Journal of Banking and Finance 73 (2016) 165–181

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

The sensitivity of VPIN to the choice of trade classification algorithm

Thomas Pöppe[a,*], Sebastian Moos[b], Dirk Schiereck[a]

[a] Technische Universität Darmstadt, Hochschulstr. 1, 64289 Darmstadt, Germany
[b] Karlsruher Institut für Technologie, Kaiserstraße 12, 76131 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 25 July 2014
Accepted 22 August 2016
Available online 26 August 2016

JEL classifications:
G12
G14

Keywords:
Capital markets
Information asymmetry
Volume-synchronized probability of informed trading
VPIN
Trade classification

ABSTRACT

The VPIN metric (Easley et al. 2012b) aims to detect and predict the toxicity of order flow. This paper examines the sensitivity and robustness of VPIN to the choice of trade classification scheme, which is the major input used to compute VPIN. We compare deterministic trade-by-trade classification approaches with results computed using a newly proposed heuristic approach, bulk volume classification. We find substantial differences for all levels of aggregation: trade classification, order imbalance, VPIN and identifying "toxic periods". We also find that the detection of toxic periods does not yield consistent results in more than 60% of cases. But regression analysis can identify volume and return volatility as parameters that contribute to higher levels of sensitivity.

1. Introduction

In the trading of any security, some trades are motivated by private information, while others are purely liquidity-driven. Research on market microstructure models trading participants and their behavior stemming from different motivations. Based on these models, empirical estimation procedures aim to extract from the observed trading activity the degree to which trades are motivated by information. Easley et al.'s (1996) model and its derivatives are widely used examples, but they lack applicability in high frequency trading environments. Easley et al. (2012b) present a new evolution of this model that is intended to measure the degree of information asymmetry, particularly in high frequency trading environments—the volume-synchronized probability of informed trading (VPIN). They show that VPIN is a predictor of short-term volatility, and they demonstrate how it could have signaled the "flash crash" on May 6, 2010, hours ahead of the actual crash (Easley et al., 2011, 2012b).

This paper tests the robustness and sensitivity of VPIN to one of its key design parameters—trade classification—in order to guide researchers in its application. Computing VPIN requires separating buys from sells in the raw trading data, which is a common prerequisite for the estimation of microstructure models and hence well studied. Instead of relying on the established deterministic algorithms, however, Easley et al. (2012a) also introduce a new method to classify trades into buys and sells, called bulk volume classification (BVC). Unlike traditional classification algorithms, which operate on a trade-by-trade level, BVC defines a fraction of volume as either buyer- or seller-initiated simply on the basis of price movements within a short period of time—an inherently heuristic and potentially less accurate approach. Since the seminal work of Lee and Ready (1991), the performance of trade classification algorithms has steadily improved, reaching accuracies of above 90% (Chakrabarty et al., 2007; Ellis et al., 2000; Odders-White, 2000). We are able to confirm this finding using a proprietary subsample of data provided by Deutsche Boerse: the complex algorithms that use both trade and quote data achieve a 90.5% classification accuracy.

Our analysis is motivated by determining whether a switch to heuristics is critical, if in fact deterministic methods are accurate. The original inventors of VPIN applied the tick rule to order flow aggregated in time bars in the first working draft versions of now published and refined papers on VPIN, e.g. Easley et al. (2012b) and Easley et al. (2011). Additionally, if we assume that VPIN is a robust metric and that BVC fulfills its purpose, we can examine whether every reasonable choice of trade classification scheme actually produces roughly the same or similar results. However,

[*] Corresponding author. Fax: +49 6151 165393.
E-mail addresses: poeppe@bwl.tu-darmstadt.de, poeppe@gmail.com (T. Pöppe), sebastian.moos@student.kit.edu (S. Moos), schiereck@bwl.tu-darmstadt.de (D. Schiereck).

