Assessing Measures of Order Flow Toxicity and Early Warning Signals for Market Turbulence*

Torben G. Andersen† Oleg Bondarenko‡

This version: June 2014
First version: March 2013

Abstract

Following the much publicized "flash crash" in the U.S. financial markets on May 6, 2010, much work has been done in terms of developing reliable warning signals for impending market stress. However, this has met with limited success, except for one measure. The VPIN, or Volume-synchronized Probability of INformed trading, metric is introduced by Easley, López de Prado and O'Hara (ELO) as a real-time indicator of order flow toxicity. They find the measure useful in predicting return volatility and conclude it, indeed, may help signal impending market turmoil. The VPIN metric involves decomposing volume into active buys and sells. We use the best-bid-offer (BBO) files from the CME Group to construct highly accurate trade classification measures for the E-mini S&P 500 futures contract. Against this benchmark, the ELO Bulk Volume Classification (BVC) scheme is inferior to a standard tick rule based on individual transactions. Moreover, when VPIN is constructed from an accurate classification, it behaves in a diametrically opposite way to BVC-VPIN. We also find the latter to have forecast power for volatility *solely* because it generates systematic classification errors that are correlated with trading volume and return volatility. Controlling for trading intensity and volatility, the BVC-VPIN measure has no incremental predictive power for future volatility. We conclude that VPIN is not suitable for capturing order flow toxicity or signaling ensuing market turbulence.

JEL Classification: G01, G12, G14, G17, and C58

Keywords: VPIN, Order Flow Toxicity, Order Imbalance, Accuracy of Trade Classification, Volatility Forecasting

*We are grateful to the CME Group for providing financial support and access to data from the CME DataMine system. Andersen also acknowledges support from CREATES, Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation. We thank seminar participants at Spot Trading, the Kellogg School, the Second Annual Meeting of DAEiNA, SAC Capital, Duke University, the High-Frequency Data and High-Frequency Trading Conference at the University of Chicago, The Modeling High-Frequency Trading Activity Conference at Banff, Canada, the Fields Institute at the University of Toronto, and the London Quantitative Finance Seminar at Imperial College, London, for helpful comments.

†Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208; NBER, and CREATES; e-mail: t-andersen@kellogg.northwestern.edu

‡Department of Finance (MC 168), University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607; e-mail: olegb@uic.edu

Electronic copy available at: http://ssrn.com/abstract=2292602

1 Introduction

The trading environment for securities listed on financial exchanges worldwide have undergone dramatic changes over the last decade. Assets are now traded almost exclusively on electronic platforms and high-frequency trading firms have largely taken over the basic market making function. These developments have brought impressive gains in market quality, as measured in terms of the average bid-ask spread or trading costs associated with small transactions. On the other hand, concerns have arisen regarding market fragility. One example is the occurrence of "flash crashes," where market prices move extremely rapidly – typically downward – for a short period of time, only to then reverse trend and almost as quickly return to a level near the original price point. While most of these incidents are self-contained and have a limited impact on other securities or markets, they can spill over into related venues and ultimately disrupt the broader financial market infrastructure. This occurred, e.g., during the flash crash on May 6, 2010, which widely is believed to have originated in the E-mini S&P 500 futures contract at the Chicago Mercantile Exchange (CME), cf. the CFTC-SEC Report (2010). Such incidents may call into question the role of financial markets in providing credible signals for capital allocation and raise issues about the fairness and efficacy of the trading process itself. In particular, the events of May 6, 2010, spurred a major debate about the origin of these idiosyncratic and highly erratic market dynamics and alternative measures which may prevent them from propagating into systemic disruptions to the financial system.

Against this backdrop, the VPIN metric, developed by Easley, Lopez de Prado and O'Hara, henceforth ELO, (2011a, 2011b, 2011c, 2012a), has generated widespread attention among academics, practitioners, regulators and exchanges. It is designed to provide a real-time estimate of the toxic order flow in a financial market, i.e., the extent to which market makers (high-frequency traders) are being adversely selected by agents with private information. When toxicity is high, liquidity is likely provided at a loss and may be withdrawn in short order. Thus, a high VPIN reading should signal an elevated probability that liquidity may vanish, causing a market disturbance. Clearly, such a "warning system" would be valuable for traders, exchanges, and regulators alike. In the words of Marcos López de Prado:

> The measure would have been able to anticipate two hours in advance there was a high probability of a liquidity-induced event on May 6.
>
> This would be much more effective than a circuit breaker [which] stops the infection after the infection is already widespread. (*Bloomberg*, October, 29, 2010).[1]

In addition, ELO (2012b, 2012c, 2012d, 2013) relate to VPIN in important ways. Meanwhile, different implementations are explored in a variety of studies, e.g., Abad and Yague (2012), Bethel et al. (2012), Chakrabarty, Pascual and Shkilko (2012), Menkveld and Yueshen (2013), Yildiz, Van Ness and Van Ness (2013), Wei, Gerace and Frino (2013), Wu et al. (2013), and Moos, Pöppe and Schiereck (2014).[2] On the regulatory side, the CFTC(2013) release – seeking comments on risk controls and system safeguards for automated trading environments – cites ELO (2012a) as motivation for exploring trading intensity and volume imbalance as market quality indicators.

Nonetheless, VPIN is not without controversy. One, the metric cannot be replicated with precision without matching the entire transaction record. One must start with the identical trade and any subsequent discrepancy in the recording of a trade will imply that *all* following VPIN measures differ, even if the transaction record is otherwise identical. Thus, at each point in time, the VPIN

[1] VPIN has been featured in many other prominent news outlets, including the Wall Street Journal, it has been the topic of many key-note presentations at leading academic conferences, it is featured on a variety of web videos displaying real-time events where VPIN is interpreted as signaling extreme toxicity. Moreover, three separate VPIN related patent applications have been filed.

[2] Likewise, commentators on the general debate surrounding high-frequency trading routinely reference the ELO findings; see, e.g., Corcoran (2013) and MacIntosh (2013).

Electronic copy available at: http://ssrn.com/abstract=2292602

metric hinges on the full preceding history of the tick data.[3] Two, even more critically, the appropriate metric for gauging performance is not evident and, at least for some important predictive comparisons, it fares poorly relative to traditional forecast variables, as detailed in Andersen and Bondarenko, henceforth AB, (2014). Three, VPIN is a fairly simple metric, constructed from the real-time transaction record. If it embodies important information regarding the future return distribution, it is surprising that high-frequency traders fail to respond more actively to such signals. After all, they are known to rely heavily on real-time analysis of transaction and order book data.

AB (2014) address some of these issues, but emphasize the original VPIN metric developed in ELO (2011a, 2011b, 2011c). ELO (2012a) advocate a different implementation, arguing that the Bulk Volume Classification (BVC) procedure, outlined in ELO (2012b), improves the accuracy of trade classification relative to the approach in ELO (2011a) which applies a tick rule to volume aggregated over fixed calendar intervals. ELO motivate the focus on "bulk volume" classification with the observation that they seek to measure the systematic impact of informed trading. This is, of course, a recurrent theme in market microstructure theory, and the early literature typically relies on imbalances in the active buy and sell volume to identify informed trading. However, this is arguably problematic in the current high-frequency trading regime. In particular, as discussed in ELO (2012b), large orders are now routinely split into smaller ones, so that it is the aggregate order flow rather than the individual orders that relate to trade motivation. In addition, large orders are often executed strategically by placing multiple limit orders in the book, and then canceling and replacing them depending on the evolving market dynamics. As such, buying pressure manifests itself not only through active buying, but also by persistent submission of limit buys and cancelations of limit sells. With informed traders relying partially on limit orders, the link between the "active" side of the trade and participants with underlying information is tenuous.[4]

The existence of alternative trade classification schemes raises the question of how to assess performance, as informed trading is not directly observable. ELO address the issue in a couple of different ways. First, they refer to the evidence in ELO (2012b), who gauge the accuracy of a given scheme relative to the underlying (active) trade imbalance, determined by the trade aggressor flag for each individual trade provided by the exchange, at different levels of order flow aggregation. This presumes that the active trade imbalance lines up with the intensity of informed trading for suitably aggregated order flow. Second, ELO (2012a) emphasize the predictive power of a given VPIN metric for forecasting future short term volatility and market turbulence, i.e., the usefulness in generating a leading indicator for escalating market tensions.

In this article, we provide a detailed analysis of the factors governing the properties of the VPIN metric. We pay particular attention to the criteria advocated by ELO (2012a). First, there is only scant evidence regarding the precision of alternative trade classification schemes vis-a-vis the underlying active buy-sell imbalances for different degrees of order flow aggregation. Moreover, there is no consensus across the few existing studies.[5] We confront the issue head on by exploiting best-bid-offer (BBO) data for the E-mini S&P 500 futures to construct order imbalance measures which are almost flawless in classifying the active buy and sell volume. This enables us to monitor the precision of alternative classification techniques over time, so we can relate the performance of a specific VPIN implementation to its accuracy in this regard. Second, we recognize it may be advantageous for the order imbalance measure to deviate from the underlying imbalance between active buys and sells if the latter is not a reliable indicator of the intensity of informed trading. It then becomes important to understand when and why certain classification schemes deviate system-

[3]We reproduce the algorithms for generating alternative VPIN measures, as stated by ELO (2011c), ELO (2012a), and ELO (2012b), in our Web Appendix.

[4]This type of reasoning motivates a broader definition of order flow which includes the activity on the limit order book, see, e.g., Cont, Kukanov and Stoikov (2014) for a recent analysis along these lines.

[5]For example, it is not clear that the ELO BVC approach dominates the traditional tick rule procedure, as Chakrabarty, Pascual and Shkilko (2012) reach the reverse conclusion based on data for individual stock trades.

atically from the underlying (active) trade imbalances. Hence, we explore how trade classification interacts with other features of the market environment to generate a specific behavior of the metric. In particular, we investigate the mechanism through which alternative VPIN measures end up displaying radically different degrees of correlation with trading volume and return volatility. In short, we get "under the hood" of the VPIN algorithm and explore why different implementations produce diverging measures and support diametrically opposite conclusions.

The results are striking. We find the ELO (2012a) BVC-VPIN metric to forecast return volatility *only* because it generates classification errors that are strongly correlated with innovations to trading volume and volatility. In contrast, the VPIN metric based on the actual order imbalance is *negatively* correlated with future volatility. A related result was reported by AB (2014) and it is also consistent with observations in ELO (2011a). The latter interpret this finding as indicating that classification based on individual transactions is particularly error prone. However, this is counter-factual. We document that order imbalance measures derived from tick data classify active buys and sells better than those obtained via time or volume bar aggregation. We further establish that the VPIN metric, once we control for volume and volatility, has no residual correlation with future volatility. In particular, the ELO (2012a) VPIN measure provides no incremental explanatory power beyond what is captured by traditional volatility predictors. Consequently, the metric is not based on superior (active) trade classification and does not help in forecasting market turmoil. As such, there is no evidence that VPIN is a useful measure of order imbalances or flow toxicity.

Our analysis goes well beyond existing studies exploring the properties of VPIN. First, our accurate classification of active buys and sells removes ambiguity surrounding one important explanation for the discrepancy between alternative VPIN metrics. Moreover, it helps us establish that the classification errors are highly correlated with the variables that VPIN portends to predict and, in particular, that BVC-VPIN, by construction, will be highly correlated with concurrent realized volatility. Second, it enables us to directly assess the accuracy of the "bulk volume" classification strategy introduced in ELO (2012b). Third, we can compare the traditional cumulative signed order imbalance measure constructed via the tick rule to the imbalances obtained from accurate classification. This is helpful in assessing whether the former provides economically meaningful incremental information during stressful market conditions such as the flash crash. Fourth, we can gauge the accuracy of alternative classification schemes by exploiting the diurnal patterns in volatility and volume. These features should induce systematic, and artificial, patterns in the corresponding order imbalance measures, if the latter, indeed, are mechanically correlated with market activity variables. Fifth, we exploit a longer sample for the E-mini S&P 500 futures than used in the ELO studies of VPIN, facilitating the identification of extreme events. However, the long sample period also forces us to invoke new normalization techniques to ensure that the VPIN metric is not distorted by non-stationarity in the volume series. Taken together, our findings speak generally to the potential of *any* VPIN metric to enhance the information content of traditional order imbalance measures, because we provide an observable and transparent benchmark for the analysis.

The remainder of the article is organized as follows. Section 2 describes how we obtain accurate trade classification by combining real-time trade and order book information. Section 3 introduces the VPIN metric. Section 4 develops notation to distinguish the many variants of VPIN we explore. Section 5 shows how the trend in trading volume distorts the VPIN metric, and motivates our data-dependent detrending procedure. Section 6 analyzes classification accuracy, both in terms of average (unconditional) performance and the correlation of errors with the activity variables, while Section 7 explores how these features induce distinct properties in the associated VPIN measures. Section 8 provides predictive regressions for return volatility. We find that the VPIN metrics have no *incremental* forecast power for volatility, as they are wholly subsumed by standard realized volatility and volume measures. Section 9 explores the factors behind the surprising finding that VPIN, based on the actual trade imbalances, is strongly *negatively* correlated with future volatility, while Section 10 concludes.

2 Data

We exploit best bid-offer (BBO) files for the E-mini S&P 500 futures contract from the CME Group. Among other variables, these "top-of-the-book" files provide a complete record for the best bid, bid depth, best ask, ask depth, trade prices, and trade sizes. Quotes and trades are time stamped to the second. Moreover, the files contain a sequence indicator that identifies the order in which quotes and trades arrive to the exchange. Thus, we know the actual sequence of order arrivals within each second. The files are obtained directly from CME DataMine.

Our sample covers the period from February 10, 2006, to March 22, 2011. Hence, our analysis is based on more than five years of tick-by-tick data. The E-mini S&P 500 futures contract trades exclusively on the CME GLOBEX electronic platform. The contract expires quarterly, on the March expiration cycle. The notional value of one contract is \$50 times the value of the S&P 500 stock index and has a tick size of 0.25 index points, or \$12.50. The contract trades essentially 24 hours a day, five days a week. Specifically, from Monday to Thursday, the trading is from 15:30 to 15:15 (Chicago Time) of the following day, with a half-hour maintenance shutdown from 16:30 to 17:00. On Sunday, the trading is from 17:00 to 15:15 of the following day. We exploit the front month futures contract until it reaches eight days to expiry, when market participants roll over their positions to the next maturity contract. This ensures that we use the most actively traded futures contract throughout our analysis.

2.1 Summary Statistics

The E-mini S&P 500 futures market is among the most liquid worldwide. Table 1 provides summary statistics. Within regular trading hours, there were around 4.8 trades per second and, though the numbers are much lower, they remain impressive outside the regular hours with a trade consummated about every three seconds.[6]

The average number of trades exceeded 136,000 per day and involved 1,782,000 contracts, implying a mean transaction size of about 13 contracts. Proper interpretation of this figure requires knowledge of how the trade size is recorded in the BBO files. A trade is consummated whenever an incoming order hits a resting limit order. For example, a market buy order for five contracts will be executed at the (best) ask price if the ask depth at the top of the book is at least five contracts. On the other hand, if the current ask depth is only four contracts, the last contract of the market buy order is filled by matching it with an ask quote at a higher price, and the total trade is recorded as two successive trades in the quantities of four and one. Notice that the first four contracts always are recorded as a single trade, even if the corresponding limit orders are executed against two separate entities, each offering, say, two contracts at the best ask. That is, the BBO files report the individual trades by price from the perspective of the (active) party who demands liquidity.[7] Finally, from the bottom panel of Table 1, we see that the order size distribution is severely right-skewed. The median transaction involves only two contracts, but the typical trading day contains several individual trades involving hundreds of contracts as well.

The activity in the limit order book is of separate interest. Table 1 reveals that the number of additions or cancelations of limit orders at the top of the book is sixfold the number of trades. Meanwhile, the best bid or ask price changes once for every eleven trades, on average. This implies an extremely active order book dynamic, but nonetheless allows for a sequence of trades to execute against a given set of (best) bid-ask quotes before they shift to a new tick.

[6] We remove two trading days from our sample. May 9, 2008, has missing quote data, while January 13, 2010, has a highly unusual trading pattern where, within one second, almost 200,000 contracts were traded – representing about 17% of the typical daily volume for the preceding month. Seemingly, two parties coordinated to execute a block trade through the electronic platform. The event did not occur during a stressful trading period. All qualitative conclusions are robust to the inclusion of the latter day.

[7] Alternatively, the transaction could be recorded as two trades of two contracts at identical prices.

Table 1: **Descriptive Trading Statistics for the E-mini S&P 500 Futures Contract**

	Regular	Overnight	Holiday
# Days	1285	1285	30
Volume (1 min)	3973	208	69
# Trades (1 min)	285	23	9
# BBO Depth Changes (1 min)	1730	175	60
# BBO Quote Changes (1 min)	26	8	4
Notional Value, $Mln (1 min)	235	12	4
Trade Size	13.9	8.9	7.3
BBO Depth Changes per Trade	6.1	7.5	6.4
Trades per BBO Quote Change	10.9	3.0	2.4

Order Size: Average Daily Percentiles

	Min	10%	50%	75%	90%	99%	99.9%	Max
All	1.0	1.0	2.1	7.5	30.5	233.4	643.5	1683.5

Notes: This table reports summary statistics for the trading in the E-mini S&P 500 futures contract over the period February 10, 2006 - March 22, 2011. The data are reported separately for Regular Trading Hours (Regular, 8:30-15:15), Overnight Trading Hours (Overnight, 15:30-8:30), and Holiday Trading Hours (Holiday, exchange holidays). "BBO Quote Changes" refer to changes in the best bid/ask price, while "BBO Depth Changes" refer to changes in the bid/ask price *and/or* bid/ask size. Both measures only include changes to the top level of the limit order book.

2.2 Trade Classification from the Quote and Transaction Record

The BBO files record the changes in the best bid or offer. Importantly, the quote updates arrive in pairs, one for the bid and one for the ask, synchronized by the sequence variable. The following table illustrates the information content of the BBO files.

