a rolling window of sample size n = 50 (equivalent to 1-day of volume on average).

The VPIN metric is estimated as

\[
VPIN = \frac{\alpha \mu}{\alpha \mu + 2 \epsilon} \approx \frac{1}{nV} \sum_{r=1}^{50} |V_r^S - V_r^B|
\]

where \( r=1,\ldots,50 \) are the equal volume buckets, \( V_r^S \) is the volume classified as traded against the bid in bucket \( r \) and \( V_r^B \) is the volume classified as traded against the offer in bucket \( r \). Table 3 provides basic statistics of the VPIN metric estimates for these contracts.

<table>
<thead>
<tr>
<th>Stat</th>
<th>S&amp;PP500</th>
<th>T-Note</th>
<th>EUR/USD</th>
<th>Brent</th>
<th>Silver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.3933</td>
<td>0.4012</td>
<td>0.3269</td>
<td>0.3837</td>
<td>0.4114</td>
</tr>
<tr>
<td>StDev</td>
<td>0.0669</td>
<td>0.0874</td>
<td>0.0812</td>
<td>0.0646</td>
<td>0.0899</td>
</tr>
<tr>
<td>Skew</td>
<td>0.9675</td>
<td>0.7264</td>
<td>1.3128</td>
<td>0.6915</td>
<td>1.1877</td>
</tr>
<tr>
<td>Ex. Kurt</td>
<td>2.7965</td>
<td>0.9438</td>
<td>2.7368</td>
<td>1.0020</td>
<td>2.1465</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.9921</td>
<td>0.9953</td>
<td>0.9958</td>
<td>0.9921</td>
<td>0.9952</td>
</tr>
<tr>
<td>#Observ.</td>
<td>36914</td>
<td>36914</td>
<td>35165</td>
<td>36915</td>
<td>36914</td>
</tr>
</tbody>
</table>

Table 3 – Statistics of historical Volume-Synchronized Probability of Informed Trading (VPIN)

---

\(^{15}\) This follows directly from the Bayesian update of the estimate of the probability of an information event having occurred. See Easley and O’Hara (1987). That is, both increased (unexpected) balanced trade and increased unbalanced trade lead to an increased estimate of the probability of information based trade.
4.1. S&P 500 (CME)
Figure 8 shows the evolution of the E-mini S&P 500 futures contract (red line, expressed in terms of market value) and its VPIN metric value (green line). What is apparent is that the VPIN metric is generally stable, although it clearly exhibits substantial volatility. Of particular importance, the VPIN metric reached its highest level for this sample on May 6th 2010, providing a theoretical explanation for the qualitative observation that “liquidity evaporated”. Such high levels of the VPIN metric are consistent with virtually all order flow being one-sided for the equivalent of 1 day of transactions, and as discussed in Easley, López de Prado, and O’Hara [2011a], this excessive toxicity led market makers to become liquidity consumers rather than liquidity providers.

4.2. EUR/USD (CME)
May 6th 2010 also registered historically high flow toxicity for currencies as depicted in Figure 9. Currency futures are used for hedging purposes, but they are also used to implement macro bets for events such as competitive devaluations. This latter order flow is generally viewed as coming from informed traders such as hedge funds. The increasing levels of the VPIN metric since the beginning of the Greek crisis in January of 2010 are consistent with this view. The CFTC-SEC Staff Report noted that difficulties on May 6 first started as concerns in currency futures regarding problems in Greece, illustrating that in high frequency markets toxicity (and illiquidity) flow between markets.
4.3. T-NOTE (CBOT)

The highest VPIN metric value for the T-Note also occurred on May 6th 2010, although the level reached was not as extreme as in the previous two cases (see Figure 10). Note that there is an interesting seasonal effect in the T-Note contract around the end of the year. Specifically, volume in treasury markets slowly wanes every year after Thanksgiving, and the VPIN metric generally decreases. Thus, the lower volume is accompanied by a larger proportion of uninformed trading. What causes this seasonality is unclear, but one popular view among practitioners is that hedge fund traders generally close out their positions before the waning volume, reflecting that informed traders are “done for the year”.

Figure 9 - VPIN for EUR/USD futures

Figure 10 - VPIN for T-Note futures
4.4. BRENT CRUDE OIL (ICE)

Energy futures are also a venue in which market makers face extreme volatility in order flows. As shown in Figure 11, May 6th 2010 registered high flow toxicity, although to a lesser extent than in the case of stocks, currencies and rates. Such behavior is consistent with the fact that while the problems on May 6 were not energy related, these markets were affected by the transmission of liquidity and toxicity across markets.