http://dx.doi.org/10.1016/j.jbankfin.2016.08.006
0378-4266/© 2016 Elsevier B.V. All rights reserved.

there is one caveat to this reasoning. Easley et al. (2012b) argue that informed (high frequency) traders do not trade aggressively on their information using market orders any longer. Instead, they create price pressure through fast, sliced limit order entries combined with immediate cancellations, or similar trading strategies designed to hide their true intentions. BVC, through its focus on price changes, is intended to capture this new characteristic better than precise "theoretical" order imbalance. This crucial assertion, however, has not yet been empirically proven. Even if it is true to some extent, there are almost certainly a differences in the degree to which single securities, and especially different market places, would be affected by this change. Hence, evaluating VPIN's robustness to, in other words, varying hypotheses of how informed traders trade, is vital. We further address this concern by incorporating analysis on VPIN's incremental predictive power for future volatility.

To test the sensitivity of the VPIN model to the choice of trade classification algorithm, we compare the outcomes of the different approaches on all levels of the ladder of aggregation toward empirically relevant metrics. We begin with basic trade classification results, and then aggregate the data to compute order imbalances. Next, we compute VPIN, and, finally, we use the various versions of VPIN to compare the occurrence of "toxic periods" (Easley et al., 2012b), the major proposed application of VPIN. A practical example applies different VPIN versions to a crash of a blue-chip stock, which dropped by 24% in one day.

Few studies have explored the robustness of VPIN, or, more specifically, VPIN's sensitivity to the trade classification algorithm. But those that have studied the issue have provoked some controversy around VPIN's benefits and potential flaws, such as its incremental predictive power in comparison to autocorrelation of existing measures of volatility (Andersen and Bondarenko, 2013, 2014a, 2014b, 2015; Easley et al., 2014; Wu et al., 2013b). Chakrabarty et al. (2012) use order imbalance estimation and the detection of toxic events to compare trade-by-trade algorithms with actual BVC. Andersen and Bondarenko (2015) use a sample from S&P 500 futures with near 100% classification accuracy. They show that transaction-based classification yields better results than BVC. More importantly, they conclude that VPIN(BVC) is not a superior indicator of future volatility or toxicity, and they question VPIN's ability to detect events like the 2010 flash crash (Andersen and Bondarenko, 2014b). These contrasting results suggest that more empirical evidence is needed.

The bottom-up approach we use here covers both the foundations and the applications of VPIN. In addition, we enhance the existing evidence on VPIN over several dimensions. The empirical evidence published thus far focused on the ultra-high-frequency trading E-mini S&P 500 future on the CME Globex (Andersen and Bondarenko, 2014b; Easley et al., 2011, 2012b). We take a contrasting approach, and instead provide empirical evidence of the applicability of VPIN in the high frequency trading of equities. We would expect trading in single stocks, instead of market-wide indices, to involve a much higher share of private information. Hence, research on VPIN applicability is invaluable. The trading data in this paper comes from a large sample of the top 30 German stocks (the "DAX 30") covering the full year 2012, in contrast to the two future contracts analyzed by Easley et al. (2012b) and one contract in Andersen and Bondarenko (2014b).

It originates from a pure electronic limit order book without any market maker interference, in contrast to the WTI futures contract used in Easley et al. (2012b) where a (small) share of volume is traded via open outcry. Further more, in addition to our evaluation of the tick rule, we also examine the more complex and presumably more precise trade-by-trade classification algorithms. Extending the geography to Germany has the advantage of providing a much less fragmented order flow than in the leading U.S. and U.K. equity markets. Our trading data captures almost 70% of all trading in the respective stocks; in the US, a single venue covers, on average, barely one-third of total trading.[1]

