AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Reflecting on the VPIN Dispute

Torben G. Andersen and Oleg Bondarenko

CREATES Research Paper 2013-42

Department of Economics and Business
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Reflecting on the VPIN Dispute*

Torben G. Andersen† and Oleg Bondarenko‡

Abstract

In Andersen and Bondarenko (2014), using tick data for S&P 500 futures, we establish that the VPIN metric of Easley, López de Prado, and O'Hara (ELO), by construction, will be correlated with trading volume and return volatility (innovations). Whether VPIN is more strongly correlated with volume or volatility depends on the exact implementation. Hence, it is crucial for the interpretation of VPIN as a harbinger of market turbulence or as a predictor of short-term volatility to control for current volume and volatility. Doing so, we find no evidence of incremental predictive power of VPIN for future volatility. Likewise, VPIN does not attain unusual extremes prior to the flash crash. Moreover, the properties of VPIN are strongly dependent on the underlying trade classification. In particular, using more standard classification techniques, VPIN behaves in the exact opposite manner of what is portrayed in ELO (2011a, 2012a). At a minimum, ELO should rationalize this systematic reversal as the classification becomes more closely aligned with individual transactions.

ELO (2014) dispute our findings. This note reviews the econometric methodology and the market microstructure arguments behind our conclusions and responds to a number of inaccurate assertions. In addition, we summarize fresh empirical evidence that corroborates the hypothesis that VPIN is largely driven, and significantly distorted, by the volume and volatility innovations. Furthermore, we note there is compelling new evidence that transaction-based classification schemes are more accurate than the bulk volume strategies advocated by ELO for constructing VPIN. In fact, using perfect classification leads to diametrically opposite results relative to ELO (2011a, 2012a).

JEL Classification: G01; G14; and G17

Keywords: VPIN; PIN; High-Frequency Trading; Order Flow Toxicity; Order Imbalance; Flash Crash; VIX; Volatility Forecasting

*We are indebted to the Zell Center for Risk at the Kellogg School of Management, Northwestern University, for financial support. We are grateful to the journal editors for the opportunity to respond to the ELO rejoinder and to provide a few additional points for debate.

†Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208; NBER, and CREATES; Phone: (847) 467-1285; e-mail: t-andersen@kellogg.northwestern.edu

‡Department of Finance (MC 168), University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607; Phone: (312) 996-2362; e-mail: olegb@uic.edu

1. Introduction

Easley, Lopez de Prado and O'Hara (ELO) (2014) begin their rejoinder by asserting: " . . . AB attack a methodology we do not advocate, an analysis we never performed, and conclusions we did not draw." In short order, they state we (exclusively) comment on the (tick rule) TR-VPIN procedure. These claims are absurd and ironic. They know – very well – that our publication was delayed so that we could address also a more recent implementation strategy for VPIN. In fact, we agreed to have the initially accepted version undergo a technical retraction to allow for addition of such material. Moreover, with assistance from the journal editors, we obtained clarification on implementation details to ensure that we replicate the procedures of ELO (2011a, 2012a) as closely as possible. Thus, the published article, Andersen and Bondarenko (henceforth, AB) (2014), includes a substantial section dealing with the (bulk volume) BV-VPIN metric introduced in ELO (2012a).

We reiterate our primary concern with ELO's evidence. It focuses almost exclusively on showing that VPIN – in specific scenarios – forecast return volatility. In isolation, this exercise is vacuous. Their VPIN metric is, by construction, highly correlated with recent innovations to trading volume and return volatility. Since volume and volatility are highly correlated and display strong time series persistence, *any* variable correlated with volatility will, inevitably, possess non-trivial forecast power for future volatility. This is true for bid-ask spreads, the quote intensity, the transaction count, the (normalized) trading volume – as well as ELO's VPIN metric. This merely confirms that volatility begets volatility. To establish economic content, one must show both that the VPIN metric provides a *good indicator of underlying order imbalances* and that it has *incremental predictive power for volatility*.

ELO do not concur. They characterize the use of benchmarks or controls as irrelevant or illegitimate (a methodology they did not advocate, an analysis they did not perform). This represents a crucial disagreement. For illustration, consider the BV-VPIN procedure. Since realized volatility is highly persistent, current absolute price changes forecast future absolute price changes. ELO (2012a) impute order imbalances through a monotone function of absolute price changes and then use the associated BV-VPIN measure to forecast volatility, effectively letting absolute price changes forecast absolute price changes. The correlation is – obviously – significant but this has no bearing on any causal relationship. On the first page of his textbook, Jeff Wooldridge discusses proper procedures for establishing a causal relationship between a variable w and an outcome y. He notes that correlation between y and w does not prove that a change in w induces a change in y. We need econometric techniques to ensure we hold other relevant (control) variables fixed when measuring the (*ceteris paribus*) correlation between w and y:

> The reason we control for these variables is that we think w is correlated with other factors that also influence y. (Wooldridge (2002), page 3.)