Time	Sequence	BidPrice	BidSize	AskPrice	AskSize	TradePrice	TradeSize
17:02:58	5770	1289.50	125	1289.75	98		
17:02:58	5780	1289.50	125	1289.75	99		
17:02:59	5790					1289.75	5
17:02:59	5800	1289.50	125	1289.75	94		

At 17:02:58, the limit order book features a bid of 1289.50 and an ask of 1289.75 with depth of, respectively, 125 and 98 contracts. Within the same second, a limit sell order of a single contract arrives, raising the ask depth to 99. Note that, although the bid price and depth are unchanged, they are repeated if the other side of the book changes. At 17:02:59, five contracts are traded at 1289.75. As a result, ask depth drops by 5 contracts (sequence number 5800). Hence, we may unambiguously classify the trade of the five contracts as buyer initiated.

Effectively, the BBO files provide a snapshot of the top of the book whenever there is a change in either the quote or depth for the best bid or ask. By comparing the trade price with the preceding bid and ask, we can almost always identify the trade aggressor. The main exception occurs when trading starts up after a market closure. At these times, an initial (large) trade clears

all crossing orders in the book. These orders accumulate during a pre-opening window and execute simultaneously at the open. We assign half of this volume to the buy side and half to the sell side. These opening trades make up about 0.024% of the overall volume. In addition, during the regular course of trading, there are a small set of trades (about 0.048% of volume) consummated strictly between the last observed bid and ask quote.[8] Again, we split these equally between buys and sells. In total, we positively identify the direction for over 99.95% of the trades within the usual electronic trading environment. Our results are robust to any buy-sell assignment one may apply to the remaining trades.

The setting is exemplary for testing trade classification procedures. All trades and quotes are recorded in a single electronic system and trades can only be executed at prevailing bid or ask quotes. Finally, the sequence indicator provides a comprehensive recording of market events in real time. This eliminates problems stemming from the merger of separate files with quote and trade information, from dealing with reporting delays, and from integrating the activities across multiple venues. In practice, accurate sequencing of events is infeasible without this type of integrated system from a single electronic platform.

The only prior study exploring the accuracy of the trade classification associated with the VPIN metrics in ELO (2011a, 2012a) for the E-mini futures contract is ELO (2012b). They exploit the DataMine Market Depth files for a single year. This database contains raw message flow and can be used to construct a proxy for the order book. However, the data structure is complex and extensive data handling capabilities are required to separate the messages, disentangle the contracts for separate expiry times, sequence the transactions (stamped up to the millisecond), and separate fictitious trades (arising from the exchange algorithmic testing procedures) from actual ones. In contrast, the BBO data are cleaned and sequenced by the exchange itself, using in-house expertise and knowledge of the trading system. Hence, we analyze a significantly longer five year sample without encountering pitfalls associated with inadequate data cleaning or flaws in constructing the order book, and we can readily obtain the trade direction indicator for all transactions throughout the sample.[9]

2.3 Cumulative Order Flow Imbalance during the Flash Crash

In Figure 1, we depict the cumulative signed order imbalance and the S&P 500 futures price for the regular trading hours on May 6, 2010; the day of the "flash crash." The signed order imbalance is initiated at zero at the start of regular trading on the prior day.

The change in the cumulative signed order imbalance is clearly correlated with the price movement. Every distinct spike in the price path in Figure 1 may be linked to an innovation in the signed order imbalance. In particular, there is increasing selling pressure prior to and during the flash crash and sustained buying pressure during the recovery. For the full sample, the correlation between the S&P 500 log returns and the signed order imbalance is 0.53.[10] Evidently, the standard cumulative order imbalance measure is economically meaningful. At the same time, broader market conditions affect the sensitivity of transaction prices to shifts in the order imbalance. Some times, selling pressure causes only a minor price drop while at others prices react vigorously to negative innovations in order flow.

[8] These observations are likely generated when order book updates occur almost instantaneously and cannot be separated at the granularity of the mechanism recording the events in the BBO files.

[9] It should be noted that our database also is very large and necessitated the development of specialized storage and retrieval procedures along with the acquisition of a great deal of working memory.

[10] This correlation is measured across predetermined quantities of trading volume, corresponding to volume buckets of around $(1/50)^{\text{th}}$ of the trading day, as defined in the following section.

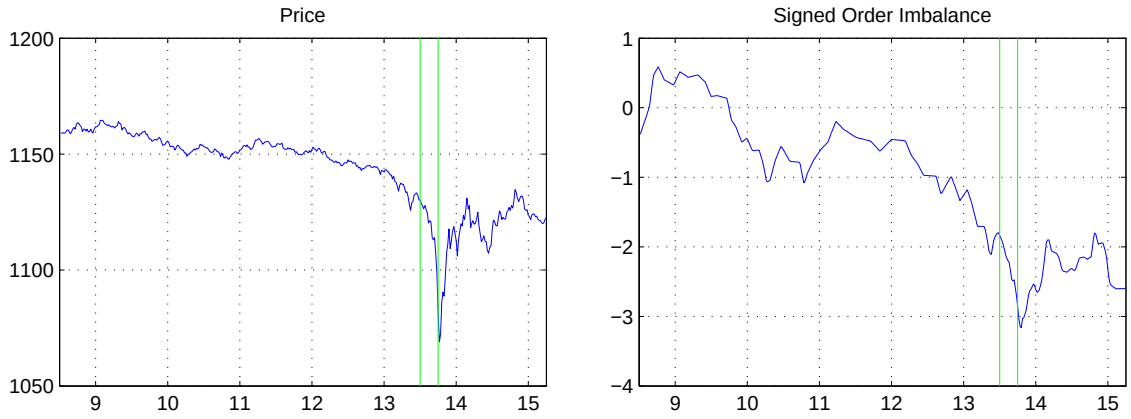

Figure 1: This figure plots the price and cumulative signed order imbalance for May 6, 2010. The solid vertical lines indicate the CTFC report timing of the "flash crash." The units for the vertical axis of the right panel are volume buckets, each representing about 2% of the expected daily trading volume.

3 The VPIN Metric

VPIN is simply defined as a rolling average of absolute order imbalance measures. As such, a key ingredient is the procedure for estimating the order imbalance. In addition, the moving average computation itself depends on the choice of measurement units and lag length. We review the procedure in detail to establish notation and identify the critical features.

3.1 Calibrating the Rolling Average of Order Imbalances.

This section reviews the construction of the VPIN metric for a fixed interval of calendar time, $[0, T]$, *given* a specific order imbalance measure. We represent the i^{th} traded contract by the time and price of the transaction, (t_i, p_i), $i = 1, \ldots, I$, where I is the total number of contracts traded. Of course, many transactions involve multiple contracts, but these may equivalently be viewed as simultaneous trades of a single contract. Hence, a trade involving $v > 1$ contracts will generate v replications of the pair (t_i, p_i) in the (unit) trade sequence. Transaction times are measured in seconds and form a non-decreasing sequence $0 \leq t_1 \leq t_2 \ldots t_I \leq T$. While many trades occur within the same second, we know the execution order from the event indicator. As such, we have a complete transactions history over the sample period.

We now define the volume bucket, V, as a fixed increment to the cumulative trading volume. We set V equal to $(1/50)^{\text{th}}$ of the volume on a regular trading day. Specifically, let T_ℓ indicate the time at which bucket number ℓ has just been filled, $\ell = 1, \ldots, \mathcal{L}$, where $\mathcal{L}$ denotes the total number of buckets in $[0, T]$. The VPIN measure is updated whenever we reach a new bucket, at times $T_1, \ldots, T_{\mathcal{L}}$, so it evolves in event time, governed by the (random) intensity of trading. Thus, by definition, each bucket encompasses V traded contracts.

VPIN is constructed from the order imbalances defined over the volume buckets. Denoting the imbalance measure for bucket ℓ by OI_ℓ, VPIN is obtained as the trailing moving average,

$$VPIN_\ell \;=\; \frac{1}{L}\sum_{j=0}^{L-1} OI_{\ell-j}\,, \tag{1}$$

where L indicates the length of the moving average. In accordance with ELO, we fix $L = 50$, so VPIN reflects the order imbalances across an average trading day. However, when the trading

volume is elevated (subdued), the buckets will fill quickly (slowly) so the value of VPIN at a specific point in time may reflect OI measures covering substantially less (more) than one day.

The order imbalances are derived from a given classification scheme, determining the number of contracts designated as (active) buys and sells, V_ℓ^B and V_ℓ^S, so $V = V_\ell^B + V_\ell^S$. It is convenient to define the proportional buy volume, $b_\ell = V_\ell^B/V$, where $0 \leq b_\ell \leq 1$, and the *signed order imbalance*, $\gamma_\ell = \left(V_\ell^B - V_\ell^S\right)/V = 2\,b_\ell - 1$, so $-1 \leq \gamma_\ell \leq 1$. The OI measure is then given as the *proportional absolute imbalance* over the bucket,

$$OI_\ell \;=\; \frac{\left|\, V_\ell^B \;-\; V_\ell^S \,\right|}{V} \;=\; \left|\, \gamma_\ell \,\right| \;=\; \left|\, 2\,b_\ell \;-\; 1 \,\right|, \tag{2}$$

Once we obtain a measure for V_ℓ^B, or equivalently b_ℓ, then γ_ℓ and OI_ℓ are also uniquely determined. Clearly, $0 \leq OI_\ell \leq 1$, with zero reflecting a perfectly balanced market and unity indicating maximal imbalance, i.e., the bucket is populated wholly by buys or wholly by sells.

3.2 Estimating the Order Imbalance.

We review alternative classification schemes which possess a few common features. First, every bucket ℓ is divided into Q_ℓ smaller units, or (time or volume) "bars." The actual classification is performed for the bars and then aggregated to the bucket level. Let the trading volume in bar (q, ℓ) be $V_{q,\ell}$, $q = 1, \ldots, Q_\ell$, so $V = V_{1,\ell} + \ldots + V_{Q_\ell,\ell}$, and the volume in each bar is further split into buys and sells, $V_{q,\ell} = V_{q,\ell}^B + V_{q,\ell}^S$. We also define the proportional buying volume, the signed order imbalance, and the volume weight for bar (q, ℓ) as,

$$b_{q,\ell} = \frac{V_{q,\ell}^B}{V_{q,\ell}}, \qquad \gamma_{q,\ell} = \frac{V_{q,\ell}^B - V_{q,\ell}^S}{V_{q,\ell}} \;=\; 2\,b_{q,\ell} - 1, \qquad \nu_{q,\ell} = \frac{V_{q,\ell}}{V}.$$

OI is then given as the (absolute) volume-weighted average of the signed OI measure for the bars across the bucket,

$$OI_\ell \;=\; \frac{\left|\, \sum_{q=1}^{Q_\ell} \left(V_{q,\ell}^B \;-\; V_{q,\ell}^S \right) \right|}{V} \;=\; \left| \sum_{q=1}^{Q_\ell} \gamma_{q,\ell} \cdot \nu_{q,\ell} \right|. \tag{3}$$

Second, trade classification is performed over bars of either fixed calendar time, known as *time bars*, or fixed trading volume, denoted *volume bars*. For the former, δ indicates the length of the time interval in seconds. For the latter, ν denotes the size of the bar relative to the volume bucket, so a volume bar contains $\nu \cdot V$ traded contracts, $0 < \nu \leq 1$. The finest level of granularity is achieved via individual transactions, $\nu = 1/V$, so that each bar consists of a single traded contract. Here, the key is to ascertain whether the buyer or seller triggered the trade, i.e., whether the trade took place at the prevailing ask or bid. This approach seems natural, but ELO (2012a) argue that classification of individual transactions in a high-frequency environment is fraught with difficulties and, inevitably, induces significant errors. Instead, they favor using larger bars that cumulate trades into "bulks" of aggregate order flow.

Third, the bar size determines the degree of aggregation applied to trades prior to classification. For volume bars, we pick ν so the bar size, $\nu \cdot V$, and number of bars, $Q_\ell = Q = 1/\nu$, are integers. Choosing $\nu = 1/V$ (and $Q = V$), we obtain a bar size of unity – or *contract-by-contract* classification. At the other extreme, for $\nu = 1$, we have only one bar with V traded contracts, representing *bucket-level* classification. An intermediate choice, $1 < Q < V$, implies multiple bars per bucket, each representing aggregated order flow. However, for volume bars, OI_ℓ always equals the simple average of the signed order imbalances, $|\gamma_{1,\ell} + \ldots + \gamma_{Q,\ell}|\,/\,Q$. Below, we consider $\nu = 0.02$ and 0.10, producing fifty and ten bars per bucket. In contrast, for time bars, Q_ℓ varies inversely

with trading intensity. We explore bar sizes from one second to several minutes, including $\delta = 60$ seconds, which is the leading case for ELO (2011a, 2012a). In the latter scenario, there are periods in which a bucket is filled in less than one minute, so $Q_\ell = 1$, while at other times it takes hours to fill the bucket, so Q_ℓ is hundred-fold larger.

Fourth, alternative trade classification schemes may be applied to the bars. The initial VPIN metric, developed in ELO (2011a, 2011b, 2011c), exploits a *tick rule*. To formalize this procedure, let $\Delta P_{q,\ell} = P_{q,\ell} - P_{q-1,\ell}$ be the price change over the bar.[11]

The Tick Rule: Let bar (q, ℓ) be designated a *buy*, i.e., $b_{q,\ell} = 1$, if:

$\Delta P_{q,\ell} > 0$ (the price change is positive), or
$\Delta P_{q,\ell} = 0$ and $b_{q-1,\ell} = 1$ (no price change, but the previous bar is a buy).

Otherwise, bar (q, ℓ) is a *sell*, i.e., $b_{q,\ell} = 0$.

ELO (2012a) instead invoke a Bulk Volume Classification (BVC) strategy applied to aggregated order flow. It assigns the proportional buy volume as a function of the price change.

The BVC Rule: Let $Z(\cdot)$ denote the cumulative distribution function of a standard normal variate. The rule assigns the proportional buy volume in bar (q, ℓ) as:

$$b_{q,\ell} \;=\; Z\left(\frac{\Delta P_{q,\ell}}{\sigma_{\Delta P}}\right) \qquad \text{and} \qquad \gamma_{q,\ell} \;=\; 2 \cdot Z\left(\frac{\Delta P_{q,\ell}}{\sigma_{\Delta P}}\right) \;-\; 1\,, \tag{4}$$

where $\sigma_{\Delta P}$ is the sample standard deviation of the price change between adjacent bars.

Thus, the BVC approach interprets no price change as balanced trading, while a large positive (negative) price change is translated into a proportionally large (small) buy volume.

A qualitatively different approach, hitherto not explored for VPIN, is to classify buys and sells using the prevailing bid and ask quotes. This procedure is reliable only if timely order book information is available. This condition is satisfied for the E-mini S&P 500 futures contract, and we exemplified the approach in Section 2.2.

Order Book Matching: Let P_q be the transaction price for contract q. It is designated a *buy* ($b_q = 1$) or a *sell* ($b_q = 0$) according the rule:

$b_q = 1$ if $P_q = P_q^a$ (the trade is consummated at the ask).
$b_q = 0$ if $P_q = P_q^b$ (the trade is consummated at the bid).
$b_q = \frac{1}{2}$ if $P_q < P_q^a$ and $P_q > P_q^b$ (no unique assignment).

where P_q^b and P_q^a indicate the last bid and ask quotes observed prior to transaction q.

As discussed in Section 2.2, order book matching allows us to, unambiguously, classify over 99.95% of the trades during the regular trading regime for the E-mini S&P 500 futures. Since we obtain the actual consummated buy and sell transactions, it provides a benchmark for alternative classification schemes, and we refer to the discrepancy between buys and sells as the "actual" (signed) order imbalance. Nonetheless, we are mindful that trade execution arises from an interaction of incoming and resting orders that is resolved through the real-time sequencing of the matching engine. The speed of the market and the differential degree of latency across market participants renders it infeasible to condition on the current state of the market. In spite of this caveat, the actual order flow provides a natural benchmark by capturing the flow of orders that end up executing against resting limit orders.

[11] Consistent with prior notation, $P_{q,\ell}$ is the last transaction price of bar (q, ℓ). For $q = 1$, i.e., the first bar in bucket ℓ, $P_{q-1,\ell} = P_{Q_{\ell-1,\ell-1}}$ is the transaction price of the last traded contract in the preceding volume bucket.

The objective, or verifiable, nature of the actual trade classification does not imply that the corresponding VPIN measures, based on this classification, are superior. It is possible that alternative schemes are preferable for capturing the private information embodied in the order flow. Later on, we test predictions in this regard put forth by ELO (2011a, 2012a).

4 Notation

We need to distinguish between several variations of VPIN and adopt the following naming convention. The first three letters of the index denote a trade classification rule:

ACT – "ACTual" trade classification via order book matching (price happens at bid or ask);

TIC – TICk rule classification, as used in ELO (2011a, 2011b, 2011c);

BKU – BulK volume classification, as used in ELO (2012a), where the proportion of buys is determined by the size of the price change relative to the average volatility over the sample (an Unconditional estimate of $\sigma_{\Delta P}$);

BKC – same as BKU, but now volatility is time-varying and estimated using a rolling window of about one week (a Conditional estimate of $\sigma_{\Delta P}$);

RND – uninformative or "RaNDom" trade classification, based on the L_2 norm, as developed in AB (2014) (labeled U2-VPIN in their exposition).

The superscript letter refers to the type of data aggregation:

R – individual trades based on tRansaction level data;

T – "Time bars" based on fixed increments to calendar time;

V – "Volume bars" based on fixed increments to trading volume.

Note that not all trade classification rules apply to all types of aggregation. In particular, Rule ACT can be used on transaction level data only. Also, even though Rule RND could be applied to volume bars, the resulting index is constant and thus uninteresting.

Finally, we use a subscript to indicate the degree of aggregation.[12] For time bar aggregation, we consider four separate frequencies, $\delta = 1$, 10, 60, and 300, indexed by subscript 1 through 4. Similarly, for volume bar aggregation, we consider $\nu = 0.02$ and 0.10, i.e., 50 and 10 volume bars per bucket, and index those by subscript 1 and 2. For example, we calculate the time bar VPIN measures TIC_1^T, TIC_2^T, TIC_3^T, and TIC_4^T, as well as the volume bar VPIN metrics BKU_1^V and BKU_2^V.[13] Table 2 summarizes our notation for the various VPIN metrics.