Our findings indicate a strong discrepancy between heuristic and deterministic trade classifications at the trade level. The discrepancy does not diminish in higher aggregated metrics, but instead increases when we compute order imbalances and especially VPIN. Moreover, although the average VPIN values are similar, the correlations of the time series are approximately 54%. In the detection of toxic periods, which is the major proposed application of VPIN, the two approaches do not give consistent results more often than in 60% of the time. Further, neither approach is consistently faster or earlier at detecting toxic periods. Regression analysis identifies volume and return volatility as parameters that contribute to a higher sensitivity of VPIN estimates to the choice of classification algorithm. These are exactly the trading conditions under which VPIN is supposed to be most useful. An examination of a crash of K+S, a German blue-chip stock that dropped 24% on July 30, 2013, reveals that VPIN can predict the crash, but only if trade-by-trade classification is used, not BVC.

The remainder of this paper is organized as follows. Section 2 reviews the trade classification algorithms and summarizes empirical results regarding their performance. Section 3 introduces the VPIN model and toxic periods, and explains how we run our evaluations. The data is described in Section 4. Section 5 presents our results, and Section 6 concludes.

2. Trade classification algorithms and their performance

Two major trends have affected the performance of trade classification algorithms in recent trading data. The computerization of trading, with precise millisecond timestamps, should benefit deterministic algorithms, because data inaccuracies, such as delayed quotes, are largely eliminated. However, high frequency and algorithmic trading may make trade classification more challenging. "Traditional" algorithms classify each single trade individually based on the preceding or succeeding trade price or, more commonly, they use the vicinity to the prevailing best bid or ask to determine trade direction. In contrast, BVC defines a proportion of the total volume traded within a given time frame as buy or sell volume based on price movements within that time frame.

2.1. Trade-by-trade classification algorithms

The tick rule is the simplest classification algorithm due to its neglect of quote data. Trades with a higher price than the previous (upticks) are classified as buys; trades with a lower price (downticks) are sells. In the case of equal prices, the last preceding price change is used. The tick rule is part of every complex trade-by-trade classification algorithm. But applying it can lead to misclassifications if quotes move in a direction opposite to the trade direction or if traded prices move through the best bid or ask levels.

The quote rule compares trade prices to the prevailing quote at the time of the trade. Trades at or above the ask are classified as buys; trades at or below the bid are classified as sells. Trades inside the spread are classified based on their proximity to the bid and ask (Harris, 1989). Trades at the midpoint of the spread cannot be classified by using the quote rule.

Sophisticated algorithms combine the quote and tick rule: Thus, trades at the midpoint are always classified by the tick rule, and trades at the best bid or ask are classified by the quote rule. The three most common algorithms differ in how they divide the remaining trades between quote and tick rule, as illustrated in Fig. 1.

[1] Source: fragmentation.fidessa.com/indexstats.

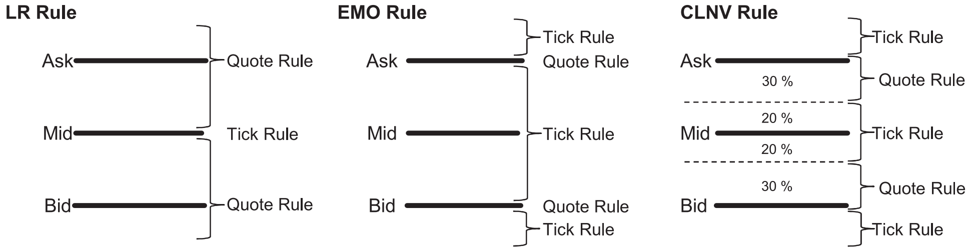

Fig. 1. Classification algorithms. This chart illustrates the functioning of three different trade-by-trade classification algorithms: LR by Lee and Ready (1991), EMO by Ellis et al. (2000) and CLNV by Chakrabarty et al. (2007).