![Figure 11 - VPIN for Brent Crude Oil futures](image)

4.5. SILVER (COMEX)

Metals such as silver or gold are typically viewed as markets with substantial information-based trading, and our estimates of VPIN in Table 3 are highest for the Silver contract. As we found with oil futures, May 6th 2010 was among the days with highest flow toxicity, although not the highest. Figure 12 does show that toxicity in this contract is increasing since the latter half of 2009, suggesting that market makers face increasing risk in liquidity provision.
VPIN AND TOXICITY-INDUCED VOLATILITY

VPIN was not created to forecast volatility but, as explained by microstructure theory, toxicity plays a role in explaining the origin of price changes and thus volatility. As order toxicity increases, market makers face potential losses and so may opt to reduce, or even abandon market making activities. This decrease in liquidity, in turn, suggests that high levels of VPIN should presage greater price volatility. In this section, we investigate this linkage. Our focus is on the very short-term effects of toxicity on volatility so we examine absolute returns over the following volume bucket. This linkage is also consistent with our hypothesis that in a high frequency context, volume-based volatility forecasts are more relevant for market makers (and traders) than their chronological-time counterparts.  

We look at three measures of this VPIN-volatility relationship: correlation, threshold correlation, and conditional probabilities. We have tested these relationships for the various asset classes and found consistent effects across all U.S. index futures. For brevity in exposition, we report results for the E-mini S&P500 Index futures for the period January 1st 2008 to November 26th 2010, only occasionally referring to other Index futures.

5.1. CORRELATION SURFACE

Deuskar and Johnson [2011] also look at the effects of order imbalance (which they term flow-driven risk) on market returns and volatility. Their analysis uses chronological time and so it is not directly comparable to what we do here. They find significant flow-driven effects on returns but note that their measure does not appear to be consistent with measuring the degree of asymmetric information. We believe this dimension is better captured by our volume-based analysis which incorporates the arrival of new information. However, both their model and ours provide strong evidence that order imbalance has important implications for market liquidity.
We begin by determining the impact that various combinations of the number of volume buckets per day and the number of buckets used to construct VPIN estimates have on the VPIN metric’s forecasting power over volatility. We first illustrate this relationship with the simple correlation between the natural logarithm of VPIN for a futures contract and the absolute return on the futures contract,

\[ \rho \left( \ln(VPIN_{\tau-1}), \frac{S_{\tau}}{S_{\tau-1}} - 1 \right) , \]

and we note that \( \tau \) indexes volume bars. We use absolute returns as a proxy for price volatility.

For E-mini S&P500 futures, this relationship is described in Figure 13. Each of the points in Figure 13 was computed using very large samples, some of over 37,000 observations. As is apparent from the figure, the correlation between VPIN and absolute future returns varies smoothly across different parameter values. The \((50, 250)\) combination seems a fairly good pair for E-mini S&P500, and it has a simple interpretation as “one week” of data (50 volume buckets per day and 5 trading days per week). For 50 buckets per day, 250 sample length, we obtain a correlation \( \rho \left( \ln(VPIN_{\tau-1}), \frac{R_{\tau}}{P_{\tau-1}} \right) = 0.213 \) on 37915 observations. For this sample \( \sigma_\rho = 0.005136 \), and 95% confidence band is given by \( \rho \left( \ln(VPIN_{\tau-1}), \frac{R_{\tau}}{P_{\tau-1}} \right) \in [0.203, 0.223]. \)

\[ 17 \text{ In the context of high frequency trading, such correlation is very significant. According to the Fundamental Law of Active Management (Grinold, 1989), } IR = IC \sqrt{BR}, \text{ where IR represents the } Information Ratio, IC the Information Coefficient (correlation between forecasts and realizations), and } BR \text{ the breadth (independent bets per year). Although a correlation of 20% may seem relatively small, the breadth of high frequency models is large, allowing these algorithms to achieve high Information Ratios. For example, an IR of 2 can be reached through a monthly model with IC of 0.58, or a weekly model with IC of 0.28, or a daily model with IC of 0.13. High frequency models produce more than one independent bet per day, thus a correlation of over 0.2 is particularly high.} \]
Figure 13 – Correlation between VPIN and future volatility for E-mini S&P500

Figure 13 suggests that there is little advantage in “over-fitting” these parameters, as there is a wide region of parameter combinations yielding similar predictive power. We use the (50,250) combination for computing the expected value of absolute returns conditional on the prior VPIN decile, $E \left[ \frac{p_t}{p_{t-1}} - 1 \middle| VPIN_{t-1} \right]$, finding the same positive relationship between greater VPINs and greater subsequent volatility.