5 Regularizing the VPIN Metric

Two features of the volume bucketing procedure render the regular VPIN implementation unsuitable for our lengthy sample. First, it is awkward that every VPIN observation depends on the complete tick record preceding it. Second, the volume trend distorts the time-bar based metrics used by, e.g., ELO (2011a, 2012a). Hence, we modify the VPIN computation to alleviate these features, while retaining the essential character of the metric.

[12] Of course, this is not applicable for transactions data, i.e., the ACT^R and TIC^R schemes, so they have no subscripts.

[13] In this terminology, the empirical work in ELO (2011a,b,c) focuses primarily on TIC_3^T, while ELO (2012a) relies on BKU_3^T. Moreover, the study of AB (2014), focusing on a shorter sample than here, analyzes the following VPIN series: TIC^R, TIC_k^T, BKU_k^T, RND_k^T, TIC_m^V, for $k = 2, 3, 4$ and $m = 1, 2$.

Table 2: **Notation for VPIN Measures**

	Trade Classification				
Data Aggregation	Actual	Tick-rule	Bulk unconditional $\sigma_{\Delta P}$	Bulk conditional $\sigma_{\Delta P}$	Random
Transactions, R	ACT^R	TIC^R	–	–	–
Time Bars, T	–	TIC_1^T–TIC_4^T	BKU_1^T–BKU_4^T	BKC_1^T–BKC_4^T	RND_1^T–RND_4^T
Volume Bars, V	–	TIC_1^V–TIC_2^V	BKU_1^V–BKU_2^V	BKC_1^V–BKC_2^V	–

5.1 Enhancing the Ability to Replicate the VPIN Metric

AB (2014) establish that the VPIN metric is sensitive to initial conditions. The starting point determines the location of the buckets throughout the sample. If it changes, the bars are split differently across the buckets. This feature hampers replication, verification, and comparison across studies. Any deviations in the reported transactions, discrepancies in the filtering of erroneous entries, or differences in the handling of transactions consummated at the daily resumption of electronic trading *at any point throughout the sample*, imply that the VPIN metrics will deviate to varying degrees.

To mitigate such effects and facilitate replication, we instead restart the bucketing process at the beginning of each trading day (normally, at 15:30 of the previous calendar day). At the end of a given day, the last bucket will usually contain less than V contracts and we simply discard it.[14] As a result, one may verify the VPIN computations for a given trading day by having access to the identical transactions record for that specific day (and the fifty preceding volume buckets necessary for the initial VPIN computation) as well as knowledge of the size of the volume bucket only.

5.2 Inducing Stationarity in the Volume Series

ELO (2011a, 2012a) fix the volume bucket, V, at $(1/50)^{\text{th}}$ of the average daily trading volume in the sample. In our case, this would imply setting V to about 36,500 contracts.[15] We refer to this approach as "constant V". This may be sensible for short samples bereft of persistent variation in daily volume, but it is problematic if there is a pronounced trend in volume. Since its introduction in 1997, the volume in the E-mini contract has risen steadily. The average daily volume was about 17 thousand contracts in 1998, about 970,000 in 2006, and more than 2 million in 2010. We also note that the theory motivating the VPIN construction in ELO (2012a) is cast within a stationary market environment.

Figure 2 shows that the volume displays high-frequency spikes that are strongly correlated with the VIX index during turbulent market conditions. When trading intensifies, there are less time bars per bucket and, as documented in AB (2014), this tends to inflate the time-bar based VPIN metric. In our sample, the bucket of 36,500 contracts is filled, on average, within 22 minutes of regular trading in the lowest volume month, but in just 4.5 minutes during the highest volume month. ELO (2012a) scale the bucket size to the trading intensity to ensure compatibility of the corresponding VPIN series, when comparing different futures markets. The argument for a similar adjustment for a pronounced volume trend *within* each series is even more compelling. Only if the metric exploits a similar number of bars across the sample can we meaningfully interpret VPIN

[14] Alternatively, we could add the volume of the last incomplete bucket to the previous one and weigh the contribution of this bucket accordingly in the VPIN computation. We have confirmed that the discrepancy between these two approaches is negligible.

[15] This compares to $V = 39{,}351$ contracts in ELO or $V = 40{,}000$ in AB (2014) for more recent and shorter samples. The average volume in the early part of our sample is considerably lower than in later years.

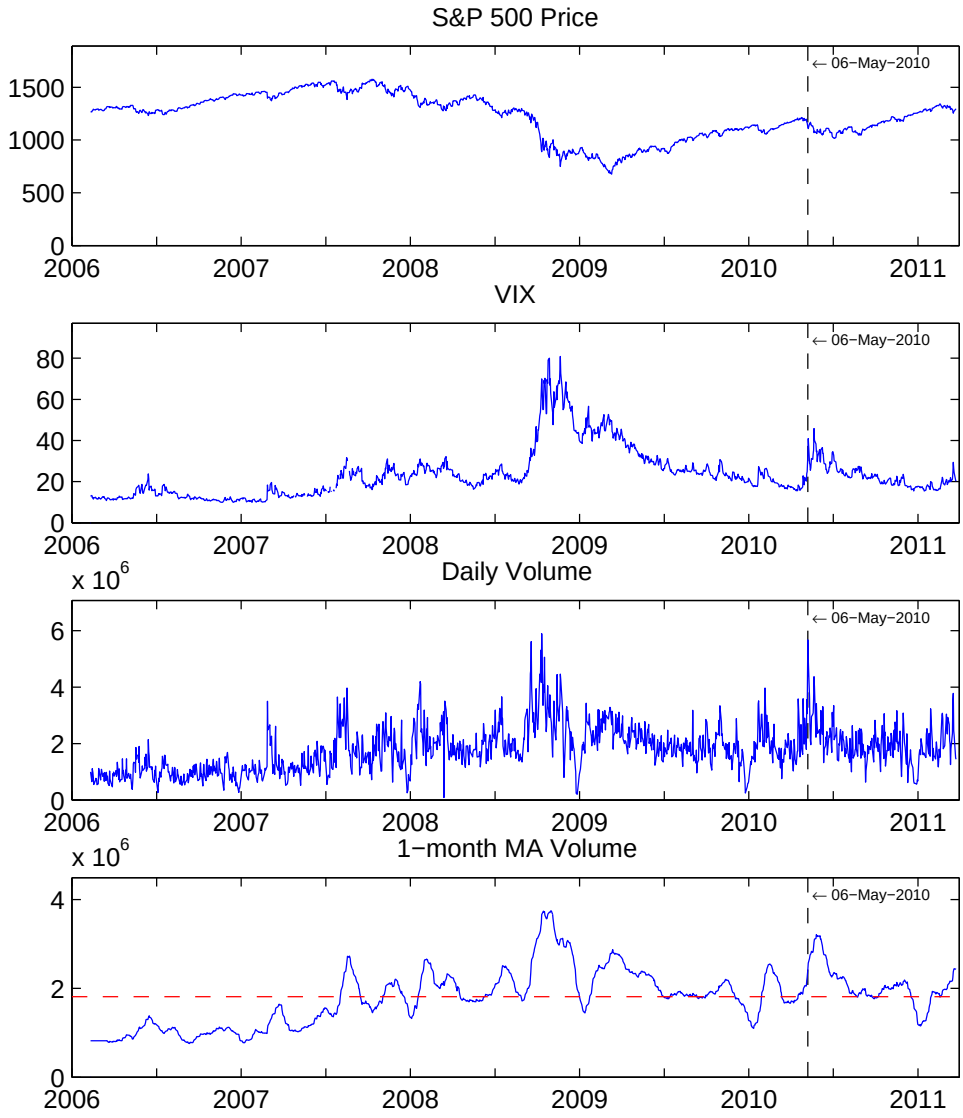

Figure 2: This figure depicts daily values of the S&P 500 index, VIX, trading volume, and a 21-day moving average of the daily volume over the entire sample, February 10, 2006 - March 22, 2011. The dashed line in the bottom panel indicates the average daily volume over the full sample.

values at different times. In particular, if a secular increase in trading doubles the average daily volume, the bucket size should also be doubled to render the early and late stages of the sample comparable.

Hence, we set V equal to $(1/50)^{\text{th}}$ of the daily volume *over the preceding month.*[16] The choice of a one-month moving average helps smooth out seasonal volume fluctuations associated with, e.g., declines around major holidays, which are evident from the third panel of Figure 2. We refer to this as the "detrended V" approach below. The bottom panel of Figure 2 contrasts the one-month moving average of the daily volume used in this study with the dashed line, representing the daily sample mean, used in the constant V approach.

To illustrate the effect of our volume regularization, we contrast the daily maxima of the BKU_3^T-VPIN series obtained from the constant and detrended V approach in panel one and three of Figure

[16]On any given day, we round V to the closest multiple of 100. This ensures that when the volume bar aggregation is used, the bars we consider always contain a whole number of contracts.

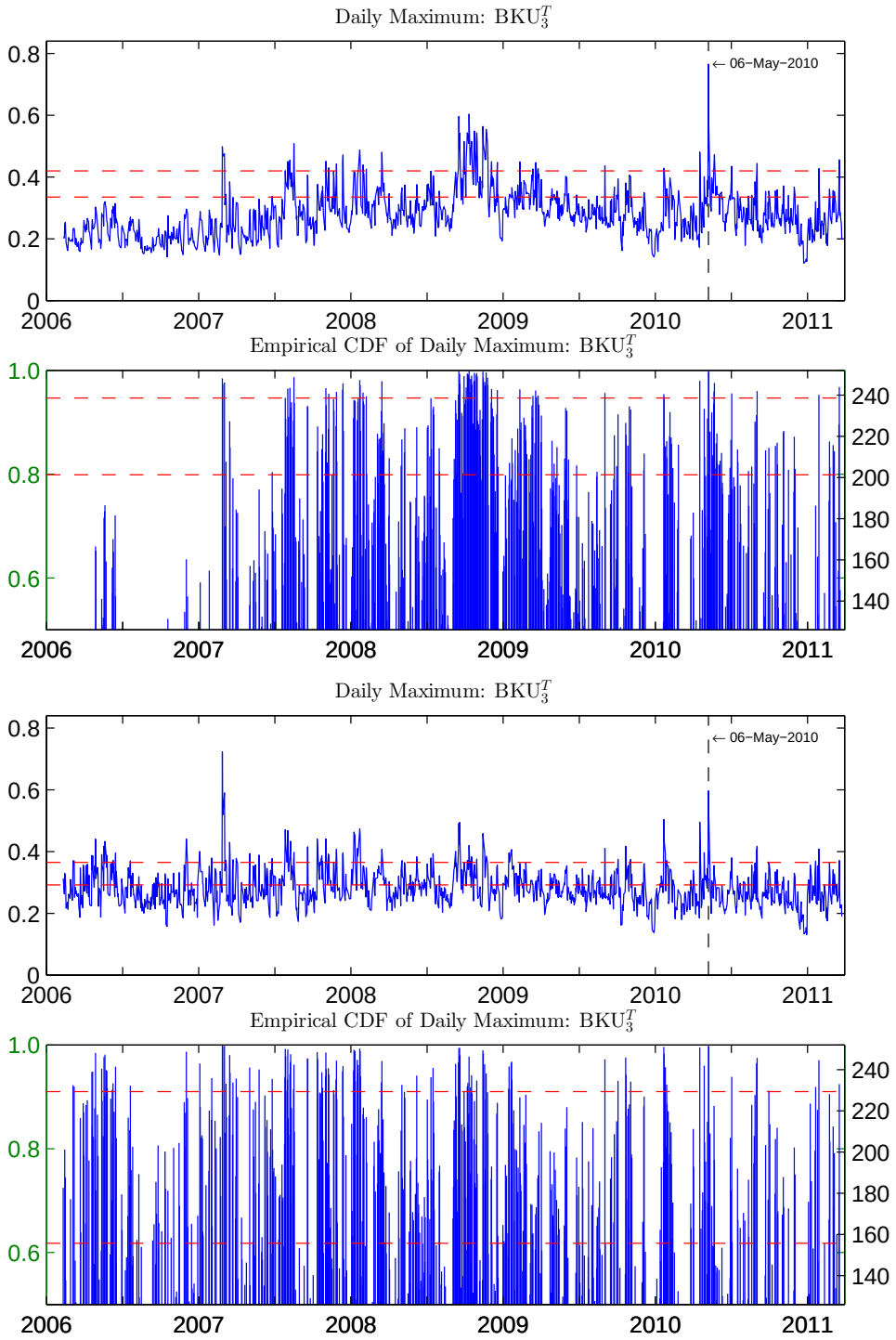

Figure 3: The two top panels concern the VPIN measure under the constant V approach. Panel one depicts the daily maximum value of BKU$_3^T$-VPIN. Panel two depicts the CDF (left scale) and corresponding number of days per year with a (maximum) VPIN level below that for the given day (right scale). The horizontal dashed lines indicate the level of VPIN on the Flash Crash day at 12:30 (the lower one) and 13:30 (the upper one). The two bottom panels provide the identical series for BKU$_3^T$-VPIN, but under the detrended V approach.

3.[17] For the first year, the level and variation in the BKU$_3^T$ series in the top panel is noticeably

[17]We use BKU$_3^T$ for the illustration as it is most closely aligned with the VPIN metric adopted in ELO (2012a),

lower than for the remainder of the sample. In addition, BKU_3^T only attains extreme values when volume and volatility are elevated, (refer back to Figure 2). While toxicity may tend to be high during turbulent market conditions, the pronounced pattern is troublesome. The theory underlying VPIN is developed for a stationary setting, but it seems that even exceptional days in the early part of the sample fail to stand out due to the apparent dampening of the metric when (trend) volume is low. In contrast, the detrended V series for BKU_3^T in panel three appears stationary. The financial crisis no longer stands out as one prolonged period of elevated toxicity. This is consistent with the market fluctuations at that time being driven primarily by observable shocks to economic policy and fundamentals rather than by private information filtering into the market through trading. The extreme VPIN value during the flash crash is evident from both BKU_3^T series. However, for the detrended series, February 27, 2007, represents an even larger outlier. On this date, the S&P 500 index tumbled 50 points in its largest drop in about four years. Moreover, two days later, on March 1, 2007, the VPIN value almost matches that of May 6, 2010, as well. In short, the VPIN series is drastically transformed by the standardization of the volume series.

ELO (2012a) argue that extreme VPIN values should be gauged from the CDF. The empirical CDF for the daily maximal BKU_3^T-VPIN readings are reported on the left vertical axis of panel two and four. On the right, we provide the number of days per year that realizes a maximum below the current value. For example, the second panel shows that the BKU_3^T-VPIN metric one hour prior to the flash crash (lower dashed line) was exceeded on about 20% of the days, or about 50 times a year. Likewise, the BKU_3^T-VPIN value at the onset of the crash (upper dashed line) was topped 13 times a year, or about once a month. Moreover, these events were all observed post 2006, with massive clustering during the financial crisis. In contrast, the extreme VPIN events are almost uniformly distributed once we control for the volume trend. This is consistent with a stationary environment where toxicity rears its head randomly. Thus, the following results are based on the detrended volume series.[18]

6 Misclassification Measures for Order Flow Imbalance

We now explore the properties of alternative trade classification procedures, using the accurate active buy-sell classification based on the ACT^R scheme from Section 2.2 as benchmark.

6.1 Order Flow Aggregation and Trade Classification

The determination of whether a given trade in an electronic limit order book system is an active sell or buy is clearly contingent upon the circumstances surrounding the individual transaction: was the trade consummated at the existing bid or ask quote. Since the ACT^R scheme establishes the latter with extremely high accuracy, it is in principle straightforward to evaluate the precision of alternative trade classification procedures.

Nonetheless, there are complications. Most importantly, for the construction of the VPIN metric, we rely on the relative amount of active buying over volume buckets, i.e., for quite highly aggregated order flow. It is possible that some schemes are superior for transaction-by-transaction classification while others dominate for estimating the proportion of active buying for the cumulative volume across a large set of separate transactions.[19]

Consider the following illustrative example. The bucket covers 2 minutes and there are 4 trades of one contract within each minute. The actual trade sequence is BBBSSSSB, which we represent via a buy indicator sequence and compare to a candidate classification rule C:

but similar results are obtained for alternative VPIN measures based on time bars.

[18] Qualitatively similar, but less transparent, findings obtain if we rely on the raw volume figures.

[19] In fact, ELO (2012a, 2012b) argue forcefully that the "bulk" volume approach is superior to tick-by-tick identification due to the noise associated with classification of individual trades in a high-frequency environment.

Rule	Trade Sequence	Misclassification Rate Individual	1 minute	2 minute
Actual (A)	$(1,1,1,0,0,0,0,1)$	–	–	–
Candidate (C)	$(1,0,1,0,1,0,1,0)$	0.50	0.25	0.00

In terms of individual contracts, or volume bars for $\nu = 1/8$, rule C misclassifies four out of eight, or 50%. However, if we focus on one-minute bars (or $\nu = 1/2$), then rule C misclassifies only two contracts ($|2-3| + |2-1|$), so the inaccuracy rate drops to 25%. Finally, at the bucket (two-minute bar) level, the classification is perfect ($|4-4| = 0$).

The example highlights the point that *order flow aggregation improves accuracy.* This happens because of error diversification. Formally, for V_q indicating the trade volume in bar q within a given bucket, $q = 1, \ldots, Q$, it follows directly from the triangle inequality that,

$$\left| \sum_{q=1}^{Q} \left(\hat{V}_q^B - V_q^B \right) \right| \;\; \leq \;\; \sum_{q=1}^{Q} \left| \hat{V}_q^B - V_q^B \right| . \tag{5}$$

Hence, the overall classification accuracy for aggregated order flow, obtained by cumulating over individual transactions or bars, is inevitably better than if precision is measured for the smaller units. Consequently, active trade classification schemes must be compared at the same aggregation level to avoid unduly favoring the procedure using the more aggregate order flow.

Apart from TIC^R (and ACT^R), all our classification rules exploit bar-level aggregation. This involves a loss of granularity but, as noted above, it does not necessarily imply that bar-level classification is inferior. Ultimately, the relative usefulness of alternative schemes is an empirical question which we begin to address in Section 6.3.