The first of this group uses the quote rule for all trades except midpoint trades (Lee and Ready, 1991, "LR"). Interestingly, it took research almost 10 years to turn this principle up-side-down and apply the tick rule to all trades except those at the best bid and ask, which resulted in slightly better performance (Ellis et al., 2000, “EMO”). Another variation divides the spread into three parts, and uses the quote rule for trades close to the bid and ask, while applying the tick rule to the 20% band around the midpoint, and for trades outside the spread (Chakrabarty et al., 2007, “CLNV”). We use these three algorithms, in addition to the tick rule, to compare with BVC.

2.2. Bulk classification

Similarly to the tick rule, bulk classification only requires tick data to infer trade direction. Instead of classifying on a trade-by-trade basis, however, bulk classification determines the share of buys and sells of a chunk of aggregated trading volume. Trades are aggregated in either time or volume bars. Using 1 min time bars, as proposed by Easley et al. (2012b), we aggregate trades in each minute by summing their volumes and then assign the last trade price in that minute to the bar. Bars with zero trading volume are excluded. Using price changes ΔP_i between time bars and the standard deviations of price changes, $\sigma_{\Delta P}$, we define buy and sell volume of a time bar i as:

$$V_i^B = V_i \cdot Z\left(\frac{\Delta P_i}{\sigma_{\Delta P}}\right), \tag{1}$$

$$V_i^S = V_i \cdot \left(1 - Z\left(\frac{\Delta P_i}{\sigma_{\Delta P}}\right)\right) = V_i - V_i^B$$

where V_i^B and V_i^S are the respective buy and sell volume and V_i is the volume of a bar. The cumulative distribution function (CDF) of the standard normal distribution, denoted by Z, weights the volume toward either buys or sells. If the price change is zero, the volume is equally weighted between buy and sell ($Z(\frac{\Delta P_i}{\sigma_{\Delta P}} = 0) = 0.5$). If the price increases (decreases), it will be weighted more toward buy (sell) volume, with the weight dependent on the value of the price change.

The procedure of bulk classification with 1-minute time bars is illustrated in Table 1. Panel A depicts a sample of raw tick data using time, price, and volume. These trades are aggregated and classified in panel B. Table 1 shows how the size of the price change influences the buy and sell volume. Relative to all price changes, $\Delta P_1 = -0.19$ is a large decline, and consequently the entire volume in the first time bar is classified as sell. Within the next few minutes after the first opening trades, when volatility is typically higher, price changes “normalize” and volume is then divided between buy and sell.

2.3. Performance of traditional trade classification algorithms

Motivated by the need for classified trading data to estimate microstructure models, various algorithms have been proposed and their performance evaluated on markets worldwide. We summarize the results of the most important empirical studies that use actual classification data in Table 2. Most studies evaluate their algorithms by using trading data from NASDAQ and the NYSE. The LR algorithm performs fairly well, with a 73–91% accuracy depending on the study and the market. Based on an analysis of the classification accuracy of trades within the spread on a NASDAQ data set, Ellis et al. (2000) propose a new classification algorithm, EMO, that improves the accuracy on their data to 81.87%. Chakrabarty et al. (2007), in turn, propose another variation, CLNV, that outperforms EMO on NASDAQ data by 0.72 percentage points (76.52%). These two slightly more complex algorithms outperform LR by a few percentage points in three out of five studies. On the market where VPIN has been studied thus far, the Chicago Mercantile Exchange (CME), the tick rule comes close to a 90% accuracy rate (Andersen and Bondarenko, 2015; Easley et al., 2012a).

In order to study a new trading venue using data from equity markets, we contacted Deutsche Boerse to help validate one of our key assumptions—that trade-by-trade algorithms perform reasonably well even under recent market conditions. Deutsche Boerse provided us with a proprietary sample of trading data for 10 trading days from 12/3/2012 to 12/14/2012 for 10 stocks from our original sample.[2] This timeframe is outside the sample period of our main empirical analysis. The data is comprised of all the matching, limit or market orders, that make up each trade. The millisecond timestamp of each order allows us to identify the order that came last and thereby led to the execution of the trade, i.e. the aggressor side. For about 10% of the trades in the Deutsche Boerse data, we find that the timestamps of the constituting orders are identical on a millisecond level. Thus, we cannot distinguish the true trade initiator for these trades and exclude them from our evaluation. Data from ThomsonReuters is acquired accordingly for the same sample, and processed in exactly the same way as the trading data for the main empirical analysis (see Section 4). The timestamps between the two sources deviate by a range of milliseconds. But we are nevertheless able to match each single trade from Deutsche Boerse to our trading data, because the ordering of trades and their prices and volumes match 100%.