Table 4 reports these correlations decile-by-decile for all of the contracts we consider. Note that for each contract, future volatility increases in almost all cases as we move through increasing VPIN deciles.

<table>
<thead>
<tr>
<th>VPIN(t-1) cut-off</th>
<th>ES1 Index</th>
<th>DM1 Index</th>
<th>RTA1 Index</th>
<th>NQ1 Index</th>
<th>FA1 Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.146%</td>
<td>0.112%</td>
<td>0.218%</td>
<td>0.164%</td>
<td>0.141%</td>
</tr>
<tr>
<td>2</td>
<td>0.151%</td>
<td>0.128%</td>
<td>0.187%</td>
<td>0.154%</td>
<td>0.157%</td>
</tr>
<tr>
<td>3</td>
<td>0.153%</td>
<td>0.130%</td>
<td>0.184%</td>
<td>0.167%</td>
<td>0.164%</td>
</tr>
<tr>
<td>4</td>
<td>0.160%</td>
<td>0.144%</td>
<td>0.182%</td>
<td>0.187%</td>
<td>0.163%</td>
</tr>
<tr>
<td>5</td>
<td>0.157%</td>
<td>0.146%</td>
<td>0.190%</td>
<td>0.184%</td>
<td>0.176%</td>
</tr>
<tr>
<td>6</td>
<td>0.158%</td>
<td>0.147%</td>
<td>0.200%</td>
<td>0.163%</td>
<td>0.164%</td>
</tr>
<tr>
<td>7</td>
<td>0.171%</td>
<td>0.155%</td>
<td>0.179%</td>
<td>0.168%</td>
<td>0.166%</td>
</tr>
<tr>
<td>8</td>
<td>0.173%</td>
<td>0.173%</td>
<td>0.189%</td>
<td>0.209%</td>
<td>0.193%</td>
</tr>
<tr>
<td>9</td>
<td>0.202%</td>
<td>0.216%</td>
<td>0.220%</td>
<td>0.243%</td>
<td>0.247%</td>
</tr>
<tr>
<td>10</td>
<td>0.324%</td>
<td>0.296%</td>
<td>0.333%</td>
<td>0.243%</td>
<td>0.257%</td>
</tr>
</tbody>
</table>

Table 4 – Prior VPINs effect on the following absolute return

Figure 14 provides a plot of the return on the E-mini S&P500 futures over the next 10 volume buckets (1/5 of an average day’s volume) sorted by the level of VPIN. The graph spreads out vertically suggesting that higher toxicity levels, as measured by higher VPIN, lead greater absolute returns.
5.2. THRESHOLD CORRELATION
To understand better the response of volatility to greater levels of the VPIN metric, we computed
\[ \rho \left( \ln(\text{VPIN}_{t-1}), \frac{S_t}{S_{t-1}} - 1 \right) \]
for within deciles of VPIN. The first subset contains observations within the first decile of VPIN, the second subset includes observations within the first two deciles, and so on. Table 5 provides additional support for the hypothesis that high absolute returns follow high values of VPIN.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Ln(VPIN) cut-off value</th>
<th>L-L Correl</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-1.06</td>
<td>0.01</td>
</tr>
<tr>
<td>0.2</td>
<td>-1.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>0.3</td>
<td>-0.99</td>
<td>-0.02</td>
</tr>
<tr>
<td>0.4</td>
<td>-0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.94</td>
<td>0.03</td>
</tr>
<tr>
<td>0.6</td>
<td>-0.92</td>
<td>0.04</td>
</tr>
<tr>
<td>0.7</td>
<td>-0.90</td>
<td>0.05</td>
</tr>
<tr>
<td>0.8</td>
<td>-0.86</td>
<td>0.08</td>
</tr>
<tr>
<td>0.9</td>
<td>-0.81</td>
<td>0.11</td>
</tr>
<tr>
<td>1.0</td>
<td>-0.51</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Table 5 – Lead-Lag Threshold Correlations for E-mini S&P500 futures

5.3. CONDITIONAL PROBABILITIES
A more revealing approach to understanding the relationship between lagged VPIN and volatility is to examine the relationship in terms of conditional probabilities. Table 6 counts the number of buckets in our sample that fell within each
\( (VPIN_{t-1}, \frac{P_t}{P_{t-1}} - 1) \) bin where VPIN levels have been grouped in 5%-tiles and absolute returns have been grouped into bins of size 0.25%.