In other words, we must control for, e.g., current realized volatility when assessing the relation between VPIN and future volatility. This observation should not be controversial but it represents the crux of our disagreement with ELO.

We also note that VPIN is a moving target. ELO (2011a) relies on TR-VPIN with time bars. ELO (2012a) adopts a time bar BV-VPIN procedure using a normal CDF transformation of absolute price changes. The latest version of ELO (2012b) explores BV-VPIN employing a Student-t CDF transformation instead. In the rejoinder, ELO stress the evidence of Wu, Bethel, Gu, Leinweber, and Ruebel (WBGLR) (2013) who explore only volume bar BV-VPIN. Nonetheless, they imply that all major conclusions are robust. We address these claims in a comprehensive study, AB (2013), concluding that all critical points identified in AB (2014) carry over to that context. For the remainder of this response, we address the issue at hand, namely the VPIN evidence in ELO (2011a, 2012a), although we refer to AB (2013) for illustration of some broader points.

2. The Main Points of Dispute – An Initial Response

ELO make a number of specific assertions. For one, they insist VPIN reached an extreme level prior to the flash crash and thus provided a strong signal of the potential for impending turmoil. They argue this is best seen from the value attained by the (empirical) CDF of VPIN. At different stages, they mentioned thresholds ranging from 90% to 99% as an indicator of extreme value. However, even a threshold of 99% is too low to indicate true extreme values in their setting. ELO generate fifty VPIN values per day on average. Thus, across large samples, we expect such values to be observed at a frequency corresponding to once every two days. On the other hand, the VPIN values are highly correlated so the average frequency is also a highly problematic metric for extreme realizations. An inherently more meaningful measure for the occurrence of outliers of a given size is the CDF of the daily maximum VPIN value. This provides an invariant (daily) horizon for comparisons. Using this transparent measure, we document an elevated, but not unusual, VPIN value prior to the flash crash. In fact, it is trivial to obtain more striking indicators of extreme volatility than VPIN before the crash using directly observable market indicators. In short, VPIN provides neither a unique nor reliable crash indicator.

A second point stressed by ELO is the theoretical foundation for the VPIN metric. However, the VPIN implementation is largely bereft of any connection to theory. For example, the properties of VPIN vary dramatically with the type and length of the time or volume bar. ELO provide no systematic answer to these observations, except to suggest that trade classification at high frequencies is noisy. In contrast, we provide a broad rationalization of these patterns, emphasizing typical regularities associated with the volatility dynamics and the effect of microstructure features beyond those invoked by ELO.

A related point centers on the precision of the trade classification. ELO assert that the BV procedure is more suitable than the regular tick rule applied to individual transactions. In reality – using accurate buy-sell indicators as the metric for performance – the standard tick rule uniformly dominates the BV classification. This point is documented carefully in Chakrabarty, Pascual and Shkilko (CPS) (2012) and AB (2013). Moreover, CPS find the misclassification of BV to be correlated with overall market activity. Likewise, AB (2013) find that the BV errors are strongly correlated with the volatility level. Thus, the VPIN metric becomes increasingly distorted as the absolute price changes grow larger. This generates spikes in the toxicity measure which bear little resemblance to the actual order imbalances in the market. Instead, large absolute price changes (volatility) begets a high VPIN reading as well as high future volatility – thus explaining the correlation between VPIN and future short-term volatility.

A fourth issue is the findings of related independent research. The citations of ELO (2014) provide not a single reference of relevance for the pertinent issues (CPS is not cited). ELO emphasize the studies by authors affiliated with the Lawrence Berkeley National Laboratories. Marcos López de Prado is a Research Affiliate of this institution. One of his co-authors participates in the execution of both studies. Leaving aside the question of "independence," the studies provide a useful service by undertaking a large-scale implementation of VPIN along the lines outlined by ELO. They confirm that high VPIN readings are correlated with future volatility which, mechanically, they must be. Unfortunately, there is no investigation of the underlying source of correlation. Hence, they are silent on the causal issues we address. Similarly, none of the remaining papers investigate the relation between VPIN and actual order imbalances or compare the ELO VPIN forecast performance to benchmark predictors. Instead, the list of references underscores the attention VPIN has garnered, reminding us yet again why it is critical to seek additional clarity regarding the origin of the VPIN-volatility correlation.

3. The Disputed Points – Detailed Responses

3.1 Did VPIN Hit a Historical High Prior to the Flash Crash?

VPIN was developed in the wake of the flash crash as a tool for gauging the likelihood of a liquidity induced market malfunction. Thus, it is of interest to establish whether the metric would have signaled a truly extreme event prior to the crash on May 6, 2010, if it had been operational at the time. In some sense, this is a minimal requirement: no such metric would be put forth without an in-sample back-test to validate that it would have served as an effective warning indicator prior to the flash crash. Of course, not many such measures have been proposed, suggesting this may be a difficult task. The (TR-)VPIN metric certainly gained instant recognition upon its appearance:

> "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," said Lopez de Prado, head of high-frequency futures at Tudor. [1]

> "All morning long on May 6 order flows were becoming increasingly unbalanced, and volumes were huge," said O'Hara. "An hour or more before the flash crash our measure hit historic levels." [2]

The basic point is reiterated in the original ELO (2011a) publication (using Eastern Time):

> By 11:55 a.m. on May 6 the realized value of the VPIN metric was in the 10% tail of the distribution (it exceeded a 90% CDF(VPIN) critical value). By 1:08 p.m., the realized value of the VPIN metric was in the 5% tail of the distribution (over a 95% CDF(VPIN)). By 2:30 p.m., the VPIN metric reached its highest level in the history of the E-mini S&P 500. At 2:32 p.m., the crash began, according to the CFTC-SEC Report time line.