The last row of Table 2 gives our results for the trade classifications. A more detailed breakdown is in columns 3–6 of Table 3. The two most complex algorithms, EMO and CLNV, achieve the best results, with above 90% classification accuracy on a single trade level. The performance of the simpler tick rule varies more between stocks and achieves a lower accuracy of 82%. A comparison with BVC is possible after aggregating trades on either a time bar or a bucket level. We can measure the accuracy by relating the number of false classified trades per time bar or bucket to total volume, as follows:

$$Acc_{Algo} = 1 - \frac{1}{nV_t} \sum_{t=1}^{n} \left| (B_{t,Algo} - B_{t,True}) \right| \tag{2}$$

[2] We thank Deutsche Boerse and in particular Dr. Miroslav Budimir, Dr. Alexandra Hachmeister and Thomas Katzschner for their advice and support.

Table 1
Bulk volume classification of trading data. Panel A depicts a short excerpt of Adidas tick data from July 2, 2012 to illustrate bulk classification within 1-min time bars. Only a few trades at the start and end of every minute are reported. Panel B shows how the trades are aggregated into 1 min time bars, and then classified using the cumulative distribution function of previous price changes. P_i is the price of the last trade within the time bar. ΔP_i is the price change from bar $i-1$ to bar i; and CDF is the normalized cumulative distribution function of the previous price changes. The corresponding estimate of the standard deviation of the price changes is 0.0455. V_i is the total volume traded within bar i. $V_i{}^B$ and $V_i{}^S$ are the resulting volumes classified as either buy or sell, respectively, within time bar i.

Panel A: Trading data			Panel B: Time bar aggregation								
Time	Price	Volume	Time bar	Start	End	P_i	ΔP_i	CDF	V_i	$V_i{}^B$	$V_i{}^S$
9:01:12.920	55.73	30	1		9:02:00.000	55.73					
	…more trades		2	9:02:00.001							
9:02:57.378	55.54	8	2		9:03:00.000	55.54	−0.19	≈ 0	9959	0	9959
9:03:00.125	55.51	73	3	9:03:00.001							
9:03:02.160	55.50	500	3								
	…more trades		3								
9:03:55.795	55.49	228	3								
9:03:56.035	55.49	111	3		9:04:00.000	55.49	−0.05	0.136	7195	978	6217
9:04:00.517	55.51	230	4	9:04:00.001							
	…more trades										
9:04:59.096	55.57	5	4		9:05:00.000	55.57	0.08	0.961	14240	13679	561

Table 2
Accuracy of traditional classification algorithms. This table shows the accuracy of trade classification algorithms by empirical studies. Columns 2 and 3 denote the original exchange and the first year of the considered data set. The remaining columns show the reported performance for each algorithm used: the tick rule (Tick), the reverse tick rule (Re. Tick), the quote rule (Quote), the Lee–Ready algorithm (LR), the Ellis et al. algorithm (EMO), and the Chakrabarty et al. algorithm (CLNV). Peterson and Sirri (2003) examine two periods during a NYSE tick change. Chakrabarty et al. (2012a) report the accuracy of short sales (1) and long sales (2). Data of Aitken and Frino (1996) excludes trades occurring between the spread and the quote changes that preceded the trades. Savickas and Wilson (2003) exclude midpoint trades in their sample. CME futures trades in Easley et al. (2012b) could only occur at quotes. The final row presents results from this paper.