Table 6 – Joint distribution of \((VPIN_{t-1}, \frac{P_t}{P_{t-1}} - 1)\)

<table>
<thead>
<tr>
<th>VPIN percentiles</th>
<th>Absolute return between two consecutive buckets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>81.21% 15.86% 2.33% 0.33% 0.13% 0.07% 0.07% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.10</td>
<td>79.15% 16.06% 3.66% 0.67% 0.40% 0.07% 0.07% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.15</td>
<td>80.48% 15.86% 2.53% 0.73% 0.13% 0.20% 0.07% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.20</td>
<td>80.23% 15.85% 3.66% 0.67% 0.07% 0.07% 0.00% 0.00% 0.07%</td>
</tr>
<tr>
<td>0.25</td>
<td>79.41% 16.72% 2.53% 1.00% 0.20% 0.07% 0.00% 0.00% 0.07%</td>
</tr>
<tr>
<td>0.30</td>
<td>79.48% 16.19% 3.26% 0.60% 0.47% 0.00% 0.00% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.35</td>
<td>79.61% 15.72% 3.46% 0.93% 0.07% 0.20% 0.00% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.40</td>
<td>78.56% 17.24% 3.06% 0.87% 0.13% 0.00% 0.13% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.45</td>
<td>77.68% 17.92% 3.46% 0.67% 0.07% 0.13% 0.07% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.50</td>
<td>78.01% 16.79% 3.00% 1.40% 0.27% 0.33% 0.07% 0.07% 0.07%</td>
</tr>
<tr>
<td>0.55</td>
<td>76.35% 17.72% 4.20% 1.20% 0.27% 0.13% 0.13% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.60</td>
<td>73.77% 20.11% 3.93% 1.80% 0.20% 0.00% 0.13% 0.07% 0.07%</td>
</tr>
<tr>
<td>0.65</td>
<td>72.75% 20.25% 4.93% 1.47% 0.33% 0.13% 0.07% 0.00% 0.07%</td>
</tr>
<tr>
<td>0.70</td>
<td>73.68% 19.72% 4.86% 1.20% 0.40% 0.00% 0.13% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.75</td>
<td>72.42% 21.39% 4.20% 1.27% 0.67% 0.07% 0.00% 0.00% 0.00%</td>
</tr>
<tr>
<td>0.80</td>
<td>72.84% 19.17% 4.86% 1.60% 0.93% 0.33% 0.13% 0.07% 0.07%</td>
</tr>
<tr>
<td>0.85</td>
<td>71.89% 19.39% 5.00% 1.93% 0.87% 0.27% 0.20% 0.13% 0.33%</td>
</tr>
<tr>
<td>0.90</td>
<td>62.29% 23.45% 8.39% 3.00% 1.20% 0.67% 0.27% 0.40% 0.33%</td>
</tr>
<tr>
<td>0.95</td>
<td>56.03% 25.18% 10.73% 4.60% 1.67% 0.87% 0.53% 0.13% 0.27%</td>
</tr>
<tr>
<td>1.00</td>
<td>47.44% 28.11% 12.92% 6.13% 2.86% 0.93% 0.53% 0.33% 0.73%</td>
</tr>
</tbody>
</table>

Table 7a – Prob \( (\frac{P_t}{P_{t-1}} - 1) | VPIN_{t-1} \)
Table 7b – Prob \left( \frac{P_t}{P_{t-1}} - 1 \right)\)

Easley, López de Prado and O’Hara [2011a] showed that 2 hours of persistently high levels of VPIN foreshadowed a 5% price collapse in E-mini S&P500 on May 6th 2010. These two tables present evidence of a more general result: high volatility tends to appear once the market’s tolerance for toxicity reaches a saturation point. This toxicity-induced volatility suggests that the VPIN metric can play a risk management role in a high frequency world.

The insurance value of VPIN can be best understood by looking at the upper quartile of the Prob \left( \frac{P_t}{P_{t-1}} - 1 \right) distribution. This region contains over 60% of all absolute returns greater than 0.75%. If VPIN had no insurance value against large price moves, that section of the conditional distribution would contain only 25% or less of the events. Table 8 illustrates this piece of the conditional distributions. The significance of these proportions (p) relative to what would be expected from a uniform distribution (u) is expressed in terms of their z-scores (z). In summary, for any reasonable significance level we can assert that the proportion of high absolute returns in the upper VPIN quartile is much greater than what would be derived from a uniform distribution.