These early citations (necessarily) refer to time bar TR-VPIN. In more recent accounts, it is often unclear what notion of VPIN is invoked, but the assertions continue to be definitive:

> One hour before the *flash crash*, order flow toxicity was the highest in recent history. [3]

ELO (2014) now back off, however slightly, from this claim, but still argue:

> This is also why AB's comment that VPIN was higher at isolated points in the past is not relevant here. We showed that VPINs has sustained levels above their 99% level for more than an hour before the crash.

This characterization is inaccurate. Our VPIN measures indicate that, on May 6, 2010, TR-VPIN (BV-VPIN) reached a CDF level of 95.3% (78.7%) at 12:30. This value was surpassed on 71 (189) preceding days, constituting 11.7% (31.2%) of the pre-crash sample. Likewise, only by 13:30 – the onset of the crash – had TR-VPIN (BV-VPIN) exceeded the 95.3% (78.7%) threshold for one full hour. There were 39 (131) preceding days, representing 6.4% (21.7%) of the pre-crash sample, where this value was surpassed for at least one hour. In comparison, the TR-VPIN (BV-VPIN) CDF level at 13:30 was 98.5% (94.7%), a value topped on 26 (49) preceding days, or 4.3% (8.1%) of the pre-crash sample. If anything, the one hour condition *weakens* the evidence that VPIN reached an extreme pre-crash level.

[1] "'Toxic' Orders Can Predict Likelihood of Stock Market Crashes, Study Says," Bloomberg, Oct 29, 2010.

[2] "'Flash' Crashes now Predictable, Thanks to Cornell-Developed Metric," Cornell University Chronicle, Dec 1, 2010.

[3] Wikipedia: http://en.wikipedia.org/wiki/2010_Flash_Crash, July 25, 2013.

In summary, our finding that VPIN was far removed from the historical high is robust. This signifies a jarring inconsistency, contradicting the oft-repeated claim – highlighted by main-stream media, ELO (2011a), and Wikipedia – that VPIN would have signaled extreme stress prior to the crash.

3.2 What Constitutes an Extreme VPIN (CDF) Value?

ELO use the empirical CDF of the VPIN metric to indicate extreme realizations. However, the interpretation of the CDF VPIN value in the current context is both subtle and deceptive. For example, we may have a sample of VPIN values covering 500 trading days (about two years) and employ a CDF threshold of 99%. If there is only one VPIN value recorded per day, then only five days contain a recorded VPIN value exceeding the threshold. But if there are an average of 50 VPIN observations per day – as is the case with ELO VPIN – there are 250 VPIN values exceeding the threshold in the 500-day sample. The interpretation of this number hinges critically on the distribution of the extreme VPIN readings across separate days in the sample. They could be highly clustered and all be recorded within a span of five days. On the other hand, they could be dispersed and be observed across, say, 100 different days. In the former case, the 99% threshold is breached only on a handful of days per year, while in the latter it is surpassed, on average, once a week. Thus, is the threshold violated once every five months or every week? The 99% CDF level is silent on this critical issue.

These observations motivate our emphasis on the number, or percentage, of days for which the threshold is breached. For each trading day, we record the highest VPIN reading and then compute the CDF of those daily maximum values. To illustrate how this statistic empirically is related to the ELO CDF values for our E-mini S&P 500 futures sample, the following table converts the CDF(VPIN) into the corresponding daily CDF(MAX VPIN) for the BV-VPIN metric.

CDF of VPIN	90.0	95.0	99.0
CDF of daily max VPIN	83.8	91.6	96.9

The table implies that the 99% BV-VPIN threshold is attained on 3.1% of the days, or about 8 times a year. In our computation, BV-VPIN was around the 95% CDF mark at the onset of the crash, implying that a corresponding VPIN level was hit on more than 8% of the days, or about once every two weeks.[4]