Study	Data	Year	Tick	Re. Tick	Quote	LR	EMO	CLNV
Aitken and Frino (1996)	ASX	1992	74.4					
Lee and Radhakrishna (2000)	NYSE	1990				93.0		
Odders-White (2000)	NYSE	1990	78.6		75.0	85.0		
Tanggaard (2004)	NYSE	1990	81.5		73.4	84.7		
Finucane (2000)	NYSE	1990	83.3	72.1		84.4		
Ellis et al. (2000)	NASDAQ	1996	77.7		76.4	81.1	81.9	
Theissen (2001)	FWB	1996	72.2			72.8		
Peterson and Sirri (2003)	NYSE	1997(1)				91.0	90.8	
	NYSE	1997(2)				87.4	88.2	
Savickas and Wilson (2003)	NASDAQ/ NYSE	1995	60.8		79.0	79.0	74.4	
Chakrabarty et al. (2007)	NASDAQ	2005	75.4			74.4	75.8	76.5
Chakrabarty et al. (2012)	NASDAQ	2005(1)				78.6		
	NASDAQ	2005(2)				78.3		
Easley et al. (2012a)	CME	2011	86.4					
Anderson and Bondarenko (2015)	CME	2006–2011	88.4					
The current paper	XETRA	2012	82.0			89.6	90.4	90.5

Table 3
Evaluation of trade classification algorithms with true trade initiator. This table the presents results of an evaluation of trade-by-trade classification algorithms based on a proprietary data sample provided by Deutsche Boerse specifically for this paper. The data covers a random sample of 10 stocks of our original sample and 10 trading days in December 2012 (December 3 - December 14). The second column counts the number of trades matched between the two sources Deutsche Boerse and ThomsonReuters, which equals all trades from Deutsche Boerse where a trade initiator flag is available. Columns 3–6 present the share of trades correctly identified as buy or sell by the four trade-by-trade classification algorithms on single trade level. Columns 7–11 and 12–16 present classification accuracy on time bar and bucket levels, respectively, for trade-by-trade algorithms and for BVC.

Security	Trades in sample	Trade level accuracy				Time bar level accuracy					Bucket level accuracy				
		Tick	LR	EMO	CLNV	BVC	Tick	LR	EMO	CLNV	BVC	Tick	LR	EMO	CLNV
ADSGn	23,195	.82	.90	.91	.91	.76	.87	.92	.93	.93	.87	.92	.94	.95	.95
DB1Gn	23,415	.82	.92	.92	.92	.78	.87	.93	.93	.94	.89	.92	.95	.95	.95
DBKGn	85,443	.85	.90	.91	.91	.79	.91	.92	.94	.94	.90	.95	.95	.96	.96
FMEG	23,890	.81	.90	.90	.90	.77	.86	.91	.92	.92	.88	.92	.94	.95	.95
HNKG	27,347	.81	.89	.89	.90	.78	.86	.92	.92	.92	.88	.92	.95	.95	.95
IFXGn	33,349	.83	.88	.90	.90	.77	.88	.90	.92	.92	.88	.93	.94	.95	.95
LING	16,781	.76	.88	.89	.89	.77	.82	.90	.91	.91	.88	.89	.93	.94	.94
SAPG	42,186	.79	.88	.89	.89	.78	.86	.91	.92	.92	.90	.93	.95	.96	.96
SDFGn	31,166	.84	.92	.93	.93	.78	.88	.92	.93	.93	.89	.93	.94	.95	.95
TKAG	41,476	.81	.89	.90	.90	.77	.88	.91	.93	.92	.86	.92	.94	.95	.95
Total	348,248	.82	.90	.90	.91	.78	.88	.91	.93	.93	.88	.93	.94	.96	.95

$B_{t,\,Algo}$ is the number of buys per time bar or bucket t from algorithm *Algo*, $B_{t,\,True}$ is the actual number of buys, V_t is the total volume in each bucket or time bar, and n is the number of time bars or buckets. All algorithms, and especially BVC by construction, benefit from false buys and sells offsetting each other in aggregation. Nevertheless, as Table 3 shows, trade-by-trade algorithms consistently outperform BVC across all stocks by up to 15 percentage points on a time bar level, and by 5–8 percentage points on a bucket level. Hence, trade-by-trade algorithms can serve as a precise benchmark for a heuristic classification procedure, and for testing the robustness of VPIN.