18 For an explanation of the normal approximation to the binomial distribution, see for example Box, Hunter, Hunter and Hunter (1978), p.130.
The previous conditional probability analysis was conducted on our standard \((50,250)\) parameter combination. This choice of parameters maximized \(p \left( \ln(VPIN_{t-1}), \frac{P_t}{P_{t-1}} \right)\), but it is not necessarily the one that maximizes a particular “risk scenario”, say for example, \(\text{Prob} \left( \text{CDFVPIN}_{t-1} > \frac{3}{4} \left| \frac{P_t}{P_{t-1}} - 1 \right| > 0.75\% \right)\). The following figure shows the effect of parameter choices on VPIN’s insurance value against absolute returns of over 0.75%.

Table 8 – Upper quartile of the \(\text{Prob} \left( \frac{P_t}{P_{t-1}} - 1 \right)\) distribution

<table>
<thead>
<tr>
<th>VPIN Pctiles</th>
<th>1.00%</th>
<th>1.25%</th>
<th>1.50%</th>
<th>1.75%</th>
<th>&gt;1.75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>4.99%</td>
<td>8.24%</td>
<td>7.35%</td>
<td>5.26%</td>
<td>4.17%</td>
</tr>
<tr>
<td>0.85</td>
<td>6.03%</td>
<td>7.65%</td>
<td>5.88%</td>
<td>7.89%</td>
<td>14.58%</td>
</tr>
<tr>
<td>0.90</td>
<td>9.36%</td>
<td>10.59%</td>
<td>14.71%</td>
<td>10.53%</td>
<td>22.92%</td>
</tr>
<tr>
<td>0.95</td>
<td>14.35%</td>
<td>14.71%</td>
<td>19.12%</td>
<td>21.05%</td>
<td>12.50%</td>
</tr>
<tr>
<td>1.00</td>
<td>19.13%</td>
<td>25.29%</td>
<td>20.59%</td>
<td>21.05%</td>
<td>33.33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stats</th>
<th>1.00%</th>
<th>1.25%</th>
<th>1.50%</th>
<th>1.75%</th>
<th>&gt;1.75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>481</td>
<td>170</td>
<td>68</td>
<td>38</td>
<td>48</td>
</tr>
<tr>
<td>(p)</td>
<td>53.85%</td>
<td>66.47%</td>
<td>67.65%</td>
<td>65.79%</td>
<td>87.50%</td>
</tr>
<tr>
<td>(u)</td>
<td>25.00%</td>
<td>25.00%</td>
<td>25.00%</td>
<td>25.00%</td>
<td>25.00%</td>
</tr>
<tr>
<td>(z)</td>
<td>14.610</td>
<td>12.487</td>
<td>8.122</td>
<td>5.807</td>
<td>10.000</td>
</tr>
</tbody>
</table>

Figure 15 - \(\text{Prob} \left( \text{CDFVPIN}_{t-1} > \frac{3}{4} \left| \frac{P_t}{P_{t-1}} - 1 \right| > 0.75\% \right)\) for various parameter combinations (surface)
Figure 15 shows that a range of parameters with 250 or more buckets per day and a sample length of at least 300 are consistent with $\text{Prob} \left( CDF(\text{VPIN}_{t-1}) \left| \frac{p_t}{p_{t-1}} - 1 \right| \right)$ approaching 80%. Consequently, optimization of the VPIN parameter calculation can lead to significant forecasting ability over short term, large price movements.

6. TOXICITY-INDUCED VOLATILITY VS. GENERAL VOLATILITY
That the trading of an asset may be linked to its future volatility is an idea investigated by numerous authors. There is a large literature demonstrating linkages between trading volume and volatility (see, for example, Gallant et al [1992]; Conrad, Hameed, and Niden [1994]; Lo and Wang [2000]). Subsequent research by Chan and Fong [2000] examined the components of trading volume, concluding that “order imbalance plays a role in the volatility-volume relation”. Chordia, et al [2002] find that market returns are affected by daily order imbalances, while Deuskar and Johnson find strong effects of intraday order imbalance on volatility.