6.2 Defining Misclassification Measures

This section formalizes the "misclassification" measure adopted in Section 6.1. We develop the measure for volume buckets and then extend it to the individual contract level as well.

6.2.1 The Misclassification Measure for Volume Buckets

We initially focus on a given bucket, ℓ, comprising V traded contracts, allocated to Q separate bars. The volume in bar q is $V_{q,\ell}$ and the (true) proportion of buys is $b_{q,\ell}$. The corresponding estimates of active buys and proportion of buys are, respectively, $\hat{V}_{q,\ell}$ and $\hat{b}_{q,\ell}$.

Upon defining $\nu_{q,\ell} = V_{q,\ell}/V$ and aggregating by suitable volume-weighting, we arrive at a bucket-wide misclassification measure, MB_ℓ, for a given classification scheme,

$$MB_\ell \;\; = \;\; \frac{\sum_{q=1}^{Q} |\hat{V}_{q,\ell}^B - V_{q,\ell}^B|}{V} \;\; = \;\; \sum_{q=1}^{Q} |\hat{b}_{q,\ell} - b_{q,\ell}| \cdot \nu_{q,\ell} \, .$$

The measure simplifies further for volume bars, where $\nu_{q,\ell} = 1/Q$,

$$MB_\ell \;\; = \;\; \frac{1}{Q} \sum_{q=1}^{Q} |\hat{b}_{q,\ell} - b_{q,\ell}| \, .$$

Finally, we note that MB_ℓ equals $(1 - Ar_\ell)$, where Ar_ℓ is the accuracy measure of ELO (2012b) for bucket ℓ. Thus, the two concepts are formally equivalent.

The MB_ℓ measure references a single volume bucket. Our sample contains thousands of buckets and we construct the overall measure by averaging across all buckets,

$$MB \;=\; \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} MB_\ell \,.$$

6.2.2 Misclassification at the Individual Contract Level

Trade-by-trade classification is critical for a host of microstructure applications, but the bar-based rules are not designed for this purpose. Nonetheless, if one decides to assign a trade classification to individual contracts on the basis of the bar-based measures, the natural procedure is self-evident. We simply assign a probability of an active buy to each contract that equals the overall estimated proportion of buys for the bar.

We exemplify the approach for the tick-rule and BVC strategies applied to one minute bars. Consider the actual trade sequence (A) from above, spanning two one-minute bars with four trades within each bar, $(1, 1, 1, 0, 0, 0, 0, 1)$. If the price rises over the first and drops over the second bar, the (one minute) tick rule classifies all contracts in the first bar as buys and those in the second as sells. If the price increase over the first bar exceeds the drop over the second, BVC may assign a buy ratio of 0.90 to the first bar and 0.20 to the second.

		Misclassification Rate		
Rule	Trade Sequence	Individual	1 minute	2 minute
Actual	$(1, 1, 1, 0, 0, 0, 0, 1)$	–	–	–
Tick	$(1, 1, 1, 1, 0, 0, 0, 0)$	0.250	0.250	0.000
BVC	$(0.9, 0.9, 0.9, 0.9, 0.2, 0.2, 0.2, 0.2)$	0.325	0.100	0.050

The (one-minute bar) tick rule produces a misclassification rate of 25% (2/ 8) for individual trades. Note, however, that the overall inferred relative buy volume perfectly matches the actual proportion of buys, so for two-minute bars the misclassification rate is 0%. In the case of BVC and one-minute bars, the misclassification rate for individual contracts is 32.5%.[20] In contrast, the BVC rate is only 10% when assessed on the basis of the aggregated buy volume within the one-minute bars.[21] Both scenarios illustrate the effect of interpolating proportional order flow measures from bars to individual contracts – it induces reverse error diversification. Most importantly, we must avoid comparing, say, the tick rule misclassification rate for individual contracts (25%) to the aggregate one-minute bar BVC rate (10%), and instead compare it to the BVC rate for contract-to-contract classification (32.5%).[22]

Formally, for I total trades, the misclassification measure at the contract level is,

$$MC \;=\; \frac{1}{I} \sum_{i=1}^{I} |\hat{b}_i - b_i|, \tag{6}$$

where $\hat{b}_i$ and b_i denote the estimated and true buy indicator for contract i.

[20] The rate is computed as $(\,3\,|1-0.9| + |0-0.9| + 3\,|0-0.2| + |1-0.2|\,)\;/\,8$.
[21] This rate is computed as $(\,|0.90-0.75| + |0.20-0.25|\,)\;/\,2$.
[22] Alternatively, perform the comparison explicitly at the one-minute or two-minute bar level for all schemes.

6.3 Empirical Inaccuracy Measures for Alternative Classification Schemes

This section compiles inaccuracy measures for the various classification techniques across our full sample for the E-mini S&P 500 futures. We contrast our results to the only existing evidence concerning this contract, namely ELO (2012b, 2012c), but the findings also complement a large literature on empirical trade classification, originating with Lee and Ready (1991).[23]

6.3.1 Misclassification at the Transaction Level

The top panel of Table 3 shows that the tick rule applied at the contract level misclassifies 11.6% of the transactions over our five year sample (entry TIC^R). This is roughly consistent with ELO (2012b), where a failure rate of 13.6% (1-0.864) is reported. Our interpretation is entirely different, however. We stress the striking improvement relative to the tick rule applied to aggregated order flow. Already at the one-second level, the error rate reaches 24.8% (entry TIC_1^T), and the trend continues with failure rates of 38.7% and 43.1% at the 60- and 300-second levels (entries TIC_3^T and TIC_4^T). Finally, the best time-bar bulk volume procedure is BKU_3^T with an error rate of 23.2% at the transaction level. In terms of classifying individual trades, the approaches in ELO (2011a) and ELO (2012a) are an order of magnitude worse than the traditional tick rule applied at the transaction level.

Table 3: **Error Rates for Alternative Trade Classification Rules**

Panel A: MC

Rule	$\square^R$	$\square_1^T$	$\square_2^T$	$\square_3^T$	$\square_4^T$	$\square_1^V$	$\square_2^V$
ACT	0.000						
TIC	0.116	0.248	0.330	0.387	0.431	0.339	0.388
BKU		0.283	0.242	0.232	0.245	0.233	0.170
BKC		0.282	0.242	0.240	0.259	0.229	0.170

Panel B: MB

Rule	$\square^R$	$\square_1^T$	$\square_2^T$	$\square_3^T$	$\square_4^T$	$\square_1^V$	$\square_2^V$
ACT	0.000						
TIC	0.023	0.038	0.072	0.141	0.264	0.065	0.121
BKU		0.040	0.045	0.083	0.161	0.043	0.047
BKC		0.040	0.045	0.086	0.171	0.041	0.043

Notes: This table reports misclassification rates at the contract level (MC) and for volume buckets (MB). Rows represent alternative trade classifications: ACT (actual), TIC (tick rule), BKU (BVC with unconditional $\sigma_{\Delta P}$), and BKC (BVC with conditional $\sigma_{\Delta P}$). Columns represent the type of data aggregation: R (transactions), T (time bars, with $\delta = 1$, 10, 60, and 300), and V (volume bars with $\nu = 0.02$ and 0.10). For example, the procedure in ELO (2012a), BKU_3^T (BVC with unconditional $\sigma_{\Delta P}$ for one-minute bars) appears in the third row, fourth column (BKU and $\square_3^T$).

[23]The two ELO citations refer to sequential versions of the same working paper. We reference both, because the first deals with the BVC approach as implemented in ELO (2012a), while the second explores additional features that are critical for our discussion, but relies on a different BVC procedure. The voluminous literature on trade classification includes, e.g., Aitken and Prino (1996), Asquith, Oman and Safaya (2010), Boehmer, Grammig and Thiessen (2007), Chakrabarty, Li, Nguyen and Van Ness (2007), Chakrabarty, Moulton and Shkilko (2012), Ellis, Michaely and O'Hara (2000), Finucane (2000) and Odders-White (2000).

6.3.2 Misclassification at the Bucket Level

The bottom panel of Table 3 reveals that the bucket level error rates are sharply lower than at the contract level, indicating the strength of the error diversification effect. For instance, the tick rule now misclassifies only 2.3% versus the 11.6% at the individual contract level. Nonetheless, the relative rankings remain largely intact. As before, the tick rule based on individual trades outperforms both the tick rule or BVC approach applied to aggregated order flows, and the error rates increase monotonically with aggregation. Hence, it is always better to use tick data relative to one-second data as well as ten-second data rather than one-minute data. The adverse impact of applying the tick rule to aggregated order flow is not surprising: within a bar the transactions are classified as either all buys or all sells. Since active buys and sells alternate quite rapidly, this is increasingly counter-factual as the bar size grows. BVC provides a more balanced assignment of buys and sells which, almost tautologically, is better for larger bars. This explain why BVC underperforms the tick rule for small bars, e.g., TIC^R and TIC_1^T versus BKU_1^T, while the result reverses for TIC_2^T and TIC_3^T versus BKU_2^T and BKU_3^T.

Notably, our findings are at odds with the only prior study of this type for the E-mini futures: " ...using either time bars or volume bars is more accurate than standard tick-rule classification schemes based on individual trade data" (ELO (2012b), page 3).[24] They reach this conclusion by comparing the *contract level* tick rule to the BVC approach, assessed at *the bar level*. ELO report a 12.4% failure rate for time bars of 120 seconds and 10.0% for volume bars of 8,000 contracts (relative to 13.6% for the transaction tick rule). For us, the corresponding "improvement" is reflected in an error rate at the bucket level of 8.3% for 60-second time bars and 4.7% for $\nu = 0.1$ volume bars (BKU_3^T and BKU_2^V under the MB panel).

The problem is that bar and transaction level error rates are *incompatible*. As documented above, precision is enhanced, realization-by-realization, as we aggregate the order flow.[25] In fact, for identical levels of order flow aggregation, the transaction level tick rule performs, by far, the best. In short, the conclusions of ELO are turned upside down.

6.4 Systematic Time-Variation in Misclassification Rates

We now explore whether the accuracy of the classification varies systematically over time. This is important. If precision deteriorates when return volatility rises, say, then the VPIN metric may be distorted exactly when it is of most interest.

The two top panels of Figure 4 provide error rates for the BKU_1^T and BKU_3^T schemes. The upper panel refers to individual transactions and the second to volume buckets. Aggregation clearly enhances the performance of BKU_1^T greatly, while the improvement for BKU_3^T is more modest. At the bucket level, BKU_1^T dominates uniformly although the discrepancies are most pronounced during high volatility periods (compare with Figure 2).

The TIC^R series is included in each panel to facilitate comparisons to the traditional tick rule. In either scenario, the tick rule uniformly dominates. At the contract level, the precision of the TIC^R approach is near constant. At the bucket level, BKU_1^T starts mimicking the properties of TIC^R and the accuracy appears to improve as volatility increases, inducing a distinct negative correlation with BKU_3^T. It is evident that the active trade classification of BKU_3^T is poor throughout and particularly error prone during volatile market conditions.

The bottom two panels portray misclassification rates for the volume bar BVC procedures. Clearly, the use of volume bars improves accuracy. Moreover, in analogy to our prior findings, BKU_1^V – based on the smaller bar size – inherits some features of the TIC^R series at the bucket

[24] ELO (2012a) cites this piece for details on BVC and its accuracy. This working paper was subsequently replaced by ELO (2012c) which analyzes a different implementation of the BVC procedure.

[25] This point is also documented carefully by Chakrabarty, Pascual and Shkilko (2012).

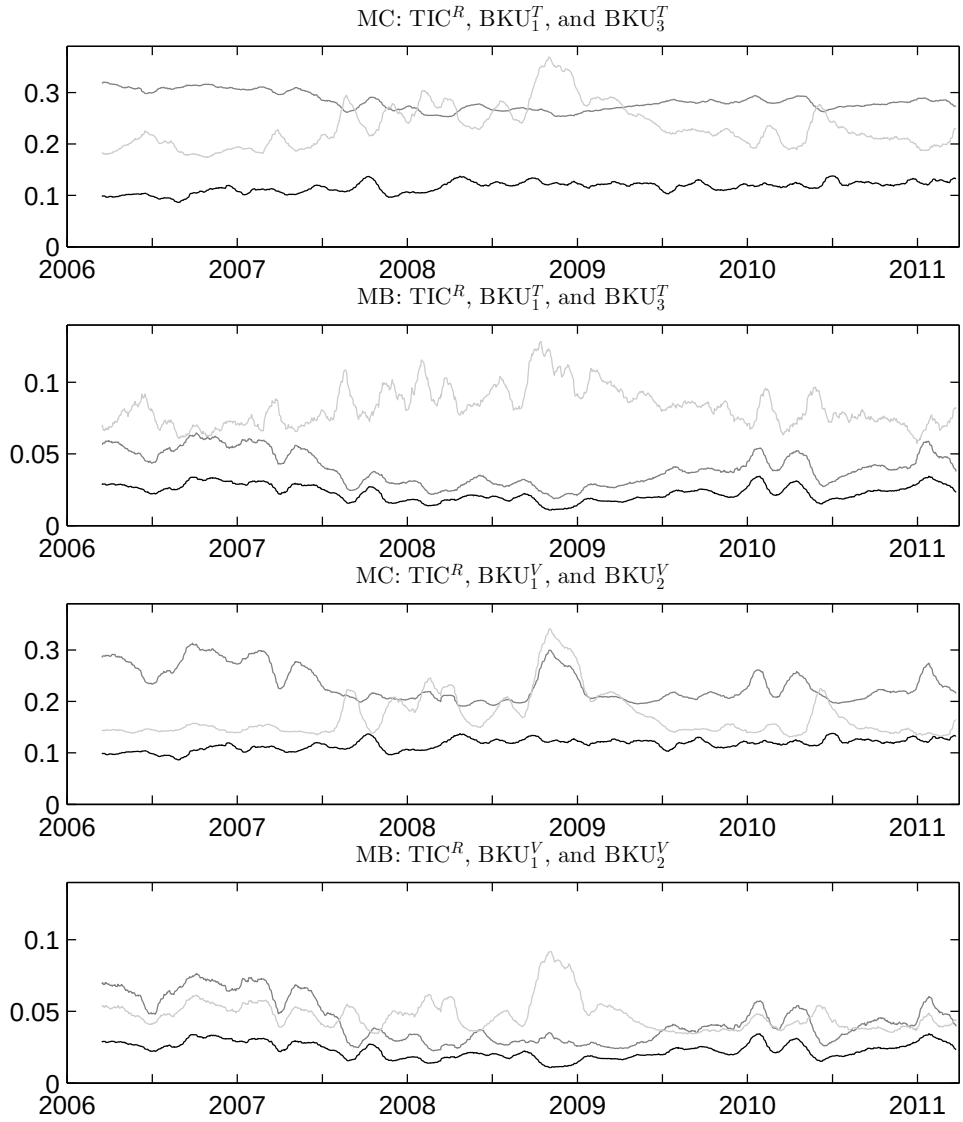

Figure 4: This figure depicts 21-day moving averages of the misclassification rates for different trade classification procedures. The top two panels depict MC and MB for TIC^R, BKU$_1^T$, and BKU$_3^T$, while the bottom two panels show MC and MB for TIC^R, BKU$_1^V$, and BKU$_2^V$. MC stands for misclassification at the contract level and MB denotes misclassification at the volume bucket level. In each panel, the order of the plots is black (first), gray (second), and light gray (third).

level. Furthermore, BKU$_2^V$ retains the qualitative features observed in BKU$_3^T$, but the fluctuations are dampened. Overall, the main conclusion stands: the standard tick rule uniformly dominates the BVC procedures at the bucket level. Moreover, the most unstable approach, in terms of the accuracy being compromised during volatile market episodes, is BKU$_3^T$. This is also the series most closely aligned with the procedure of ELO (2012a).

7 The Empirical Behavior of Alternative VPIN Metrics

The different trade classification schemes generate distinct VPIN metrics, as the associated order imbalances differ. This section explores the properties of alternative VPIN metrics.

7.1 The Evolution of Alternative VPIN Series over a Five-Year Span

Figure 5 depicts a set of different VPIN series across our sample. ACT^R-VPIN in the top panel serves as our benchmark, as it reflects the actual active order flow. We note that TIC^R-VPIN evolves very similarly to ACT^R, although TIC^R almost invariably is at a higher value. Evidently, TIC^R generates order imbalances that are slightly more volatile than they should be. However, the errors are small and homogeneous so the two VPIN series display near identical variation over time, and both reach their *minimum* during the onset of the financial crisis.

Inspecting the lower three panels, we find, quite naturally, that VPIN constructed from BKU_1^T and BKC_1^T resembles ACT^R the most. However, the low frequency variation in BKU_1^T and BKC_1^T is notably smaller than for the transaction-based measures. In comparison, the BKU_3^T and BKC_3^T series display much more short-term variation, and the former is prone to erratic movements in either direction. Overall, there is little coherence between the BKU_3^T or BKC_3^T series and the VPIN measures derived from contract level classification. Finally, the bottom panel shows that there are dramatic differences in the volume-bar BVC-VPIN series, depending on whether the price changes are normalized by an unconditional or conditional volatility factor. BKC_2^V is basically flat, apart from high-frequency oscillations, while BKU_2^V is erratic, attaining extremely highs during the 2008 crisis, and hitting extreme lows in 2006 and 2010. Visually, BKU_2^V, appears inversely related to our benchmark ACT^R series. These observations are confirmed by Table 4, which shows that the ACT^R series is mildly positive correlated with BKC_2^V, but strongly negatively correlated with BKU_2^V.

7.2 The Association of VPIN with Trading Volume and Return Volatility

Section 7.1 shows that the VPIN metric is sensitive to the choice of design variables, such as whether time or volume bars are used, the length of the bars, and whether the price changes are normalized by an unconditional or conditional volatility factor. Section 6 documents that there are substantial discrepancies in the accuracy and stability of the different trade classification schemes. We now explore how the alternative VPIN measures are associated with key market activity variables, namely trading volume and return volatility. We use the "model-free" implied volatility index, VIX, and a realized volatility measure, RV, as proxies for the latter.