Finally, ELO compare the CDF of VPIN to that for VIX, while acknowledging that VIX is a proxy for monthly volatility expectations. For an evolving liquidity event, one may anticipate a severe, but temporary, market disruption. This should manifest itself in a sharp increase in VIX relative to its recent past, reflecting the near-term potential for turmoil, while also allowing for a reversal over the remainder of the month-long window. In short, the *relative increase* in the VIX is the variable of interest. Figure 1 displays the empirical CDF of the VPIN metrics from ELO (2011a, 2012a) and a normalized VIX measure, defined as the current VIX value divided by its past 21-day (monthly) moving average, on the day of the flash crash. The normalized VIX measure is elevated throughout and generates high CDF values much earlier than VPIN (note the discrepancy in scale). For example, at 13:00, the BV-VPIN CDF is about 80% while the normalized VIX CDF is close to 99%. By no means does this imply that normalized VIX is a toxicity proxy, but it does exemplify the need for proper benchmarks in gauging whether VPIN is a unique crash indicator. The claim that VPIN is superior in signaling increased toxicity/volatility is simply not borne out by the data. Note also that VPIN is calibrated over a host

[4] In ELO (2011b), they exemplify regulatory use of VPIN via a rule that orders a temporary market halt when VPIN exceeds the 90% CDF threshold – a value that BV-VPIN would exceed more than once per week in our sample.

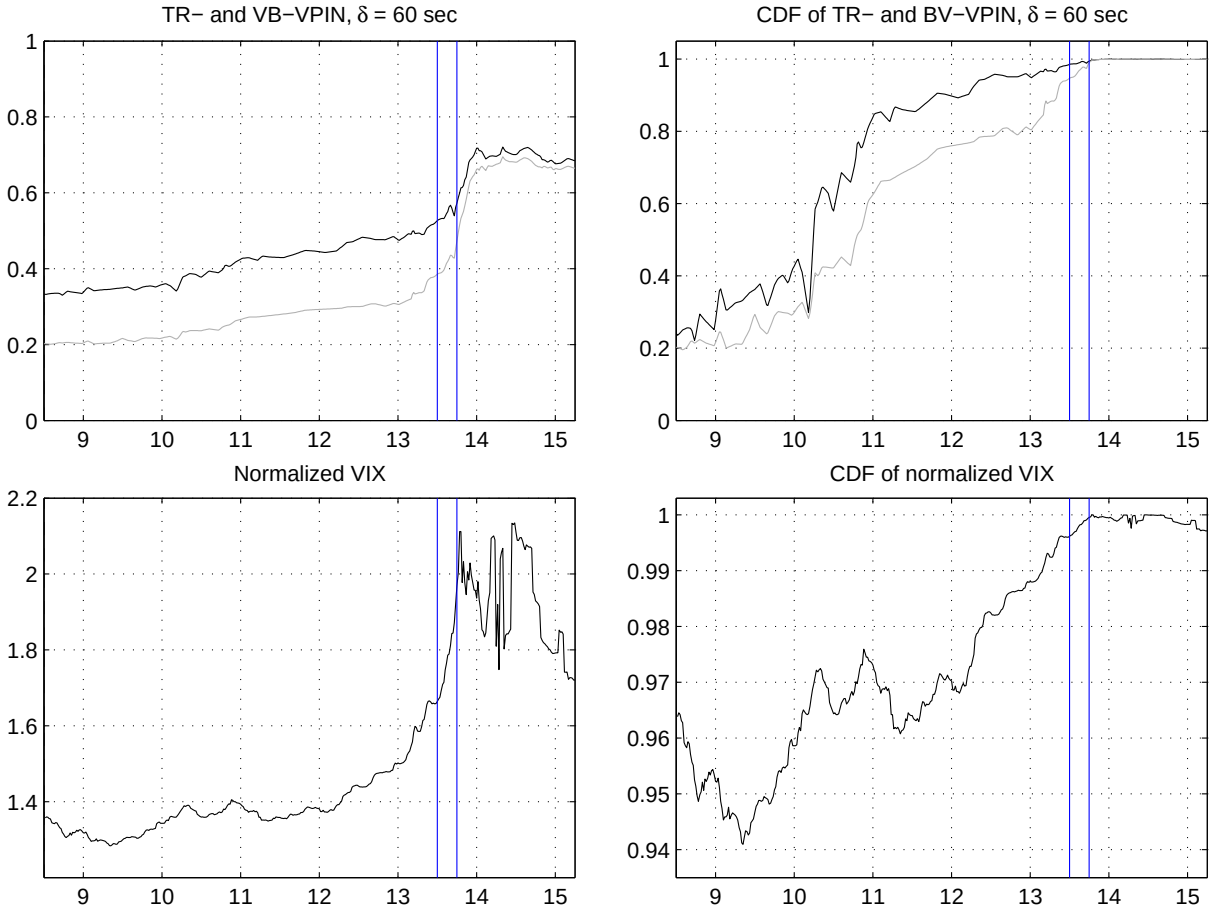

Figure 1: **The evolution of two VPIN metrics and normalized VIX for May 6, 2010.** The top panels show TR-VPIN (black) and VB-VPIN (gray). The bottom panels show the normalized VIX, which equals VIX divided by its 21-day moving average. The vertical lines indicate the timing of the flash crash.

of design variables. We did not experiment with alternative moving average values for VIX, so our evidence merely reflects one sensible, but non-optimized, choice for the lag window.