One of the conclusions of Easley and O’Hara (1996) is that asymmetric information may be a source of such volatility. This occurs because when market makers detect order flow toxicity they adjust the quotes at which they are willing to provide liquidity to accommodate an increased probability of adverse selection. However, not all sources of volatility are liquidity related. For example, volatility can be forecasted based on intraday seasonal patterns (NYSE opening and closing), or as a result of scheduled releases of information (economic releases, dividend announcements, auctions, corporate issuances, and the like). There is also a very large literature eschewing specific causes for volatility in favor of applying time series models to forecast volatility.

The purpose of VPIN is not to forecast volatility, and yet, some forecasting power should be expected if toxicity is one of multiple sources of volatility. As we demonstrated in the previous section, toxicity–based volatility is more related to large price movements, relating as it does to the unwillingness of high frequency market makers to provide liquidity when toxicity is high. It may be argued that VPIN is a worse predictor of volatility than a combination of indicators, for example VIX and volume. Our concern, however, is not with the best predictor of volatility, or for that matter volatility in general, but rather with liquidity-induced volatility, which the SEC-CFTC official study points out as the cause behind the ‘flash crash’. Consequently, it is a useful exercise to see how our toxicity-based VPIN measure differs from more standard volume-based linkages with large price movements.

It makes little sense in the high frequency domain to compare a time-based measure $\text{Prob} \left( v_{t-1} \left| \frac{p_t}{p_{t-1}} - 1 \right| \right)$, where $v_{t-1}$ is the volume in the prior time bar $t$, with our volume-based measure $\text{Prob} \left( \text{VPIN}_{\tau-1} \left| \frac{p_t}{p_{\tau-1}} - 1 \right| \right)$, where $\tau$ indexes volume bars. This is because market makers must adopt decisions in transactional time. Forecasts generated in equal time steps, regardless of the trade intensity, are of little use to
them. Moreover, such a calculation misses the point of how toxicity impacts volatility.

There is a comparison, however, that can make this distinctive role of toxicity-induced volatility clearer. VPIN is computed on equal volume buckets and so

$$\text{Prob}(\text{VPIN}_{t-1} \mid \frac{p_t}{p_{t-1}} - 1) = \text{Prob}(\frac{p_{t-1}}{\text{VPIN}_t} - 1)$$

as $V_t$ is a constant

The right-hand expression can be interpreted as capturing absolute returns per contract traded. We can use this same scaling approach to calculate $\text{Prob}(V_{t-1} \mid \frac{p_{t-1}}{p_t} - 1)$, i.e. the conditional probability of volume in the prior time bar (t-1), conditioned on absolute return per contract traded in the current time bar (t). This probability is not the same as $\text{Prob}(V_{t-1} \mid \frac{p_t}{p_{t-1}} - 1)$, as the scaled measure better captures the meaning of liquidity-induced volatility (i.e., larger expected absolute return per contract traded), while the latter is related to general volatility.

Table 9b reports $\text{Prob}(V_{t-1} \mid \frac{p_{t-1}}{p_t} - 1)$, i.e. the distribution of volume in the prior time bar conditional of the return per contract traded. This is the analogue to Table 7b, where we have replaced $\text{VPIN}_{t-1}$ (VPIN’s reading at volume bucket τ-1) with $V_{t-1}$ (volume traded in time bar t-1). The rows are the 5%-tiles of $V_{t-1}$, while the columns are the deciles of $\frac{p_{t-1}}{p_t} - 1$. For comparison to Table 7, we have also included $\frac{p_{t-1}}{V_{t-1}}$ in Table 9a.
The results from Table 9b are the opposite of those in Table 7b. While the highest returns (per contract traded) were preceded by VPINs within the upper quartile (see Table 8), we now find out that the highest returns per contract traded are preceded by volume within the lower quartile. Both VPIN and volume are positively correlated with higher volatility. However, when we look at the volatility per contract traded, it becomes evident that VPIN addresses a particular type of volatility, the one that is triggered by elevated readings of toxicity.

Because liquidity-induced volatility is a relatively rare event (although associated with dramatic episodes, like the ‘flash crash’), it is misleading to judge VPIN on the basis of its forecasting power over general volatility. However, to our knowledge, VPIN is the only high frequency gauge designed to measure toxicity, which we have confirmed triggers a particular type of volatility not explained by traditional indicators.