7.2.1 Actual VPIN versus Return Volatility and Trading Volume

We start with a key finding. Table 4 reveals a remarkably strong *negative* association of the ACT^R- and TIC^R-VPIN measures with both trading volume and return volatility. This is eye-opening. It implies that VPIN computed from the underlying active (ACT^R) order flow, henceforth labeled *actual VPIN*, has a fundamentally different relation to volatility and volume than the BVC-VPIN of ELO (2012a). Moreover, TIC^R-VPIN mimics the main features of ACT^R-VPIN very well, suggesting that we may use the former as a proxy for the latter and obtain results that are consistent with those derived from actual VPIN.[26]

This striking finding raises two separate questions. First, why do the VPIN measures based on trade classification using aggregate order flow produce diametrically opposite results to the ones exploiting the actual order flow imbalance? Second, why is there such a pronounced negative

[26] VPIN measures are defined over volume buckets, which are not aligned with regular calendar time. In contrast, variables such as VIX, RV, returns, and volume are usually measured in calendar time units. To compute correlations and run predictive regressions, we convert VPIN and other bucket based measures from the volume to the one-minute calendar grid. This can be done in two ways: (1) use the last available reading prior to given calendar time t, or (2) linearly interpolate between two closest readings straddling time t. The first approach uses slightly "stale" VPIN measures, while the second involves a minor "look-ahead" bias but ,empirically, the difference are negligible. We raise the odds of VPIN providing incremental predictive power by adopting the second approach. Moreover, given the break in the activity measures between regular and overnight trading hours, we confine our analysis to data from regular the trading hours.

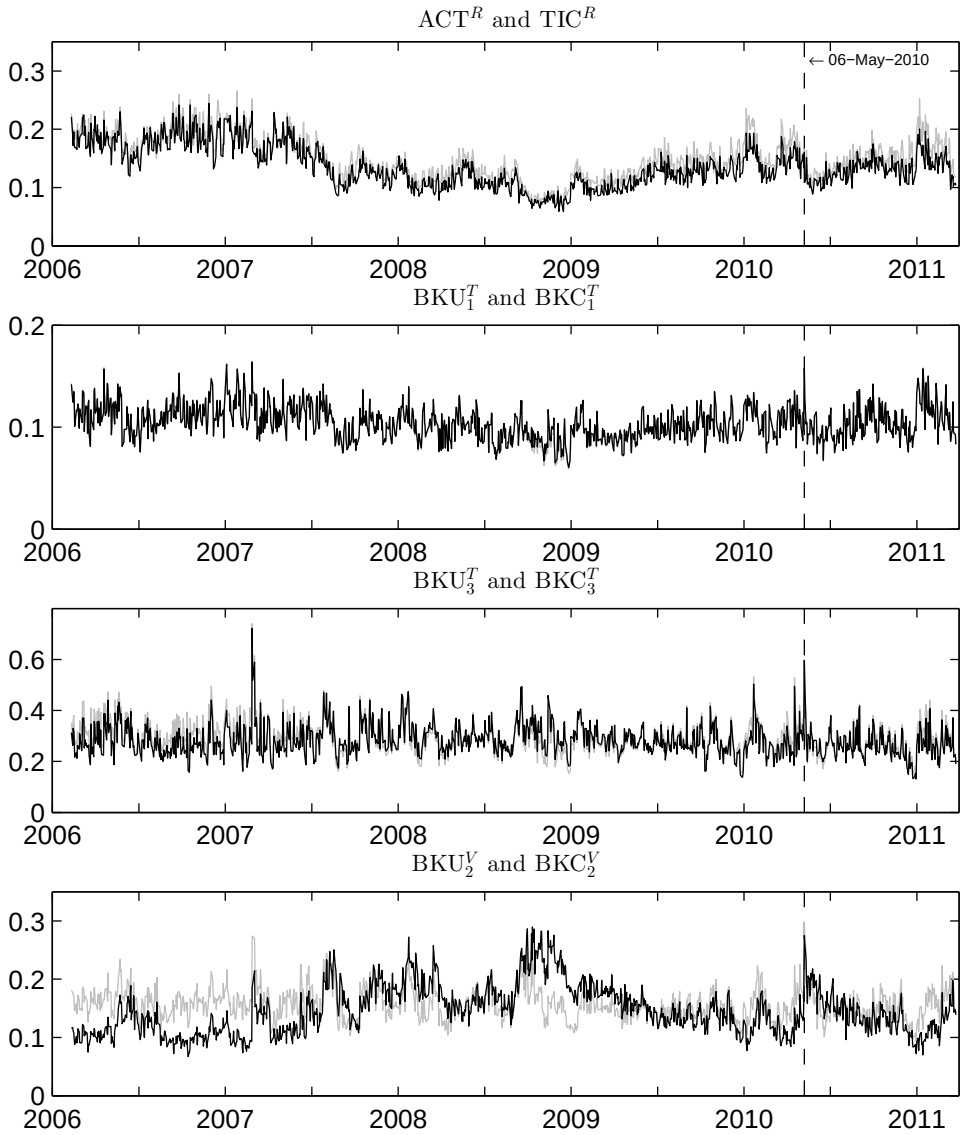

Figure 5: This figure plots the daily maximum values of various VPIN measures, February 10, 2006 - March 22, 2011. In each panel, the order of the plots is black (first) and gray (second).

association between (actual) VPIN and market activity variables? The remainder of this section explores the first issue. A full answer to the second question will take us beyond the current work, but we identify some of the factors behind this phenomenon in Section 9.

7.2.2 Time-Bar VPIN versus Return Volatility and Trading Volume

Table 4 reports correlations of activity variables and VPIN measures based on time bars. TIC^T is much less correlated with volatility than established in AB (2014). This stems from the elimination of the volume trend. Once we control for persistent shifts in volume, TIC^T-VPIN is *negatively* related to volatility, as is the case for the transaction based measures. Second, the correlation of these indices with ACT^R-VPIN declines with the aggregation level, suggesting they deviate more strongly from the actual VPIN metric when classification accuracy deteriorates, as documented in Section 6. Third, this effect is also evident in increasing correlations between TIC^T-VPIN and the activity variables (volume, VIX and RV) as a function of bar size. This is consistent with the

assertion in AB (2014) that persistent yet mean-reverting fluctuations in volume severely distorts time-bar VPIN measures. This stems from the induced variation in the number of bars in the volume buckets. Less bars lead to lower error diversification within the bucket and larger inferred order imbalances.

For additional evidence, we turn to RND^T-VPIN, constructed as in AB (2014). RND^T uses the same bars as the other time-bar VPIN measures, but the trade classification is randomized to generate an i.i.d. series with an equal chance that any bar is a buy or sell. Hence, *ceteris paribus*, it serves as a control for the impact of the trade classification on VPIN. If the dramatic shift in the properties of TIC^T-VPIN across time bars is due primarily to the time variation in trading intensity, and not trade classification per se, the RND^T-VPIN series should display features akin to TIC^T-VPIN. In fact, RND^T-VPIN has nearly the same correlation with ACT^R and TIC^R, and displays the identical correlation patterns with volume, volatility, BKU^T- and BKC^T-VPIN across bar sizes, as TIC^T-VPIN, corroborating our hypothesis.

Moving on to BKU^T- and BKC^T-VPIN, we should encounter the same type of volume effect although the dependence on the price change adds a new feature. Table 4 confirms that the 60- and 300-second bar BKU^T measures are even more *positively* correlated with volume than TIC^T. Moreover, BKC^T is much less correlated with volume due to the normalization of the price change by a realized volatility measure which dampens the inferred order imbalances during periods of (expected) high market activity. Hence, controlling for predictable variation in volatility alters the BKU^T-VPIN series substantially. Moreover, remarkably, RND^T is more highly correlated with ACT^R than BKU^T and BKC^T at every aggregation level. Thus, constructing order imbalances from the size of price changes, as for BKU^T and BKC^T, leads to a weaker association with actual VPIN than what is obtained via random trade assignment.

In summary, our controls for the volume trend and the impact of trade classification, via RND^T-VPIN, enables us to validate the close association between TIC^T-VPIN for small time bars and actual VPIN. Hence, the divergent properties of TIC^T-VPIN for large time bars are consistent with the hypothesis that time-variation in market activity is increasingly more distorted as the order flow is more highly aggregated. We observe similar forces impacting the BKU^T- and BKC^T-VPIN metrics, but the use of absolute price changes in classifying order imbalances introduces a new element which we explore below.

7.2.3 Volume-Bar VPIN versus Return Volatility and Trading Volume

We now focus on VPIN metrics constructed from volume bars. The bottom panel of Figure 7.1 suggests that the BKU^V measure is highly correlated with volatility, while BKC^V is more stable. The correlations assembled in Table 4 corroborate the point. The BKU^V metrics have correlations with VIX and RV exceeding 0.74, while BKC^V is decidedly less correlated with volatility and the correlation with volume shrinks rapidly with bar size. The stronger alignment of BKC^V with RV relative to VIX and the stronger correlations for smaller bars suggest that BKC^V has some commonality with the high-frequency variation in volatility, even if the measures are largely unrelated over longer horizons. Finally, the TIC^V measures display the same negative association to volatility and positive relation with ACT^R-VPIN as documented for small bar TIC^T-VPIN. As a result, their overall properties and correlation with volatility and volume can hardly be more disparate relative to BKU^V- and BKC^V-VPIN.

How do we rationalize the striking disparity in the behavior of TIC^V-VPIN relative to BKU^V- and BKC^V-VPIN? By construction, the TIC^V series annihilates the effect of time-variation in trading volume. Thus, the dominant source of distortion has been removed, and for small bar sizes it should be reasonably compatible with actual VPIN. In fact, for $\nu = 0.02$ the correlation of TIC^V-VPIN with ACT^R- and TIC^R-VPIN is high, and the correlations with BKU^V- and BKC^V-VPIN as well as volume and volatility are remarkably similar to those for ACT^R-VPIN. This reaffirms the

Table 4: **Correlations for VPIN Measures**

Data Aggregation	Rule	ACT^R	TIC^R	Volume	VIX	RV
Transactions	ACT^R	1.00	0.96	-0.53	-0.72	-0.62
	TIC^R	0.96	1.00	-0.53	-0.73	-0.64
Time Bars	TIC_1^T	0.90	0.92	-0.42	-0.66	-0.55
$\delta = 1$ sec	BKU_1^T	0.63	0.62	-0.08	-0.39	-0.24
	BKU_1^T	0.65	0.65	-0.12	-0.44	-0.29
	RND_1^T	0.83	0.87	-0.48	-0.72	-0.64
Time Bars	TIC_2^T	0.76	0.75	-0.16	-0.49	-0.34
$\delta = 10$ sec	BKU_2^T	0.35	0.30	0.30	-0.04	0.15
	BKU_2^T	0.58	0.54	0.08	-0.29	-0.08
	RND_2^T	0.76	0.76	-0.16	-0.53	-0.38
Time Bars	TIC_3^T	0.61	0.60	0.04	-0.34	-0.19
$\delta = 60$ sec	BKU_3^T	0.06	0.01	0.55	0.21	0.38
	BKU_3^T	0.44	0.40	0.27	-0.15	0.05
	RND_3^T	0.55	0.52	0.16	-0.28	-0.10
Time Bars	TIC_4^T	0.46	0.44	0.21	-0.20	-0.04
$\delta = 300$ sec	BKU_4^T	-0.20	-0.26	0.71	0.42	0.55
	BKU_4^T	0.34	0.30	0.38	-0.07	0.14
	RND_4^T	0.40	0.37	0.31	-0.14	0.05
Volume Bars	TIC_1^V	0.84	0.86	-0.45	-0.63	-0.53
$\nu = 0.02$	BKU_1^V	-0.66	-0.71	0.69	0.76	0.80
	BKU_1^V	-0.19	-0.24	0.55	0.28	0.46
Volume Bars	TIC_2^V	0.65	0.66	-0.28	-0.44	-0.35
$\nu = 0.1$	BKU_2^V	-0.65	-0.69	0.68	0.74	0.78
	BKU_2^V	0.12	0.07	0.37	0.04	0.26

Notes: This table reports correlation coefficients computed at the one-minute frequency over regular trading hours. "Volume" is the one-day backward trading volume, "RV" is the realized volatility over the previous day. The VPIN measures are converted to the calendar grid as explained in footnote 26.

critical impact of time-variation in volume on TIC^T-VPIN.

In contrast, for BKU^V- and BKC^V-VPIN, the size of the short-term price changes is a driving force. As such, they may share features with regular realized volatility measures. For concreteness, our discussion focuses on BKU^V-VPIN which is most closely aligned with RV.

In general, equations (1) and (3), plus the fact that $\nu_{q,\ell} = 1/Q$ for volume bars, imply,

$$\text{BKU}^V\text{-VPIN}_\ell \;=\; \frac{1}{L}\sum_{\ell=1}^{L}\left|\frac{1}{Q}\sum_{q=1}^{Q}\gamma_{q,\ell}\right|. \tag{7}$$

Suppose now that the signed order imbalance measure, $\gamma_{q,\ell}$, is a monotone increasing function $f(r)$ of the return over the bar and $f(0) = 0$, i.e., $\gamma_{q,\ell} = f(r_{q,\ell})$, where

$$r_{q,\ell} \;=\; \frac{P_{q,\ell} \,-\, P_{q-1,\ell}}{P_{q-1,\ell}} \;=\; \frac{\Delta P_{q,\ell}}{P_{q-1,\ell}}$$

is the simple return over bar q. We note that the signed order imbalance for BKU^V is an example

of a measure that satisfies the above characterization,

$$\gamma_{q,\ell} \;=\; 2 \cdot Z\left(\frac{P_{q-1,\ell}}{\sigma_{\Delta P}} \cdot r_{q,\ell}\right) - 1 \;=\; f\left(\frac{P_{q-1,\ell}}{\sigma_{\Delta P}} \cdot r_{q,\ell}\right).$$

Of course, a similar characterization applies for realized volatility measures. Specifically, if $f(r) = r$ or $f(r) = \ln(1+r)$, then $\gamma_{q,\ell}$ equals the simple or log returns, respectively.[27]

Since $\gamma_{q,\ell}$ always has the same sign as the return $r_{q,\ell}$, the VPIN metric,

$$\text{BKU}^V\text{-VPIN}_\ell \;=\; \frac{1}{L}\sum_{\ell=1}^{L} |\,\gamma_\ell\,| \;=\; \frac{1}{L}\sum_{\ell=1}^{L}\left|\frac{1}{Q}\sum_{q=1}^{Q} f(r_{q,\ell})\right| \tag{8}$$

will be closely related to realized volatility. To quantify this relation, we measure the coherence between the log-return and the bucket-wide signed order imbalance.[28] For BKU_1^V and BKU_2^V, the correlations of γ_ℓ with $\ln(1+r_\ell)$ are 0.86 and 0.84, respectively. In contrast, for the actual trade classification ACT^R, the correlation is substantially lower at 0.53. Hence, the design of the BVC measures almost mechanically boosts the correlation with return volatility, while it weakens the association with the underlying order imbalances, cf. Section 6.

Lastly, we recall that BKC^V normalizes the price changes with a factor reflecting recent realized volatility. Hence, BKC^V controls for *expected* volatility and is largely immune to persistent shifts in volatility, but remains sensitive to volatility *innovations*. Therefore, we expect a weaker association with volatility for BKC^V than BKU^V. Moreover, BKC^V should be more strongly correlated with RV than VIX, as the former captures volatility realizations more directly. These predictions are consistent with the correlations reported in Table 4.

In short, the BKU^V- and BKC^V-VPIN series inherit many properties from realized volatility. Moreover, since the BKU^T- and BKC^T-VPIN measures share the dependence on price changes, they naturally display similar features, albeit in a weaker form, as, in addition, they are subject to direct volume distortions arising from the time bars. Hence, arguably, the BVC-VPIN metrics are more closely related to RV measures than actual order flow imbalances. Whether these nonlinear filtering procedures and toxicity measures, based on the evolving volume and price change series, are warranted cannot be answered a priori. It is possible they extract valuable information regarding the intensity of informed trading from the real-time transaction record. Ultimately, it hinges on their incremental predictive power for episodic market turbulence beyond what is embedded in standard volatility or volume measures.

7.2.4 Intraday Order Flow Imbalances

Trading volume and return volatility display pronounced intraday patterns. In addition, we have argued that a number of VPIN measures obtained from aggregated order flow, almost mechanically, will be correlated with volume and volatility. It follows that the order flow imbalance measures underlying these VPIN metrics should inherit the same type of intraday activity pattern. Below, we document that this is, indeed, the case. Moreover, we find that the identical qualitative intraday pattern is operative for order imbalances generated by a benchmark classification scheme which exploits the same aggregated order flow but assigns trade direction randomly. The latter is, by construction, driven exclusively by the prevailing trading pattern and cannot, by design, contain information regarding informed trading beyond the trading pattern itself. This implies that the order imbalances are governed by the trading pattern and not the trade classification. Hence, for the

[27] The scaling factor $\frac{P_{q-1,\ell}}{\sigma_{\Delta P}}$ inside the transformation function for BKU^V reflects the use of price changes, and not returns, and the normalization by the standard deviation of the price change in the BVC scheme.

[28] Note, γ_ℓ equals the log return over the entire bucket, i.e., $\ln(1+r_\ell) = \ln\frac{P_{Q,\ell}}{P_{0,\ell}}$.

associated VPIN measures, it is the market activity that matters, while actual order imbalances are secondary. Nonetheless, this is still potentially consistent with these VPIN metrics capturing the intensity of informed trading, as long as the latter also covaries strongly with volume and volatility.

Lastly, we find the actual order flow imbalance to follow a distinctly different intraday pattern. Thus, the order imbalance measures discussed above deviate systematically from the actual imbalance in a manner that is linked to the trading patterns, but not the trade classification rule. There are two possible explanations. Either the order imbalance measures, exploiting highly aggregated order flow, deviate from the overall order flow imbalance due to systematic misclassification which is correlated with market activity, but has no direct association with informed trading. Alternatively, the BVC style order imbalance measures do, indeed, reflect the intensity of informed trading, and the latter is inherently correlated with market activity, while trade classification – surprisingly – is irrelevant in this respect.

We now turn to the evidence regarding the intraday activity patterns. The alternative hypotheses concerning the VPIN measures are explored in the following section.