3.3 Is the ELO VPIN Metric Theory Driven?

ELO (2011a, 2012a) rely on the static PIN model in motivating VPIN. The driving force of that model is a set of exogenous arrival rates for good/bad news and uninformed/informed traders. In taking the theory to the high-frequency setting, the model is unaltered, but the data are sampled using a volume-based scheme. The key insight is that the probability of informed trading, or order flow toxicity, may be derived from the expected order imbalance. From this point, however, the theory ceases to provide much guidance. There are a number of design variables involved in computing VPIN that are *ad hoc* but crucial for the empirical properties of the measure. In order to be credible, the empirical findings should be robust across a variety of designs. As amply documented in AB (2013a, 2013b), we find the metric failing on this criterion, and we only point to a few examples here.

First, the theory does not advocate mixing business time (volume buckets) and calendar time (time

bars). ELO stress the importance of volume-based sampling, but ELO (2011a, 2012a) use time bars for the main empirical illustrations. We document that time bar TR-VPIN is severely distorted by variation in the trading volume. This is established via the benchmark variables, U1-VPIN and U2-VPIN. ELO fundamentally misinterpret the role of these metrics. They are not candidate toxicity measures, but controls that help identify the incremental impact of the order imbalance measure – relative to the pure volume effect – on TR-VPIN.[5] We also document that moving from time to volume bars reverses the findings of ELO (2011a). We provide a microstructure based explanation of this fact, while ELO – though never directly refuting our findings – insist their results remain valid for TR-VPIN computed via volume bars. For BV-VPIN, we explain how volume bars allow the metric to approximate (tick-sampled) realized volatility, while time bars introduce auxiliary correlation with trading volume. As such, the metric with best predictive power for future short-term volatility is volume bar BV-VPIN – but this is achieved by mimicking realized volatility, not by inferring order flow imbalances from transactions data.

Second, the theory is silent on the length of the (time or volume) bars. AB (2014) show that the ELO results gradually shift towards the diametrically opposite as the bar size shrinks. ELO deem such evidence proof that trade classification is worsening. On the contrary, the classification improves with smaller bars. In fact, transaction-based classification is, by far, superior to the BV procedure or the tick rule applied to aggregated transactions, as explained in the following section. We provide intuitive arguments in favor of this point in AB (2014) and document it formally in AB (2013), while ELO argue purely heuristically, using examples to illustrate that individual trades may be misclassified by a tick rule. However, in the aggregate only systematic biases matter – idiosyncratic errors diversify effectively across the large number of trades for transaction-based procedures.

Third, in the rejoinder, ELO seem to advocate selecting the design variables based on in-sample optimization of the forecast power of extreme VPIN realizations for short-term volatility, as in WBGLR. This is the ultimate theory-less approach, searching over thousands of designs for a constellation that maximizes the given objective. Inevitably, it selects a design for which VPIN correlates strongly with the best-known predictor for future volatility, namely concurrent realized volatility. Moreover, note that, as the design variables are varied during optimization, the imputed order flow changes as well, but there is no constraint that the latter is aligned with actual trade imbalances. Consequently, this risks substituting a pseudo-realized volatility measure in lieu of an order imbalance indicator, rendering it even more critical to use realized volatility as a control. Upon doing so, AB (2013) find the predictive power of BV-VPIN to be annihilated – it is fully subsumed by realized volatility. Rather than being guided by theory, the VPIN implementation is largely unconstrained, allowing the measure to morph into a pure activity measure with only weak linkages to the underlying order flow.

3.4 Is Bulk Volume Trade Classification Accurate?

Chakrabarty, Pascual and Shkilko (CPS) (2012) undertake a comprehensive study of two classification methods, the standard transaction tick rule, which we label TTR, and bulk volume (BV) classification, using a sample of hundreds of stocks traded on NASDAQ's INET platform. They find TTR to outperform BV in all tests – for small and large stocks, for time and volume bars of all sizes, for the precision

[5] ELO claim Table 6 of AB (2014) implies superiority of BV-VPIN. This comment reflects a fundamental misunderstanding of our empirical strategy. The table displays univariate regressions. The predictive power of BV-VPIN falls well short relative to VIX. While it does improve upon U1-VPIN, the latter is a mere benchmark, reflecting a VPIN metric that fails to exploit the information regarding either trade classification or volume heterogeneity across bars. A minimum requirement for VPIN is to improve significantly on this "uninformed" benchmark. As such, the more striking feature is that TR-VPIN falls short of this uninformed benchmark, rendering it a truly inferior predictor, albeit the one that propelled VPIN into the media limelight. As discussed below, the information content of BV-VPIN is fully subsumed by realized volatility.

of order imbalance measures, and for the accuracy in computing VPIN. In particular, the best BV specification misclassifies 20.3% of the trades, and TTR only 9.2%. CPS also document that the BV accuracy is adversely impacted by episodes of elevated market activity, as captured by high volatility, trading frequency, and hidden volume. Hence, overall, CPS find BV to be less accurate and less stable than TTR in estimating order imbalances. This matters for performance. For TTR-VPIN metrics, they find 91% to 93% correct identification of toxic events, whereas BV-VPIN identifies only 64% to 70% of these events. They also conclude that the TTR estimates have significantly lower dispersion and lower Type-II errors of over-identifying toxic events than BV-VPIN.