2.4. Evaluation of bulk volume classification and VPIN

Easley et al. (2012a) was the first to introduce an alternative procedure for classifying trades, as a prerequisite for VPIN. Instead of using a trade-by-trade method, they use a heuristic approach by approximating the share of buys and sells from aggregated volume and returns. In their comparison of traditional tick rule with new bulk classification, the tick rule was able to identify 86.43%, 67.18%, and 78.95%, respectively, of the volume of E-mini S&P 500 Futures, WTI crude oil futures, and gold futures correctly. If trades are aggregated into “bars” of several seconds, of up to 1-minute in length, the performance of both the tick rule and BVC increases to ranges of 70–98% (tick) and 93% (BVC). Larger bars lead to greater accuracy as more classification errors offset each other. Overall, however, the tick rule still outperforms BVC in terms of pure classification accuracy, a point analyzed in detail by Andersen and Bondarenko (2015).

In parallel to our research, Easley et al. (2016) compare BVC and the tick rule more thoroughly than in their initial publication, which introduced BVC and used it to directly calculate VPIN (Easley et al., 2012b). However, the trading data used is the same as in Easley et al. (2012b), that is, it contains no equity data. Furthermore, apart from a different methodological approach, it again only uses the tick rule in comparison to BVC, while the literature review and current results show that the tick rule is usually the worst performing algorithm among the well-known trade-by-trade classification algorithms and hence may not be an appropriate benchmark.

The major argument in favor of bulk classification is not raw buy-sell classification accuracy, however. Instead, it is the desire to capture the behavior of informed traders, who supposedly do not trade aggressively on their information using market orders any longer, but instead use a more subtle approach by creating price pressure through sliced limit orders and quick order cancellations (Easley et al., 2012a, 2016). By incorporating price change directly, BVC is intended to also capture this aggressive, limit order-driven price pressure. Consequently, whether to prefer BVC over the tick rule depends on one’s belief about an informed trader’s actual trading behavior. Easley et al. (2012a) argue favor of BVC, but they provide only limited empirical evidence, at least for markets other than those studied in their paper. Hence, whether VPIN produces consistent results with either approach remains an important question.

Using signed NASDAQ OMX data from 2005 and 2011, Chakrabarty et al. (2012) conduct a similar analysis of BVC. Again, the tick rule outperforms BVC on an aggregated bulk level, both for raw trade classifications and order imbalances. In addition, VPIN estimates of the bulk tick classification are more strongly correlated with “actual” VPIN values based on the signed trading data. BVC is more successful at classifying large and more frequently traded stocks, while the tick rule suffers from increased trading frequency in volume bars. Note that, inconsistent with Easley et al. (2012a), Chakrabarty et al. (2012) find that the bulk tick rule is a better indicator of order imbalance for all volume and time bars. The divergence of these conclusions could arise from the different methods used to estimate the accuracy of order flow imbalance. Overall, Chakrabarty et al. (2012) focus on answering “who is right” and “who is faster” at classifying trades and less on the implications for calculating and applying VPIN, as we do here. Moreover, Chakrabarty et al. (2012) consider only the tick rule, but not more advanced trade classification algorithms, as we also do here.