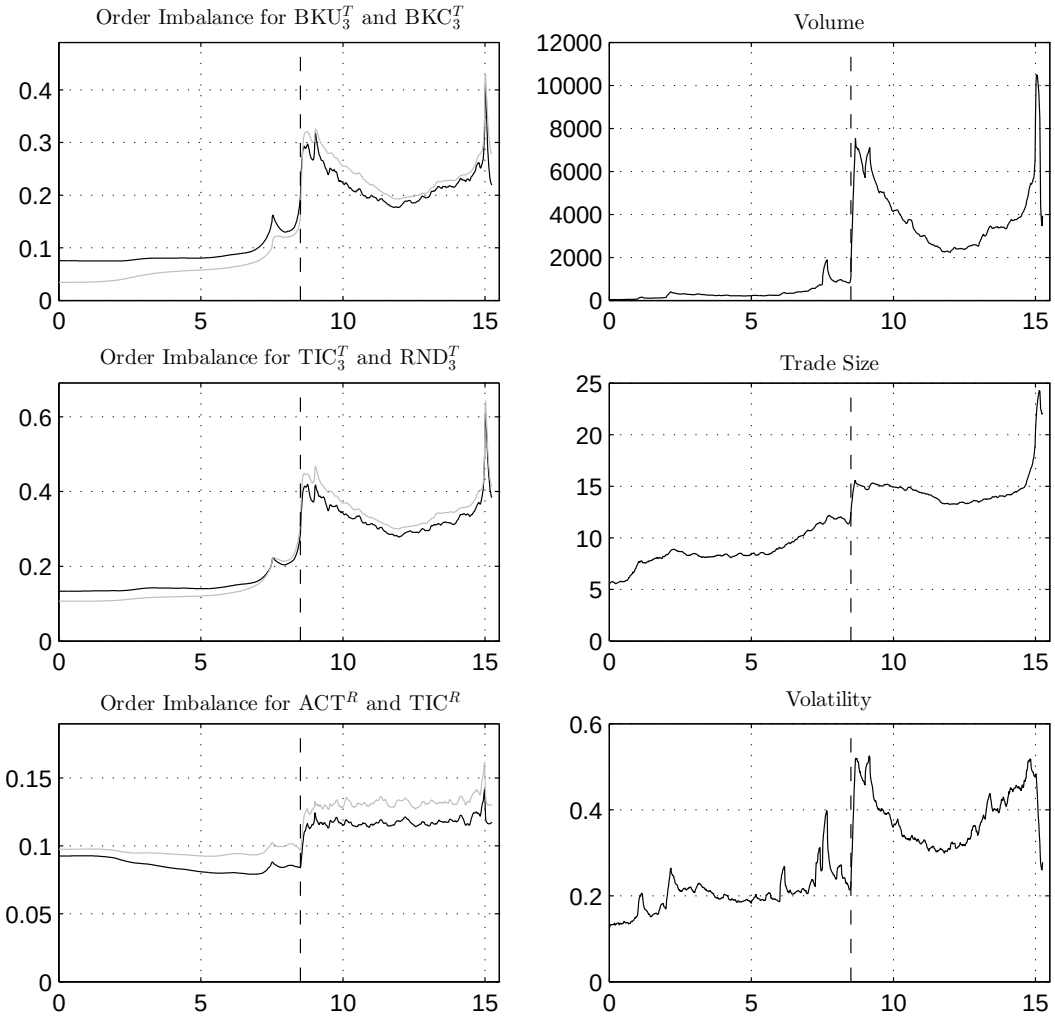

Figure 6: This figure depicts various intraday statistics from 0:00 to 15:15, averaged across all trading days from February 10, 2006 to March 22, 2011. The vertical dashed line represents the start of regular trading hours. Left panels: Order Imbalance (OI) for the classification schemes BKU_3^T, BKC_3^T, TIC_3^T, RND_3^T, ACT^R, and TIC^R, where the order of the plots is black (first) and gray (second). Right panels: trading volume, trade size, and volatility (square root of average return squared), where the statistics are computed at the one-minute frequency and then smoothed over a 10-minute window.

Figure 6 depicts the average intraday evolution of a set of activity measures, including alternative order imbalance measures as well as trading volume, trade size and return volatility. All measures based on one-minute time bars in the two top panels of the left column display a pronounced U-shaped pattern across regular trading hours. These curves resemble a mixture of the corresponding intraday variations in trading volume and trade size in the top panels of the right column. The close association between TIC^T and RND^T corroborates the findings of AB (2014) – the random classification underlying RND^T-OI induces the same intraday behavior as observed for TIC^T-OI. More strikingly, we observe the same pattern for BKU^T and BKC^T as well. That is, the accuracy of the order imbalance measure is immaterial. Purely randomized classification generates the identical pattern, corroborating the hypothesis that variation in overall market activity is the primary determinant of these VPIN measures.

Moving to the ACT^R-OI series in the bottom panel on the left, we find the intraday evolution of the actual order imbalance to be radically different from the depiction above. It is low and stable prior to the regular trading day. Then there is an initial jump at the open, followed by a stable level throughout regular trading hours. Just prior to the equity market closure at 15:00, the ACT^R-OI spikes, presumably due to pressure on traders to reach their desired positions. Subsequently, during the 15:00-15:15 post-trading in the futures market, the OI measure falls back to the prevailing level for the regular trading session. Finally, we note that TIC^R-OI again provides a reliable proxy for the corresponding ACT^R series.

In summary, using the actual underlying order flow, we find no association between the intraday patterns of order imbalances and trading intensity. In contrast, the order imbalances constructed from time bars display a pronounced pattern that mimics the diurnal features in the trading volume and trade size series. This "cross-sectional" finding complements the time series evidence in identifying the variation in trading intensity and volatility as major factors in driving – and potentially distorting – the evolution of time bar VPIN.

8 The Incremental Predictive Power of VPIN

We have demonstrated that VPIN, constructed from actual order imbalances, is strongly *negatively* correlated with return volatility. At the same time, the VPIN metrics proposed by ELO (2011a, 2012a) display a pronounced *positive* correlation with volatility. Our analysis suggests that the discrepancy arises from systematic misclassification of the actual order imbalance by the VPIN implementations relying on aggregate order flow for trade classification. These errors are correlated with market activity, e.g., trading volume and return volatility, thus inducing the positive correlation between VPIN and volatility for large time or volume bars.

These observations render the interpretation of existing empirical evidence ambiguous. The ability of VPIN to predict future return volatility and serve as a crash indicator – if indeed factual – is clearly not linked to the actual underlying order imbalances. Instead, a potential explanation is that the nonlinear VPIN filtering procedure extracts useful information regarding toxic order flow – not regular (non-toxic) order flow – from the observed trade and price patterns. Thus, the true test regarding the usefulness of VPIN, as developed in ELO (2011a, 2012a), is the extent to which the measure provides auxiliary information beyond realized volatility and volume measures in forecasting future volatility and market disruptions.

In this section, we pursue a few different strategies. We test whether the VPIN, in general, has incremental predictive content. We also employ a dummy variable to isolate the predictive power during periods when the VPIN metric is extremely elevated. This amounts to a test of whether extreme VPIN values serve as useful signals regarding future market turmoil. Finally, in the spirit of an event study, we revisit the flash crash and investigate whether the VPIN measure potentially could have helped predict the impending market breakdown.

8.1 Predictive Regressions

This section explores volatility forecast regressions to further assess the properties of the alternative VPIN measures. ELO (2012a) use the positive association between BKU_3^T-VPIN and future volatility to argue that VPIN may serve as a proxy for order flow toxicity.

Table 5 provides predictive regression results for five minutes and one day ahead volatility. The first couple of univariate regressions involve ACT^R and TIC^R. The transactions-based VPIN metrics are robustly and negatively related to future volatility, consistent with our prior evidence. AB (2014) reach the same conclusion and conjecture it arises from a thinning of the order book during volatile periods. We explore this feature further in Section 9.

The next forecast variable, TIC_3^T, is closely related to the VPIN measure studied by ELO (2011a, 2011b, 2011c). AB (2014) document that this metric is correlated with volatility, but the information content regarding future return variation is fully subsumed by the (observable) VIX measure. Moreover, AB found a simple "uninformative" metric to dominate TIC^T in forecasting future volatility although it, by construction, has no relation to order imbalance. Instead, this association arises from the largely mechanical correlation of the measure with trading volume induced by the time bars. Here, we annihilate most of this correlation by volume detrending. The impact is evident in Table 5. Neither TIC^T nor RND^T are positively associated with future absolute returns and their explanatory power is essentially nill. The last VPIN metrics investigated are BKU_3^T and BKC_3^T. Not surprisingly, BKU_3^T has significant predictive power for future volatility – it inherits some of the realized volatility features associated with the BKU^V series. Nonetheless, the explanatory power is limited with R^2 values of about 10% and 12% for the two horizons.[29] Moreover, normalizing the price changes by a measure of recent volatility annihilates the predictive power, as evidenced by the regressions for BKC_3^T. This suggests that BKU_3^T has predictive power due to its correlation with realized volatility which arises from the use of price changes to infer order imbalance. This is consistent with the vastly superior performance of either trading volume, VIX, or RV. For example, volume provides a three-fold and RV a six-fold improvement in explanatory power relative to BKU_3^T at the five-minute horizon, and the discrepancy is even larger at the daily level.

Finally, we include an encompassing regression where BKU_3^T and RV are joint explanatory variables for future volatility. The coefficient on BKU_3^T is now insignificant. It adds nothing to the predictive power of RV. Thus, in general, BKU_3^T-VPIN is irrelevant for predicting future volatility: traditional real-time indicators like realized volatility and trading volume are vastly superior and subsume the information content of VPIN.

8.2 Reverse Causality: Realized Volatility as a Predictor of Future VPIN

The prior section documents that the information content of time bar VPIN is subsumed by realized volatility in terms of predicting future volatility. Since VPIN is a sluggishly moving variable, constructed as a rolling moving average over 50 volume buckets, this result has even stronger implications. It suggests that realized volatility can predict the future VPIN and, furthermore, the predicted component, VPIN^P, is the one with forecast power for future volatility, while the residual component, VPIN^R, possesses little, if any, predictive power.

To test this conjecture, we first construct a predictive regression for VPIN, based solely on RV observations obtained $\Delta = 5$ minutes earlier,

$$\text{VPIN}_t^P \;=\; a_0 \,+\, a_1 \cdot \text{RV}_{t-\Delta} \,+\, \text{VPIN}_t^R. \tag{9}$$

Next, we generalize the predictive regression from Table 5 involving BKU_3^T-VPIN by jointly including the predictive and residual component of BKU_3^T-VPIN as regressors. The findings are reported

[29] Ironically, the forecast power associated with the *negative* relation between TIC^R and future volatility is substantially higher that implied by the positive correlation between BKU_3^T and future volatility.

Table 5: **Forecast Regressions for Average Absolute Return**

Panel A: 5-Minute Forecast

Reg	Const	ACT^R	TIC^R	TIC_3^T	BKU_3^T	BKC_3^T	RND_3^T	Vol	VIX	RV	$\bar{R}^2$
(1)	0.10	-0.50									19.71
	(25.50)	(-17.90)									
(2)	0.11		-0.52								22.35
	(25.43)		(-18.47)								
(3)	0.06			-0.07							1.32
	(14.90)			(-6.25)							
(4)	-0.02				0.22						9.91
	(-4.03)				(13.04)						
(5)	0.03					0.05					0.49
	(8.12)					(3.50)					
(6)	0.05						-0.04				0.27
	(9.66)						(-2.71)				
(7)	-0.01							0.26			33.42
	(-4.27)							(19.66)			
(8)	-0.01								0.22		48.42
	(-8.31)								(30.66)		
(9)	0.00									0.13	61.35
	(3.17)									(98.96)	
(10)	0.00				0.00					0.13	61.35
	(1.81)				(0.54)					(92.26)	

Panel B: 1-Day Forecast

Reg	Const	ACT^R	TIC^R	TIC_3^T	BKU_3^T	BKC_3^T	RND_3^T	Vol	VIX	RV	$\bar{R}^2$
(1)	0.10	-0.50									33.30
	(26.14)	(-18.34)									
(2)	0.11		-0.52								37.40
	(25.98)		(-18.84)								
(3)	0.07			-0.08							3.36
	(16.27)			(-7.82)							
(4)	-0.01				0.18						11.87
	(-1.94)				(12.14)						
(5)	0.04					0.02					0.09
	(11.04)					(1.28)					
(6)	0.06						-0.06				1.15
	(11.48)						(-4.49)				
(7)	-0.01							0.24			48.02
	(-2.66)							(18.97)			
(8)	-0.01								0.21		77.85
	(-8.33)								(32.87)		
(9)	0.00									0.18	82.30
	(0.56)									(26.33)	
(10)	0.00				0.00					0.18	82.30
	(0.06)				(0.46)					(24.60)	

Notes: This table reports OLS forecast regressions for the average absolute one-minute return (AAR), defined as $AAR(t, t+T) = 1/T \sum_{i=1}^{T} |r_{t+i}|$ and $r_t = 100 \ln(P_t/P_{t-1})$ is the one-minute return. Forecasts are formed every 5 minutes during regular trading hours. The forecast horizon $T = 5$ minutes or 1 day (defined as 405 one-minute intervals during regular trading). "Vol" is the one-day backward trading volume, "RV" is the realized volatility over the preceding one hour (Panel A) or one day (Panel B). The VPIN measures are converted to the calendar grid as explained in footnote 26. The t-statistics in parentheses reflect HAC-standard errors with 81 lags.

in the first row of Table 6. The anticipated component is highly significant and the residual component is insignificant. That is, the only relevant information in BKU_3^T-VPIN is what, a priori, may be distilled from past realized volatility. Furthermore, RV explains only a small part of the overall variation in VPIN, with an R^2 of about 14%, while the explanatory power of equation (9) is almost identical to that of RV itself (R^2 of 82%) in Table 5. Hence, the variation in RV gets encoded almost one-for-one in BKU_3^T-VPIN, but the measure is primarily driven by separate forces unrelated to future RV. In sum, the variation in VPIN not directly associated with past RV only serves to obfuscate the relevant information, i.e., the source of predictive power in VPIN is the variation stemming from lagged realized volatility.

Table 6: **Two-Stage Predictive Regressions for one-day-ahead Realized Volatility**

Reg	Const	VPIN^P	VPIN^R	$\text{VPIN}^P\cdot\text{D}$	$\text{VPIN}^R\cdot\text{D}$	$\bar{R}^2$
(1)	-1.41	6.55	-0.04			81.97
	(-25.94)	(29.38)	(-1.33)			
(2)	-1.41	6.57	-0.04	-0.21	0.30	82.00
	(-25.68)	(29.07)	(-1.30)	(-1.62)	(1.77)	

Notes: The table reports the results of a two-stage predictive regression for one-day-ahead realized volatility (RV). In the first stage, VPIN is projected onto 5-minute lagged RV to obtain the "projected" and "residual" components of VPIN, i.e., $\text{VPIN} = \text{VPIN}^P + \text{VPIN}^R$. Specifically, we obtain $\text{VPIN}^P = 0.22 + 0.14\,\text{RV}$, with $\bar{R}^2 = 14.02\%$. In the second stage, VPIN^P and VPIN^R are used to forecast future RV. The forecast horizon is $T = 1$ day (defined as 405 minutes of regular trading hours). Forecast are formed every 5 minutes during regular trading hours. For VPIN, we use the BKU_3^T metric, which is converted to the calendar grid as explained in footnote 26. In specification (2), we also introduce the dummy variable D, which equals one when VPIN exceeds the $99th$ percentile, and zero otherwise. Shown in parentheses are t-statistics based on HAC-standard errors with 81 lags.

These findings suggest reverse causality: it is the increase in market activity that drives up the VPIN metrics generated from highly aggregated order flow. Simultaneously, the persistence in volatility and volume implies that high market activity in the recent past predicts elevated future volatility. Hence, there must also be a significant correlation between current VPIN and future volatility, but this predictable component is entirely accounted for by past realized volatility. As such, realized volatility drives VPIN and not the other way around.

8.3 Extreme Realizations of VPIN

So far, we have studied general properties of VPIN. One specific goal of ELO (2011a, 2012a) is to develop a warning signal for impending market turmoil. As such, the association between extreme readings of VPIN and subsequent market conditions is of particular interest.

8.3.1 General Statistical Evidence

To explore whether extreme VPIN readings are particularly useful as signals for future market turmoil, we include a dummy variable in the two-stage predictive regression, indicating times at which VPIN takes values in the top 99% percentile of the distribution. The second row of Table 6 shows that these observations carry no additional predictive power for future return volatility over-and-above what is already encoded in the "predictive" component associated with the past variation in RV. Likewise, the increase in R^2 is negligible. Hence, in general, we find no incremental information associated with truly extreme VPIN realizations.

8.3.2 An Event Study: The Flash Crash

ELO (2011a) find that VPIN attains a historical high during the day of the flash crash. Figure 3 reveals that, for our BKU_3^T-VPIN series, this distinction instead belongs to February 27, 2007. Furthermore, two days later, on March 1, 2007, we encounter another extreme VPIN value.[30] Nonetheless, to facilitate comparison with ELO, we focus on the flash crash. Qualitatively similar results are obtained for the other dates with extreme VPIN realizations.

The panels in the top rows of Figure 7 depict alternative VPIN series, while the panels in the bottom rows refer to activity variables during the regular trading hours on May 6, 2010. The figure exploits the same timing convention for VPIN as in Figure 1. The price series, replicated from Figure 1, serves as a reference for the other measures.

Initially, we note that the VIX and price series are near mirror images up through the crash. Thus, VIX responds instantaneously to the price development, but does not provide auxiliary information beyond the price path.[31] The normalized volatility depicts the standard deviation of one-minute price changes obtained over a (centered) 10-minute window divided by $\bar{\sigma}_{\Delta P}$, the unconditional volatility used to construct the BVC based VPIN measures. The normalized volume series refers to trading volume per one-minute bar divided by the size of the volume bucket. When this value exceeds unity, trading is so vigorous that buckets are filled in less than one minute, and there is only one or two (broken) time bars in each bucket.