In a separate robustness check, CPS establish that there is a striking stability in their findings between the years 2005 and 2011. This is significant as ELO hypothesize a substantial decline in the precision of TTR in the world of high-speed trading. There is no evidence for this effect in the data.

Similarly, using a sample covering more than five years for S&P 500 futures, AB (2013) document that TTR outperforms every BV scheme by a substantial margin in terms of classification accuracy. In particular, at the volume bucket level, TTR misclassifies 2.3% of trades, while the one-minute time bar BV – favored by ELO (2012a) – misclassifies 8.3%. Even more importantly, the BV errors are highly correlated with the volatility level, thus inflating the misclassification rate when markets grow turbulent. Hence, the induced toxicity measure is severely upward biased when markets are volatile – the most critical periods for the ELO short-term volatility forecast analysis.

AB (2013) also confirm that TTR induces a time series behavior in the associated VPIN that closely reflects the features of the VPIN metric generated from true trade classification. That is, TTR-VPIN series is suitable for analyzing the properties of an ideal VPIN metric. The problem for VPIN is that, relying on transaction data, all major findings of ELO are overturned. ELO raise concerns regarding whether true classification of buys and sells is the objective, as some informed trades can be passive. However, there is no alternative standard by which to assess classification accuracy – even ELO (2012b) apply the true buy-sell indicator as the criterion for success. In that study, ELO also find TTR to outperform BV classification for the E-mini S&P 500 futures, although their conclusions are less definitive.[6] Finally, AB (2013) document that the cumulative order imbalance is highly correlated with the actual price movements, implying that the (true) order imbalances convey information about directional price moves. If VPIN constructed from such series contradict the theoretical predictions, there is a fundamental issue in need of rationalization. We provide an extensive analysis, concluding that the problem stems from the construction of the VPIN metric itself. ELO has yet to address the issue in a systematic manner.

3.5 What Do We Learn from Independent VPIN Research?

ELO invoke two studies conducted at the Lawrence Berkeley National Laboratories as independent evidence that VPIN provides an early warning signal of market turbulence. Unfortunately, these studies follow the identical path of merely establishing the correlation of extreme realizations for optimized-across-design VPIN with future volatility. All parties agree this correlation should be significant (monotone transformations of absolute price changes *do* forecast short-term volatility). The issue is to explore robustness with respect to suitable benchmarks or include control variables to establish that the metric captures unique toxicity-related features of volatility, and to document how this manifests itself in a differential ability to forecast market disruptions compared to other predictor variables. Such an analysis is never undertaken, thus missing the opportunity to provide independent evidence along the

[6]ELO do not explore the accuracy of the procedure implemented in ELO (2012a), as they only adopt a Student-t CDF transformation of absolute price changes in this study. Hence, only CPS and AB (2013) directly address the precision of the approach in ELO (2012a).

more important dimensions. As an aside, we note WBGLR focus exclusively on volume bar BV-VPIN. This metric – by construction – will tend to correlate highly with tick-sampled realized volatility, as explained in AB (2013a, 2013b). Thus, the findings are as expected, but the study fails to invoke controls or benchmarks to help gauge the incremental predictive power. In short, this is a gigantic missed opportunity. They could address all major points raised above, but fail to do so.

The discussion of WBGLR offers a new opportunity to illustrate the critical role of benchmarks. They search over 16,000 design combinations, concluding:

> With appropriate parameter choices, the false positive rates are about 7% averaged over all the futures contracts in the test data set. More specifically, when VPIN values rise above a threshold (CDF > 0.99), the volatility in the subsequent time windows is higher than the average in 93% of the cases.

Is it truly remarkable? Regrettably, they apply a criterion void of a benchmark. We assert that the forecast power of volume bar BV-VPIN is tied to the (indirectly optimized) correlation with realized volatility. We conjecture that a CDF threshold for realized volatility will provide a superior signal of above-average volatility going forward. To test this hypothesis, we compute false positive rates for a similar experiment, but using current realized volatility instead of VPIN to predict the S&P 500 volatility, defined as the square-root of daily Realized Variance (RV). We adopt the daily horizon to neutralize the pronounced intraday pattern (WBGLR also include the one-day window in their analysis). We do not optimize any parameters and compute volatility in the most traditional way, see, e.g., Andersen and Bollerslev (1998). The table below summarizes our findings.

Volatility percentile	70	80	90	95	99
False positive rate	10.8	3.0	1.2	0.0	0.0

When volatility rises above the 95% threshold of its empirical CDF, subsequent volatility is higher than the sample average 100% of the time. Thus, there is not a single false positive at the 95% level versus the 7% error rate for VPIN at the 99% level! They may be confronting a more challenging set of assets but, evidently, the criterion is meaningless without a proper benchmark.