7. CONCLUSIONS
This paper discusses the connections between adverse selection, flow toxicity and PIN. After describing the theoretical framework underlying PIN, we explain the mechanism by which PIN impacts liquidity providers. Next, we present a new procedure to estimate the Volume-Synchronized Probability of Informed Trading, or the VPIN informed trading metric. A Monte Carlo study of our estimates reveals that they are accurate for any possible combination of parameters, even if the statistics are computed on relatively small samples.

An important advantage of the VPIN metric and its associated estimation procedure over previous toxicity measures and estimation procedures is that it is updated intraday with a frequency tuned to volume in order to match the speed of information arrival. It is also a direct analytic estimation procedure that does not rely on intermediate estimation of unobservable parameters, or numerical methods. We have shown that the VPIN metric has significant forecasting power over toxicity-induced volatility, and that it offers insurance value against future high absolute returns.
It is this latter property that we believe makes VPIN a risk management tool for the new high frequency world of trading. Liquidity provision is now a complex process, and levels of toxicity affect both the scale and scope of market makers’ activities. High levels of VPIN signify a high risk of subsequent large price movements, deriving from the effects of toxicity on liquidity provision. This liquidity-based risk is important both for market makers who directly bear the effects of toxicity, and for traders who face the prospect of liquidity-induced large price movements.\textsuperscript{19} We believe this is an important area for future research.

\textsuperscript{19} We stress that, in our view, VPIN is not a substitute for VIX, but rather a complementary metric for addressing a different risk. VIX captures the markets’ expectation of future volatility, and hence is useful for hedging the effects of risk on a portfolio’s return. VPIN captures the level of toxicity affecting liquidity provision, which in turn affects future short-run volatility when this toxicity becomes unusually high. For more discussion, see Easley, Lopez de Prado, O’Hara [2011b].
APPENDIX

A.1. ALGORITHM FOR COMPUTING THE VPIN METRIC

In this appendix, we describe the procedure to calculate Volume-Synchronized Probability of Informed Trading, a measure we called the VPIN informed trading metric. Similar results can be reached with more efficient algorithms, such as re-using data from previous iterations, performing fewer steps or in a different order, etc. But we believe the algorithm described below is illustrative of the general idea.

One feature of this algorithm to note is that we classify all trades within each one-minute time bar as either buys or sells using a tick test. We do not have data which directly identifies trades as buyer-initiated or seller-initiated so some classification procedure is necessary. One could classify each trade separately or one could classify trades in groups of an alternative size (based on either time or volume). Different schemes will lead to different levels of VPIN. We have used a variety of schemes and, as one would expect, cutting the data more finely leads to reduced levels of VPIN---measured trade becomes more balanced when groups of trades in say a one-minute time bar may be classified differently. However, our focus is on how rare a particular VPIN is relative to the distribution of VPINs derived from any classification scheme, and this is unaffected by the classification schemes we have examined (including trade-by-trade as well as groups based on one-tenth of a bucket). We focus on one-minute time bars as this data is less noisy, more widely available and easier to work with.

A.1.1. INPUTS

1. Time series of transactions of a particular instrument \((T_i, P_i, V_i)\):
   a. \(T_i\): Time of the trade.
   b. \(P_i\): Price at which securities were exchanged.
   c. \(V_i\): Volume exchanged.
2. \(V\): Volume size (determined by the user of the formula).
3. \(n\): Sample of volume buckets used in the estimation.

\(P_i, V_i, V, n\) are all integer values. \(T_i\) is any time translation, in integer or double format, sequentially increasing as chronological time passes.

A.1.2. PREPARE EQUAL VOLUME BUCKETS

1. Sort transactions by time ascending: \(T_{i+1} \geq T_i, \forall i\).
2. Expand the number of observations by repeating each observation \(P_i\) as many times as \(V_i\). This generates a total of \(I = \sum_i V_i\) observations \(P_i\).
3. Re-index \(P_i\) observations, \(i=1,...,I\).
4. Initiate counter: \(\tau = 0\).
5. Add one unit to \(\tau\).
6. If \(I < \tau V\), jump to step 10 (there are insufficient observations).
7. \(\forall i \in [\lceil \tau - 1 \rceil V + 1, \tau V]\), classify each transaction as buy or sell initiated.\(^{21}\)

a. A transaction $i$ is a **buy** if either:
   i. $P_i > P_{i-1}$, or
   ii. $P_i = P_{i-1}$ and the transaction $i-1$ was also a **buy**.

b. Otherwise, the transaction is a **sell**.

8. Assign to variable the number of observations classified as **buy** in step 7, and the variable the number of observations classified as **sell**. Note that $V = V^B_i + V^S_i$.