We first focus on ACT^R-VPIN. This metric is basically flat during the crash. Thus, the VPIN metric based on the actual order imbalances does *not* signal rising order flow toxicity – even if the signed order imbalance (SOI) in Figure 1, generated from the identical classification, indicates active selling pressure from around 12:00. Moreover, we observe the same qualitatively development for TIC^R-VPIN, which is constructed from trade data alone. In short, the VPIN transformation of the underlying SOI *destroys* whatever order flow information is conveyed by SOI. As such, the time bar VPIN metrics can only generate a positive association with market turbulence by misclassifying order flow in a manner that is correlated with volatility and volume. Of course, this is consistent with our general findings in Section 6 regarding trade classification based on larger sized time bars.

Inspecting the evolution of the remaining VPIN measures in Figure 7, we corroborate these predictions. As the time bars grow larger, the TIC^T, RND^T and BKU^T metrics mimic the volatility and volume patterns more closely. In addition, there is little sign of any predictive content regarding an upcoming crash. The dramatic increases in VPIN occur strictly *after* the initiation of the crash at 13:32, when both volume and realized volatility also spike. This is almost a tautology, as VPIN is generated by a moving average of past order imbalance measures. When large volume and volatility innovations occur, they receive only $(1/50)^{\text{th}}$ of the weight in the VPIN computation. Large increases in VPIN are only feasible if there is a persistent upward shift in the level of OI, and when this occurs, the effect is long-lived due to the (moving average) smoothing. If (time bar) VPIN is to predict a crash, the order imbalance measures must rise well in advance of the event. Consequently, sustained bursts in market activity precede sharp increases in (time bar) VPIN. As such, these activity variables should be superior predictors of impending volatility which is, of course, what the predictive regressions in Section 8.1 and 8.2 show. Specifically, our BKU_3^T metric is approximately 0.292 at 12:30, about an hour prior to the crash, and 0.365 at 13:30. These values fall in the percentiles 62 and 91 of the empirical distribution of daily maximum values of BKU_3^T, as displayed in Figure 3, so we would tend to observe such values eight and two times *every month*, respectively. Only the post-crash realizations are exceptional, but these values reflect, rather than predict, the crash. At the same time, ACT^R-VPIN computed from the actual order imbalance

[30] This "dethroning" of the flash crash is contingent on our "detrended V" approach. Under the ELO (2012a) "constant V" procedure, Figure 3 shows that the maximal VPIN value is attained during the flash crash.

[31] The subsequent extreme oscillations in the VIX measure arise from rapid fluctuations in the liquidity of the S&P 500 option market in the aftermath of the crash; see Andersen, Bondarenko and Gonzalez-Perez (2011).

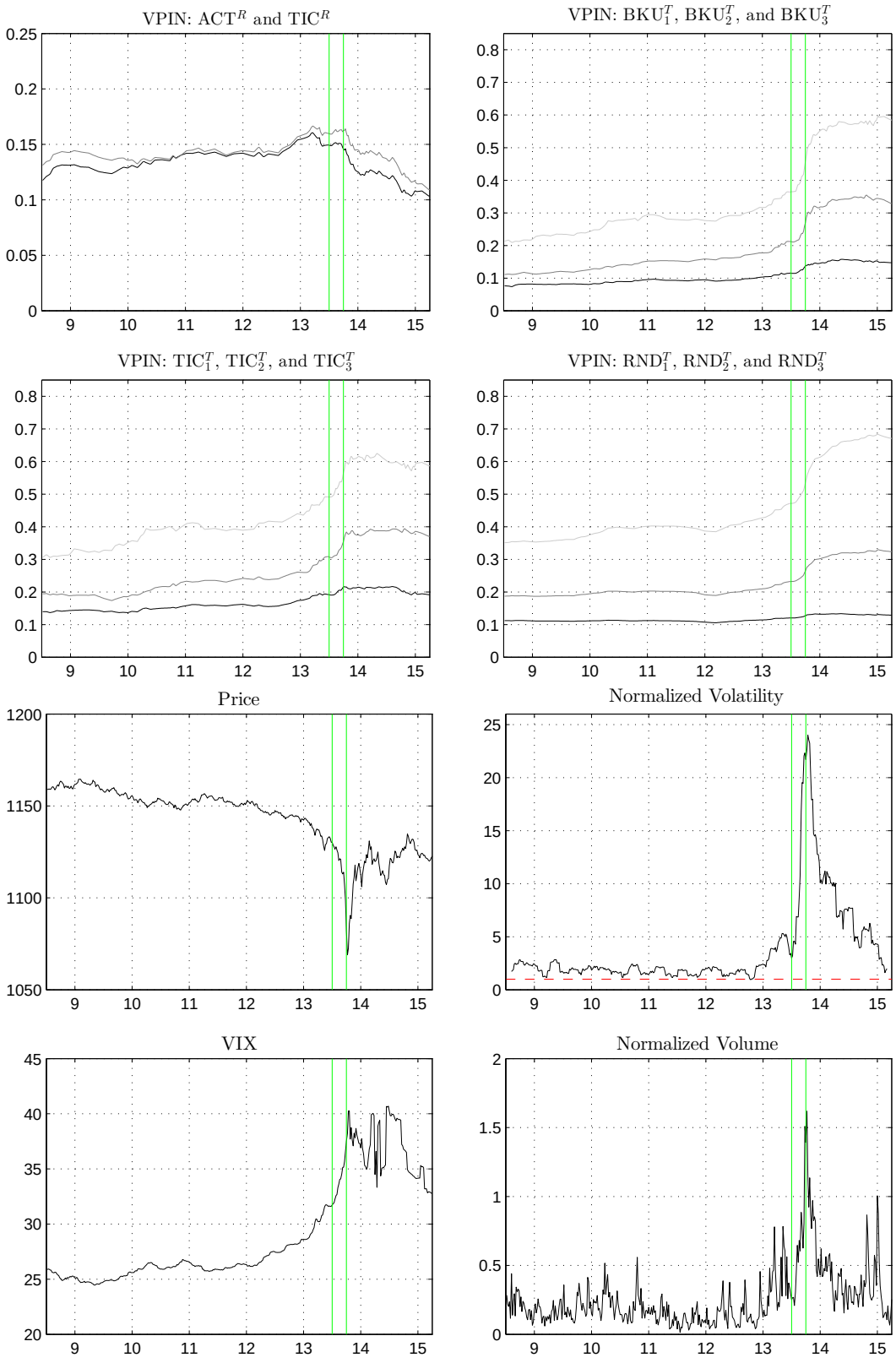

Figure 7: May 6, 2010. The solid vertical lines indicate the timing of the flash crash. In panels with multiple plots, the order of the plots is blue (first), green (second), and red (third). The normalized volatility is the ratio of the standard deviation of one-minute price changes for a (centered) 10-minute window over unconditional volatility $\bar{\sigma}_{\Delta P}$. The normalized volume is the trading volume per one-minute bar divided by the size of the volume bucket.

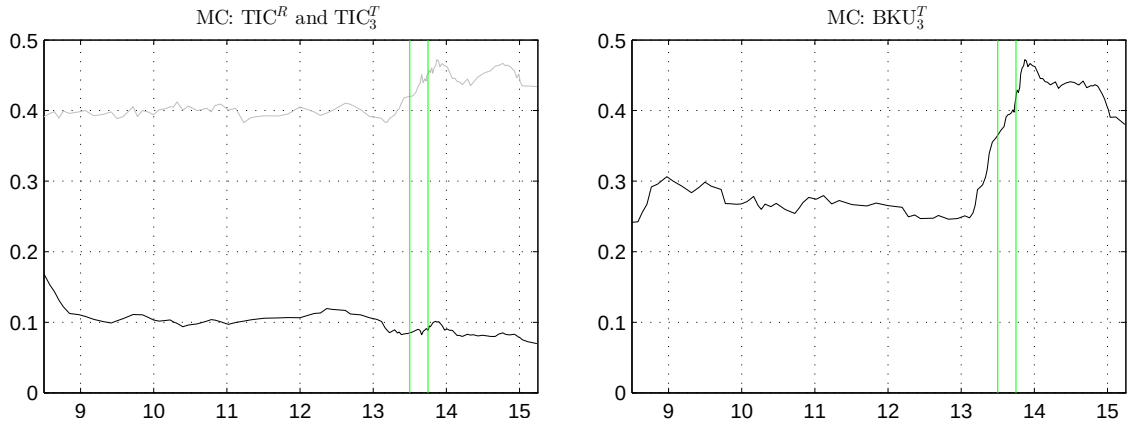

Figure 8: May 6, 2010. MC is computed as the moving average over 10 volume buckets. The solid vertical lines indicate the timing of the flash crash. MC stands for misclassification at the contract-by-contract level. In the first panel, the order of the plots is black (first) and gray (second).

measure drops to values well below the sample average. The extreme movement in BKU$_3^T$-VPIN is a testament to the volume- and volatility-induced distortions in the time-bar based order imbalance measures.

To further illustrate this effect, Figure 8 depicts the contract level misclassification rates for TIC^R and TIC$_3^T$ in the left and BKU$_3^T$ in the right panel. While the TIC$_3^T$ classification is poor throughout, there is a sharp deterioration in precision for BKU$_3^T$ just prior to the crash. This coincides with the explosive growth in the normalized volatility and volume in Figure 7. As before, the time bar BVC order imbalance measure responds primarily to concurrent volatility and volume innovations and becomes increasingly distorted as the market activity rises. As documented previously, this inflation in the VPIN metric possesses no incremental power for future volatility and is, in this respect, fully subsumed by the activity variables.

9 Why Are Actual VPIN and Volatility Inversely Correlated?

Our empirical analysis has documented a surprising fact: VPIN computed from actual order imbalance is significantly *negatively* correlated with return volatility. This must stem from systematic features that are unaccounted for by the theory behind the VPIN metric in ELO (2011a, 2012a). Combined with the lack of incremental predictive power for future volatility, this finding effectively invalidates VPIN as a useful indicator of impending market stress.

This still leaves the fundamental question of why the VPIN metric based on the actual order imbalance is inversely related to return volatility unanswered. A thorough analysis of this phenomenon will take us beyond the scope of this study. Nonetheless, we identify the empirical regularities in the tick-by-tick data that induce the negative association, thus motivating and facilitating future work on rationalizing these robust features of the data.

9.1 Trade Size and Clustering in the Trade Direction

The primary determinant behind our ACT^R-VPIN metric is, tautologically, the actual order imbalance across the volume buckets. Thus, we need to understand the clustering of active buy and sell transactions. At the micro level, a basic driver of clustering is the number of contracts exchanged per transaction.[32] When more than one contract is bought or sold, it constitutes a sequence of

[32]Recall from Section 2.1 that our trade size figures reflect the size of the transactions consummated by the active trader who effectively demands liquidity. Thus, each transaction may involve multiple counter-parties who provide liquidity by posting quotes on the limit order book.

unidirectional trades of individual contracts. Likewise, it matters whether there is a tendency towards continuation rather than reversal of the trade direction, i.e., whether a buy transaction is more likely to be followed by another buy rather than a sell. These two features, the (average) size of transactions and the probability of a continuation are, jointly, key determinants of the buy-sell error diversification. The larger the transactions and the more clustered the trade direction, the fewer effective buy and sell sequences we have over the bucket, and the larger the imputed order flow imbalances tend to be.

We explore the trade sequence dynamics empirically, using our actual order imbalance measure. For each volume bucket, we define the Average Trade Size (ATS), the Average Trade Run (ATR), i.e., the average number of consecutive trades in the same direction, and the Average Volume Run (AVR), defined as the average number of consecutive contracts traded in the same direction. These measures are closely related, as AVR = ATR $\cdot$ ATS. A large value for AVR is tantamount to fewer, but longer, buy and sell sequences in a bucket. This may occur due to a high trade size, ATS, or elevated trade continuation, ATR.[33]

Table 7: **Correlations for ATS and ATR**

		ATS	ATR	Volume	VIX	RV
Whole sample	ATS	1.00	-0.36	-0.51	-0.79	-0.66
	ATR	-0.36	1.00	0.17	0.16	-0.01
Second half	ATS	1.00	0.48	-0.28	-0.84	-0.72
	ATR	0.48	1.00	-0.45	-0.61	-0.66

Notes: This table reports various correlation coefficients computed at one-minute frequency. "Volume" is the one-day backward trading volume, "RV" is the realized volatility over the preceding day. The correlations are computed over regular trading hours only.

Table 7 documents a pronounced negative association between the average trade size and return volatility. In contrast, the trade runs are only mildly correlated with volatility, so the net effect is a strong negative correlation between volatility and AVR, and thus also ACT^R-VPIN and volatility. These findings are driven by the fact that, as volatility rises the trade size tends to drop, and the buy and sell sequences alternate more quickly within each bucket.

The mild negative correlation between ATS and ATR tends to stabilize AVR and the order imbalance measure. However, Panel B documents that, in the second part of our sample, ATR is *positively* correlated with ATS, so they reinforce each other in generating variation in AVR and VPIN. Moreover, since both are now inversely related to volatility, the negative association between volatility and ACT^R-VPIN strengthens. The pronounced negative correlation between trade size and volatility is also evident from the top panel of Figure 9.

From a purely mechanical perspective, the negative correlation between volatility and average trade size may reflect either a drop in the proportion of very large trades or a reduction in the trade size across the full range of the distribution, as volatility rises. To shed light on the issue, we introduce DTS, the dispersion in the trade size, defined as the standard deviation of the trade size across a given volume bucket. The bottom panel of Figure 9 shows that DTS is almost perfectly synchronized with ATS. Hence, when volatility rises, the trade size dispersion declines along with the trade size itself, indicating that the transactions involve fewer contracts across the board.

Consequently, beyond rationalizing and quantifying the inverse relation between volatility and trade size, a challenge for future research is to identify the forces behind fluctuations in ATR. One potential explanation for the large change in the correlation between the first and second half of the sample is the growth of high-frequency trading (HFT). HFT firms are known to trade very

[33]The probability of a trade continuation (PC) is directly related to ATR as PC = 1 - 1/ATR.

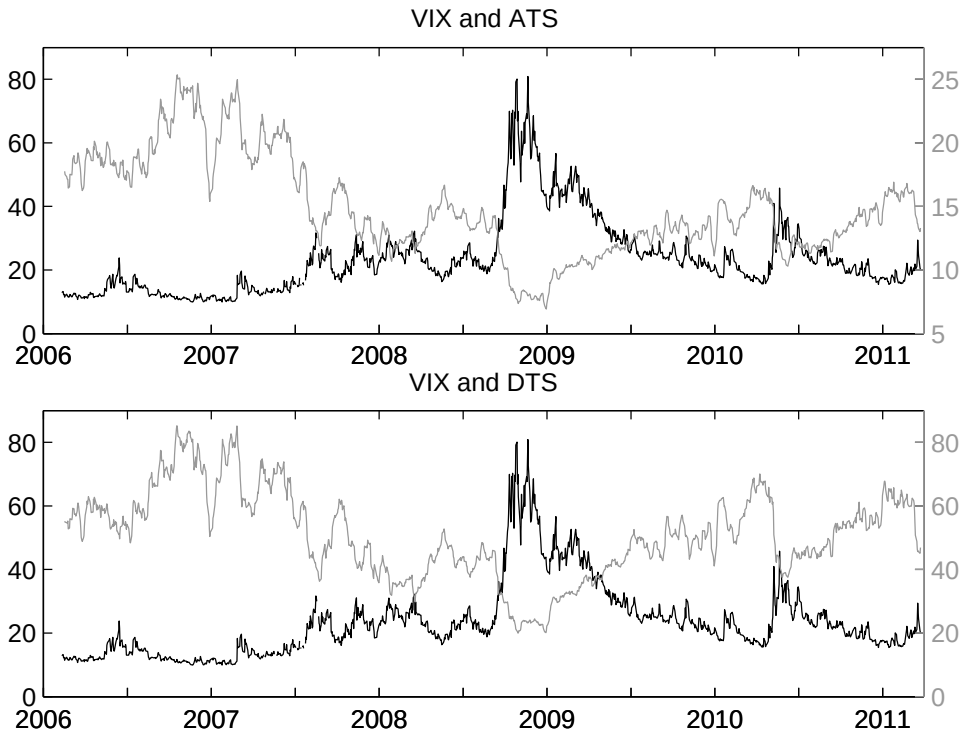

Figure 9: The top panel depicts VIX (left scale, blue) and ATS (right scale, green) and the bottom panel VIX (left scale, blue) and DTS (right scale, green). ATS is the average trade size and DTS is the trade size dispersion. ATS and DTS are obtained as five-day moving averages.

frequently while operating under strict risk limits on their net positions. If such exposure limits, more generally, are designed to reflect current risks, they should also be responsive to volatility, i.e., the position limits should tighten as volatility increases. This may involve less risk taking by lowering the size of active trades as well as more aggressive trading back towards a net balanced position in response to "passive" trades that execute against their limit orders. In this manner, an increasingly volatile setting can induce a drop in the average trade size, a drop in the dispersion of the trade size, and an elevation of the trading intensity. While this squares well with our empirical observations, we cannot verify the basic mechanism without access to individual trading accounts. Hence, it remains largely speculative to associate these developments with the presence of HFT.

9.2 Evidence from the Flash Crash

ELO (2011a) stress that historically high VPIN readings could have served as a warning signal for the flash crash. As such, our finding of a negative association between actual VPIN and return volatility in Section 8.3.2 is particularly striking. To understand whether the preceding analysis also applies to the market dynamics during this uniquely stressful episode, we depict the evolution of the relevant measures for the day of the flash crash in Figure 10.