We now turn to the remaining ELO references. Abad and Yague (2012) follow the exact ELO (2012a) BV-VPIN procedure and provide no independent analysis of relevance to the current discussion. Bohn (2011) contains no mention of VPIN at all! In his book, Corcoran (2013) merely summarizes aspects of the original ELO (2011a) time bar TR-VPIN metric – no independent analysis or implementation. Menkveld and Yueshen (2013) simply plot VPIN along with the volume share of the large seller of the E-mini S&P 500 futures for *30 minutes* covering the flash crash. It contains no independent exploration of VPIN. Finally, Wei, Gerace and Frino (2013) refer to the TR-VPIN concept from ELO (2011a), but pursue a dramatically different implementation, using trade classification from transaction data via the tick rule (the TTR procedure) and an extremely small bucket size. Clearly, their TTR-VPIN implementation is widely at odds with the recommendations of WBGLR and ELO. Moreover, it produces the opposite conclusions relative to ELO (2011a) if applied to the S&P 500 index. In short, none of these references has any bearing of the disputed issues.

In summary, the lone independent reference of relevance to the current discussion, that we are aware of, is CPS. That paper fails to make the ELO citation list.

4. Auxiliary Hypotheses Regarding the Properties of VPIN

One way to assess the validity of a specific hypothesis is to test auxiliary implications that are not directly associated with the empirical features that motivated the development of the underlying theory. Our rationalization of the properties of VPIN has a number of such implications. First, the level of time-bar VPIN should increase if volume is trending upward over time, thus falsely indicating a secular growth in average order flow toxicity. We confirm this conjecture in AB (2013).[7] Second, the pronounced intraday pattern in volatility and volume should induce a distinct intraday pattern in the order imbalance measures obtained via the bulk volume procedure, even if such a pattern is largely absent from the true order imbalances. Again, AB (2013) confirms the hypothesis. The BV-induced order imbalance measures indicate high imbalances at the open and close of the regular trading day, and much less around noon – exactly mimicking the intraday patterns of volume and volatility. In contrast, the true order flow imbalance is flat across the regular trading day. This is yet another indication of how strongly market activity variables impact ELO VPIN – in a manner quite unrelated to the true order flow imbalances. Finally, we reiterate that AB (2013) find – after controlling for the volume trend – that realized volatility fully subsumes the information content of BV-VPIN. These results add to the comprehensive set of broad empirical features that appear to contradict the interpretation of VPIN as a real-time order flow toxicity measure. Instead, it bears the hallmarks of a distorted market activity measure.

5. Postscript

Our analysis of VPIN was initiated in late 2010 when we encountered the original ELO papers, claiming that the VPIN metric, obtained from the transaction record, could forecast short-term volatility in turbulent times significantly better than alternative indicators such as the VIX. Most strikingly, VPIN was not directly using typical volatility indicators as predictors. Thus, the procedure might be exploiting insights we had failed to elicit in our own work seeking to forecast volatility via tick-by-tick data.

Upon careful inspection of the ELO procedures, we concluded, however, that the key to the forecast power of VPIN was the indirect inclusion of dynamic volume and volatility information during the implementation stage. We confirmed that VPIN was highly correlated with recent volume and volatility innovations for the E-mini S&P 500 futures. Hence, any spike in VPIN was associated with unusually large volume or volatility innovations over the preceding few hours. The issue was: *what causes what*? Even when modeling only the volatility-volume interactions, it is critical to avoid endogeneity biases arising from interpreting evidence based on univariate specifications as causal. At a minimum, such analysis must exploit control variables to account for simultaneity.[8] Upon inserting such controls for the volatility and volume effects, we found the predictive power of VPIN to drop precipitously and – for the more thoroughly designed tests – turn insignificant. That is, there appears to be no incremental predictive power in the VPIN metric beyond what is already incorporated in standard volatility forecast procedures.

Equally disturbing, we could not confirm that VPIN attained a historical high *before* the flash crash. This claim was a major factor in propelling the VPIN metric to the limelight. Throughout the following year, media, exchanges, regulators, industry associations, and academic conferences gave generous air-time to numerous presentations highlighting the metric as an indicator of order-flow toxicity and as a signal for an increasing probability of impending market turmoil.

[7] Hence, in that study, we normalize the volume buckets by a measure reflecting the daily volume experienced in the recent past in order to eliminate this mechanical inflation of the VPIN metric.

[8] A more structural approach is to express the dynamics through a set of simultaneous equations, see, e.g., Andersen (1996).

Given the attention and resources devoted to the VPIN metric, we found it important to make our findings available to the broader public. This turned out to be an unexpectedly lengthy, complex and, frankly, unpleasant undertaking. Our only intention was – and remains – to foster awareness and discussion about the underlying forces that drive the VPIN metric. From our analysis, we cannot identify a single feature that renders VPIN a genuine crisis indicator or short term volatility predictor. We understand that others may reach a different conclusion. Since the issues are quite basic, albeit hard to investigate without a considerable allocation of resources and time, we are making our VPIN series for the E-mini S&P 500 futures available to interested scholars upon request.[9] Our hope is that this will remove the main obstacles in establishing consensus regarding the more critical empirical facts under dispute. Did VPIN reach a historical high prior to the flash crash? Does VPIN systematically provide incremental forecast power for short term volatility in specific scenarios? Is the bulk volume procedure more suitable for trade classification than the regular tick-rule applied to transaction data?