10. Set $L = \tau - 1$ (last bucket is always incomplete or empty, thus it will not be used).

### A.1.3. APPLY VPIN’s FORMULA

If $L \geq n$, there is enough information to compute $\text{VPIN}_L = \frac{\sum_{\tau=L-n+1}^{L} |V^S_{\tau} - V^B_{\tau}|}{\sum_{\tau=L-n+1}^{L} (V^S_{\tau} + V^B_{\tau})} = \frac{1}{nV} \sum_{\tau=L-n+1}^{L} |V^S_{\tau} - V^B_{\tau}|$.

Note that the same results can be reached with slightly different algorithms, like reusing data from previous iterations, performing fewer steps or in a different order, etc. Alternatively, the following formula could be used:

$$\text{VPIN}_L = \frac{E[|V^S_{\tau} - V^B_{\tau}|]}{V}$$

where:

- $E[V^S_{\tau} - V^B_{\tau}] = \sigma \sqrt{\frac{2}{\pi}} e^{\frac{(-E[V^S_{\tau} - V^B_{\tau}]^2)}{2\sigma^2}} + E[V^S_{\tau} - V^B_{\tau}] \left[1 - 2Z\left(-\frac{E[V^S_{\tau} - V^B_{\tau}]}{\sigma}\right)\right]$.
- $E[V^S_{\tau} - V^B_{\tau}] \approx \frac{1}{n} \sum_{\tau=L-n+1}^{L} (V^S_{\tau} - V^B_{\tau})$.
- $\sigma^2 = E[V^S_{\tau} - V^B_{\tau} - E[V^S_{\tau} - V^B_{\tau}]]^2 \approx \frac{1}{n} \sum_{\tau=L-n+1}^{L} \left( V^S_{\tau} - V^B_{\tau} - \frac{1}{L} \sum_{\tau=L-n+1}^{L} (V^S_{\tau} - V^B_{\tau}) \right)^2$.
- $Z(\cdot)$ is the cumulative standard normal distribution.

### A.2. MONTE CARLO VALIDATION

Here we present a Monte Carlo algorithm that simulates order flow based on known parameters ($\alpha, \mu, \delta$). Our VPIN metric can be estimated on that simulated order flow, and compared with the PIN values derived from the actual parameter values, $(\alpha, \mu, \epsilon)$.

1. Set Monte Carlo Parameters:
   a. Number of Simulations: $S$
   b. VPIN: $(V,n)$
   c. PIN: $(\alpha, \mu)$

---


22 $\delta$ can be arbitrarily set, as it does not affect the value of PIN nor VPIN. This can be easily seen by re-running this Monte Carlo with different values of $\delta$. 

35
2. Set \( \varepsilon = \frac{v-a}{2}, s=0, j=0 \)

3. \( j=j+1 \)

4. Draw three random numbers from a \( U(0,1) \) distribution: \( u_1, u_2, u_3 \)

5. If \( u_1 < \alpha \):
   a. If \( u_2 < \delta \):
      i. \( V_j^B = f^{-1}(u_3, \varepsilon) \)
      ii. \( V_j^S = f^{-1}(u_3, \mu + \varepsilon) \)
   b. If \( u_2 \geq \delta \):
      i. \( V_j^B = f^{-1}(u_3, \mu + \varepsilon) \)
      ii. \( V_j^S = f^{-1}(u_3, \varepsilon) \)

6. If \( u_1 \geq \alpha \):
   a. \( V_j^B = f^{-1}(u_3, \varepsilon) \)
   b. \( V_j^S = V_j^B \)

7. If \( j = n \):
   a. \( j=0 \)
   b. \( s=s+1 \)
   c. \( VPIN_s = \frac{\sum_{j=1}^{n} |V_j^S-V_j^B|}{\sum_{j=1}^{n} (V_j^S+V_j^B)} \)

8. If \( s < S \), loop to Step 3

9. Compute results:
   a. \( E[VPIN] = \frac{1}{S} \sum_{s=1}^{S} VPIN_s \)
   b. \( V[VPIN] = \frac{1}{S-1} \sum_{s=1}^{S} VPIN_s^2 - \frac{1}{S(S-1)} (\sum_{s=1}^{S} VPIN_s)^2 \)
BIBLIOGRAPHY


DISCLAIMER

The views expressed in this paper are those of the authors and not necessarily reflect those of Tudor Investment Corporation. No investment decision or particular course of action is recommended by this paper.