The figure reveals that the factors identified above are out in full force. Prior to the crash, from 11:00-13:00, the average trade size is above 15 contracts and the average trade run just below four, generating alternating sequence of buys and sells of about 60 contracts on average. In the run-up, during, and in the aftermath of the crash, this number drops precipitously and ends up below 30, inducing a lower order imbalance measure. In reality, of course, the crash is characterized by selling pressure, as conveyed by Figure 1. However, this manifests itself through a rapid escalation in volume and only a small imbalance per bucket. The cumulative effect generated by a small

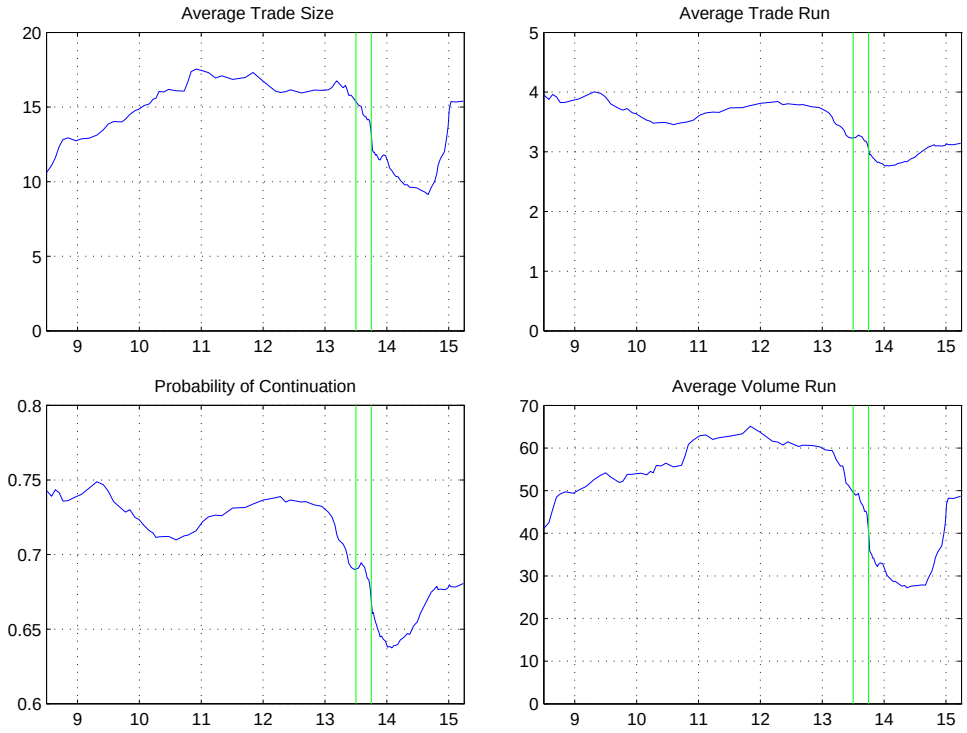

Figure 10: This figure depicts ATS, ATR, PC, and AVR on May 6, 2010. The measures are computed as moving averages over 10 volume buckets. The vertical lines indicate the timing of the flash crash.

persistent imbalance is not reflected in the actual VPIN metric, as the offset due to the drop in the trade size and trade run dominates.

In summary, the VPIN metrics compiled by ELO (2011a, 2012a) attain inflated values on the day of the flash crash for the wrong reasons. The VPIN measure constructed from the actual order imbalances *dives rather than soars*. The VPIN metrics generated from trade classification based on large bar sizes display the exact opposite correlation with volatility only because they systematically misclassify the order flow. Moreover, this does not reflect an ability to detect an increase in the toxic order flow. In fact, the randomized trade classification measure RND_3^T is, by construction bereft of information beyond the trading pattern, yet behaves similarly, as documented in Figure 7. Thus, even in the extreme case of the flash crash, the general features identified above are operative. The ELO VPIN contains no incremental information relative to the contemporaneously observed volume and volatility series prior to the crash, and therefore cannot serve any meaningful role as a supplementary warning signal.

10 Conclusion

This paper seeks to settle a brewing controversy regarding the usefulness of the VPIN metric as a real-time indicator of order flow toxicity and, by extension, predictor of impending market stress. We follow the basic construction of VPIN in ELO (2012a), but provide a few modifications to avoid excessive distortions in the metric due to the pronounced volume trend over our sample. One main innovation is to construct VPIN from a highly accurate classification scheme for active buys and sells. This metric provides a benchmark VPIN which can be used to assess the implications of alternative trade classification techniques. We document a systematic deterioration in the classification accuracy, as the order flow aggregation increases. ELO (2012a) argue that classification

via such highly aggregated order flow is necessary to obtain a useful measure of order imbalances associated with informed trading. In contrast, we find these classification schemes to be driven, almost mechanically, by the level of market activity, and not by any meaningful notion of order flow imbalance.

Our findings suggest that the VPIN metric has no useful role in capturing order flow toxicity for the S&P 500 futures market. Empirical results reaching the opposite conclusion are based on distorted metrics that induce correlation with volume and volatility, by construction. The predictive content of these measures for future volatility are subsumed by traditional real-time indicators, such as realized volatility. In fact, VPIN constructed from actual order imbalances displays a pronounced negative association with return volatility. The bottom line is that the signed cumulative order flow, based on accurate trade classification, has an economically meaningful relation to concurrent price movements. Prices rise (fall) when there is excess active buying (selling). Unfortunately, applying the nonlinear VPIN transformation to the actual signed order flow generates a metric that is negatively correlated with future volatility and has no independent explanatory power for impending market disruptions.

We briefly explore the forces generating the pronounced negative correlation between transaction-based VPIN and volatility. We find this feature to stem from a strong inverse relation between volatility and trade size, induced by a downward shift across the entire transaction size distribution. This is consistent with a deliberate reduction in net positions during times when volatility rises, thus potentially reflecting an attempt to avoid heightening the risk exposures as markets grow increasingly turbulent. Of course, such strategies should also manifest themselves in a change in the submission and cancelation of limit orders. In addition, the signed (actual) cumulative trade imbalance and the short-term price trend should also be indicative of emerging tensions in the market. Future research may find variables capturing the interaction between the price trend, order book dynamics and trade size to be useful in constructing real-time warning signals for market stress.

References

Abad, D., Yague, J., 2012. "From PIN to VPIN: An introduction to order flow toxicity," *The Spanish Review of Financial Economics* 10, pp. 74–83.

Aitken, M., and A. Frino, 1996, "The Accuracy of the Tick Test: Evidence from the Australian Stock Exchange," *Journal of Banking and Finance*, 20, pp. 1715-1729.

Andersen, T.G., 1996, "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility," *Journal of Finance*, 51, pp. 169-204.

Andersen, T.G., and T. Bollerslev, 1998, "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review*, 39, pp. 885-905.

Andersen, T.G., and O. Bondarenko, 2007, "Construction and Interpretation of Model-Free Implied Volatility," *Volatility as an Asset Class*, London: Risk Books, I. Nelken (editor), pp. 141-181.

Andersen, T.G., and O. Bondarenko, 2014, "VPIN and the Flash Crash," *Journal of Financial Markets*, 17, pp. 1-46.

Andersen, T.G., O. Bondarenko, and M.T. Gonzalez-Perez, 2011, "A Corridor Fix for VIX: Constructing a Coherent Model-Free Option Implied Volatility Measure," Working Paper.

Asquith, P., R. Oman and C. Safaya (2010), "Short Sales and Trade Classification Algorithms," *Journal of Financial Markets*, 13, pp. 157-173.

Boehmer, K., J. Grammig, and E. Thiessen, 2007, Estimating the Probability of Informed Trading – Does Trade Misclassification Matter?" *Journal of Financial Markets*, 10, pp. 26-47.

Bethel, E. W., Leinweber. D., Rubel, O., and K. Wu, 2012. "Federal Market Information Technology in the Post-Flash Crash Era: Roles for Supercomputing," *Journal of Trading*.

CFTC, 2013, "Concept Release on Risk Controls and System Safeguards for Automated Trading Environments," *Commodity Futures Trading Commission* , September 9, 2013.

CFTC-SEC, 2010, "Preliminary Findings Regarding the Market Events of May 6, 2010," *Joint Commodity Futures Trading Commission (CFTC) and the Securities and Exchange Commission (SEC) Advisory Committee on Emerging Regulatory Issues*, May 18, 2010.

Chakrabarty, B., R. Li, V. Nguyen, and R. Van Ness, 2007, "Trade Classification Algorithms for Electronic Communications Network Trades," *Journal of Banking and Finance*, 31, pp. 3806-3821.

Chakrabarty, B., P.C. Moulton and A. Shkilko (2012), "Short Sales, Long Sales, and the Lee-Ready Trade Classification Algorithm Revisited," *Journal of Financial Markets*, 15, pp. 467-491.

Chakrabarty, B., R. Pascual and A. Shkilko, 2012, "Trade Classification Algorithms: A Horse Race between the Bulk-based and the Tick-based Rules," Working Paper, SSRN, December 2012

Clark, P.K., 1973, "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," *Econometrica*, 41, pp. 135-155.

Cont, K., A. Kukanov, and S. Stoikov, 2014, "The Price Impact of Order Book Events," *Journal of Financial Econometrics*, 12, pp. 47-88.

Corcoran, C., 2013. "Systemic Liquidity Risk and Bipolar Markets: Wealth Management in Today's Macro Risk On / Risk Off Financial Environment." Wiley.

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

Easley, D., M. López de Prado, and M. O'Hara, 2011a,"The Microstructure of the "Flash Crash": Flow Toxicity, Liquidity Crashes, and the Probability of Informed Trading," *Journal of Portfolio Management*, 37 (2), pp. 118-128.

Easley, D., M. López de Prado, and M. O'Hara, 2011b, "The Exchange of Flow Toxicity," *Journal of Trading* 6 (2), pp. 8-13.

Easley, D., M. López de Prado, and M. O'Hara, 2011c, "Measuring Flow Toxicity in a High-Frequency World," Working Paper, SSRN, February, 2011.

Easley, D., M. López de Prado, and M. O'Hara, 2012a, "Flow Toxicity and Liquidity in a High-Frequency World," *Review of Financial Studies* 25, pp. 1457-1493.

Easley, D., M. López de Prado, and M. O'Hara, 2012b, "Bulk Classification of Trading Activity," Working Paper, SSRN, March 2012.

Easley, D., M. López de Prado, and M. O'Hara, 2012c, "Bulk Classification of Trading Activity," Working Paper, SSRN, August 2012.

Easley, D., M. López de Prado, and M. O'Hara, 2012d, "The Volume Clock: Insights into the High Frequency Paradigm." *Journal of Portfolio Management*, 39 (1), pp. 19-29.

Easley, D., M. López de Prado, and M. O'Hara, 2013, "Optimal Execution Horizon," *Mathematical Finance*; forthcoming.

Ellis, K., R. Michaely, and M. O'Hara, 2000, "The Accuracy of Trade Classification Rules: Evidence from NASDAQ," *Journal of Financial and Quantitative Analysis*, 35, pp. 529-551.

Epps, T.W., and M.L. Epps, 1976, "The Stochastic Dependence of Security Price Changes and Transactions Volumes: Implications for the Mixture of Distributions Hypothesis," *Econometrica*, 44, pp. 305-321.

Finucane, T.J., 2000, "A Direct Test of Methods for Inferring Trade Direction from Intraday Data," *Journal of Financial and Quantitative Analysis*, 36, pp. 553-576.

Kirilenko, A., A.S. Kyle, M. Samadi and T. Tuzun, 2011,"The Flash Crash: The Impact of High Frequency Trading on an Electronic Market," Working Paper, SSRN, May 2011.

Lee, C. and M. Ready, 1991,"Inferring Trade Direction from Intraday Data," *Journal of Finance*, 46, pp. 733-746.

MacIntosh, J.G., 2013. "High Frequency Traders: Angels or Devils?" Commentary 391, C.D. Howe Institute; Toronto, Ontario, Canada; October 2013.

Menkveld, A.J., Yueshen, B.Z., 2013. "Anatomy of the Flash Crash," Working Paper. SSRN, April 2013.

Moos, S., Pöppe, T. and D. Schiereck, 2014. "The Sensitivity of VPIN to the Choice of Trade Classification Algorithm," Working Paper. Technische Universität Darmstadt; March 2014.

Odders-White, E. 2000, "On the Occurrence and Consequences of Inaccurate Trade Classification," *Journal of Financial Markets*, 3, pp. 259-286.

Tauchen, G.E., and M. Pitts, 1983, "The Price Variability-Volume Relationship on Speculative Markets," *Econometrica*, 51, pp. 485-505.

Yildiz, S., R.A. Van Ness, and B.F. Van Ness, 2013, "Analysis Determinants of VPIN, HFTs' Order Flow Toxicity and Impact on Stock Price Variance," Working Paper, University of Mississippi; September 2013.

Wei, W. C., Gerace, D., and Frino, A., 2013. "Informed Trading, Flow Toxicity and the Impact on Intraday Trading Factors," *Australian Accounting Business and Finance Journal* 7, pp. 3–24.

Wu, K., Bethel, W., Gu, M., Leinweber, D. and Ruebel, O., 2013, A Big Data Approach to Analyzing Market Volatility," June 5, 2013, SSRN: http://ssrn.com/abstract=2274991.

Web Appendix

A TIC^T-VPIN Algorithm

This appendix reproduces algorithm for VPIN computation based on the Tick-Rule trade classification (TIC^T in our classification) as implemented in ELO (2011c) and used in ELO (2011a, 2011b).

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.

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.

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, $i = 1, \ldots, I$.
4. Initiate counter: $\tau = 0$.
5. Add one unit to τ.
6. If $I < \tau V$, jump to step 10 (there are insufficient observations).
7. $\forall i \in [(\tau - 1)V + 1, \tau V]$, classify each transaction as *buy* or *sell initiated*:
    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 V_τ^B the number of observations classified as *buy* in step 7, and the variable V_τ^S the number of observations classified as *sell*. Note that $V = V_\tau^B + V_\tau^S$.
9. Loop to step 6.
10. Set $L = \tau - 1$ (last bucket is always incomplete or empty, thus it will not be used).

1.3. APPLY VPINs FORMULA

If $L \geq n$, there is enough information to compute $VPIN_L = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{\sum_{\tau=L-n+1}^{L} (V_\tau^S + V_\tau^S)} = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{nV}$.

B BKU^T-VPIN Algorithm

This appendix reproduces algorithm for VPIN computation based on the Bulk-Volume Classification (BVC) (BKU^T in our classification) as implemented in ELO (2012c, on-line Appendix) and ELO (2012b).

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* flow toxicity 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 is that we apply a probabilistic approach to classify the volume exchanged within each 1-minute time bars. We cannot expect users to have data which unequivocally 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 aggregates of an alternative size (based on either time, number of trades or volume bars). Different schemes will lead to different levels of VPIN. We have used a variety of schemes and conclude that data aggregation leads to better flow toxicity estimates than working on raw transaction data.[34] Data granularity has an impact on the magnitude of VPIN levels, 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 1-minute time bars as this data is less noisy, more widely available and easier to work with.

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.

1.2. PREPARE EQUAL VOLUME BUCKETS

1. Sort transactions by time ascending: $T_{i+1} \geq T_i$, $\forall i$.

[34] For an in-depth analysis, please refer to ELO (2012b).

2. Compute ΔP_i, $\forall i$.
3. Expand the number of observations by repeating each observation ΔP_i as many times as V_i. This generates a total of $I = \sum_i V_i$ observations ΔP_i.
4. Re-index ΔP_i observations, $i = 1, \ldots, I$.
5. Initiate counter: $\tau = 0$.
6. Add one unit to τ.
7. If $I < \tau V$, jump to step 11 (there are insufficient observations).
8. $\forall i \in [(\tau - 1)V + 1, \tau V]$, split volume between *buy* or *sell initiated*:
   a. $V_\tau^B = \sum_{i=(\tau-1)V+1}^{\tau V} Z\left(\frac{\Delta P_i}{\sigma_{\Delta P}}\right)$
   b. $V_\tau^S = \sum_{i=(\tau-1)V+1}^{\tau V} \left[1 - Z\left(\frac{\Delta P_i}{\sigma_{\Delta P}}\right)\right] = V - V_\tau^B$
9. Assign to variable V_τ^B the number of observations classified as *buy* in step 8, and the variable V_τ^S the number of observations classified as *sell*. Note that $V = V_\tau^B + V_\tau^S$.
10. Loop to step 6.
11. Set $L = \tau - 1$ (last bucket is always incomplete or empty, thus it will not be used).

1.3. APPLY VPINs FORMULA

If $L \geq n$, there is enough information to compute $VPIN_L = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{\sum_{\tau=L-n+1}^{L} (V_\tau^S + V_\tau^S)} = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{nV}$.

C Tick-Rule and Bulk-Volume Classification

This appendix reproduces algorithms for Tick-Rule and Bulk-Volume Classification (BVC) as implemented in ELO (2012b).

1. TICK-RULE IMPLEMENTATION

Here we present a simple implementation of the Tick Rule in Python language. More efficient implementations exist, but we believe the one outlined below is the clearest.

```
queryCurs is assumed to contain the output of a SQL query such as

queryCurs.execute('SELECT Price, Volume, VolBuy FROM ' + tablename + '
ORDER BY Instrument, Time')
```

VolBuy is the field that stores the Volume from traders initiated by an aggressive buyer, as reported by the Exchange. The tick list variable will accumulate the amount matched over the entire volume. The rest of the code is self-explanatory.

```
a =queryCurs.fetchone()
flag, price, tick=1, a[0], [0,0]
while True:
   try:
      a=queryCurs.fetchone()
      # tick rule
      if a[0]>price:
         flag=1
      elif a[0]<price:
         flag=2
```

```
      if flag==1:
         tick[0]+=a[2] #correctly classified as buy
      else:
         tick[0]+=a[1]-a[2] #correctly classified as sell
      tick[1]+=a[1] #volume to be classified
      # reset price
      price=a[0]
   except:
      break
```

2. BULK VOLUME CLASSIFICATION IMPLEMENTATION

An equivalent codification of the BVC algorithm would be as follows. stDev is a real variable storing the volume weighted Standard Deviation of price changes across bars. The amount matched over the entire volume is stored in the list variable bulk.

```
a =queryCurs.fetchone()
price, bulk=a[0], [0,0]
while True:
   try:
      a=queryCurs.fetchone()
      # bulk classification
      z=float(a[0]-price)/stDev
      z=scipy.stats.norm.cdf(z)
      bulk[0]+=min(a[1]*z,a[2]) #correctly classified as buy
      bulk[0]+=min(a[1]*(1-z),a[1]-a[2]) #correctly classified as sell
      bulk[1]+=a[1] #volume to be classified
      # reset price
      price=a[0]
   except:
      break
```