References

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

Andersen, T.G., 1996. Return volatility and trading volume: an information flow interpretation of stochastic volatility. Journal of Finance 51, 169–204.

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

Andersen, T.G., Bondarenko, O., 2013. Assessing VPIN measurement of order flow toxicity via perfect trade classification. Working Paper; Kellogg School of Management, Northwestern University, and College of Business Administration, University of Illinois at Chicago.

Andersen, T.G., Bondarenko, O., 2014. VPIN and the flash crash. Journal of Financial Markets.

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, Spring.

Bohn, S., 2011. The Slippage Paradox. Working Paper; LPMA, Universite Denis Diderot (Paris 7) and CNRS, France.

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

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

[9] We have already contacted the authors behind the VPIN replication studies at the Lawrence Berkeley National Laboratories with an offer to exchange our series. We have not yet received a response.

Easley, D., López de Prado, M., O'Hara, M., 2011a. The microstructure of the "flash crash": flow toxicity, liquidity crashes, and the probability of informed trading. Journal of Portfolio Management 37 (2), 118–128.

Easley, D., López de Prado, M., O'Hara, M., 2011b. The exchange of flow toxicity. Journal of Trading 6 (2), 8–13.

Easley, D., López de Prado, M., O'Hara, M., 2012a. Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies 25, 1457–1493.

Easley, D., López de Prado, M., O'Hara, M., 2012b. Bulk classification of trading activity. SSRN, March 2012.

Easley, D., López de Prado, M., O'Hara, M., 2014. VPIN and the flash crash: A rejoinder. Journal of Financial Markets.

Menkveld, A.J., Yueshen, B.Z., 2013. Anatomy of the Flash Crash. Working Paper. SSRN, April 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, 3–24.

Wooldridge, J. M., 2002. Econometric Analysis of Cross-Section and Panel Data. MIT Press; Cambridge, MA, and London, England.

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

Research Papers 2013

2013-23: Asger Lunde and Anne Floor Brix: Estimating Stochastic Volatility Models using Prediction-based Estimating Functions

2013-24: Nima Nonejad: A Mixture Innovation Heterogeneous Autoregressive Model for Structural Breaks and Long Memory

2013-25: Nima Nonejad: Time-Consistency Problem and the Behavior of US Inflation from 1970 to 2008

2013-26: Nima Nonejad: Long Memory and Structural Breaks in Realized Volatility: An Irreversible Markov Switching Approach

2013-27: Nima Nonejad: Particle Markov Chain Monte Carlo Techniques of Unobserved Compdonent Time Series Models Using Ox

2013-28: Ulrich Hounyo, Sílvia Goncalves and Nour Meddahi: Bootstrapping pre-averaged realized volatility under market microstructure noise

2013-29: Jiti Gao, Shin Kanaya, Degui Li and Dag Tjøstheim: Uniform Consistency for Nonparametric Estimators in Null Recurrent Time Series

2013-30: Ulrich Hounyo: Bootstrapping realized volatility and realized beta under a local Gaussianity assumption

2013-31: Nektarios Aslanidis, Charlotte Christiansen and Christos S. Savva: Risk-Return Trade-Off for European Stock Markets

2013-32: Emilio Zanetti Chini: Generalizing smooth transition autoregressions

2013-33: Mark Podolskij and Nakahiro Yoshida: Edgeworth expansion for functionals of continuous diffusion processes

2013-34: Tommaso Proietti and Alessandra Luati: The Exponential Model for the Spectrum of a Time Series: Extensions and Applications

2013-35: Bent Jesper Christensen, Robinson Kruse and Philipp Sibbertsen: A unified framework for testing in the linear regression model under unknown order of fractional integration

2013-36: Niels S. Hansen and Asger Lunde: Analyzing Oil Futures with a Dynamic Nelson-Siegel Model

2013-37: Charlotte Christiansen: Classifying Returns as Extreme: European Stock and Bond Markets

2013-38: Christian Bender, Mikko S. Pakkanen and Hasanjan Sayit: Sticky continuous processes have consistent price systems

2013-39: Juan Carlos Parra-Alvarez: A comparison of numerical methods for the solution of continuous-time DSGE models

2013-40: Daniel Ventosa-Santaulària and Carlos Vladimir Rodríguez-Caballero: Polynomial Regressions and Nonsense Inference

2013-41: Diego Amaya, Peter Christoffersen, Kris Jacobs and Aurelio Vasquez: Does Realized Skewness Predict the Cross-Section of Equity Returns?

2013-42: Torben G. Andersen and Oleg Bondarenko: Reflecting on the VPIN Dispute