Andersen and Bondarenko (2014b) challenge VPIN on its designated “home turf” data, that is, high frequency futures trading in the E-mini S&P 500 and find that the incremental predictive power of VPIN compared to existing measures is not significant. However, they cannot confirm VPIN’s proclaimed detection of the 2010 “Flash Crash”, and they question the validity of its link to the theoretical microstructure work in Easley et al. (1996). In a separate study, they provide evidence that transaction-based classification is more accurate than BVC (Andersen and Bondarenko, 2015). These provocative findings have triggered a number of published replies and replies to replies, from the original authors as well as other scholars (Andersen and Bondarenko, 2013, 2014a; Easley et al., 2014; Wu et al., 2013b).

Further empirical evidence on VPIN comes from Wu et al. (2013a), who analyze its robustness in a brute-force manner by computing results for 16,000 different parameter settings. Apart from the discussion in Andersen and Bondarenko (2013), we note two striking points. First, to test VPIN’s precision, i.e., determine the true and false positives, their criterion of a true positive requires only that the maximum intermediate return to be larger than average after VPIN reaches its “alert” threshold. But the ability to predict a “larger than average” event, i.e., something that occurs roughly 50% of the time, for whatever toxicity/volatility metric, is a very weak criterion, given VPIN’s aspiration to detect events like the flash crash, which, fortunately, are not likely to happen more than a few times a year, if at all. Second, the authors already note that the choice of how to determine a price at the end of a bucket can significantly affect VPIN results: "It is somewhat surprising that computing prices differently can affect the VPIN events detected" (Wu et al., 2013a).

3. Research design

3.1. The volume synchronized probability of informed trading (VPIN)

The VPIN model of Easley et al. (2012b) builds on a microstructure trading model initially developed by Easley and O’Hara (1987). The model’s key parameters are the probability of an information event α and the arrival rates for two types of traders: μ for informed traders, who are able to observe information events and their direction, and ε for uninformed traders. The probability of informed trading (PIN) measure, initially introduced in Easley et al. (1996), is then defined as the ratio of informed trades relative to all trades:

$$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} \tag{3}$$

These parameters can be estimated with a maximum likelihood estimation that requires only the number of buys and sells of each trading day as input. In recent high frequency market environments, however, estimating the maximum likelihood function becomes difficult or impossible. VPIN addresses several shortcomings of the PIN model to make it more applicable to today’s trading activities.

There are three key steps leading from PIN to VPIN. First, VPIN uses a broader definition of information than PIN. Information events are assumed to occur frequently during a day and to differ in relevance. The arrival of relevant information triggers the arrival of trades. Therefore, a burst in volume should correlate with the degree to which information is existent and relevant. As a result, the authors opt to switch from clock time to volume time, which is the primary difference between VPIN and PIN. By using volume time, they attempt to capture the amount of information underlying the bursts in trading volume. Sampling the inputs for the model's parameters by volume time, such as arrival rates for informed and uninformed traders, implies that times with high trading activity are considered "higher resolution" than those with lower trading activity.

Second, the mathematical foundation for VPIN is laid in an approximation of the formula based on the dynamic version of the PIN model that uses the time-varying arrival rates of trades (Easley et al., 2008). The authors show that, in the fraction τ of total trading volume, the product of the arrival rate of information times the arrival rate of informed traders, $\alpha\mu,$can be approximated for as the expected absolute order imbalance $E(OI_\tau)$:

$$E(OI_\tau) = E\left[\left|V_\tau^B - V_\tau^S\right|\right] \approx \alpha\mu \tag{4}$$

Similarly, the joint arrival rate of all traders, $\alpha\mu + \varepsilon_b + \varepsilon_s$, where the uninformed trader arrival rate ε is divided between ε_b for buyers and ε_s for sellers, can be approximated for by the sum of the expected total number of trades $E[V_\tau^B + V_\tau^S]$:

$$E\left[V_\tau^B + V_\tau^S\right] \approx \alpha\mu + \varepsilon_b + \varepsilon_s \tag{5}$$

Joining these two approximations with the original definition of PIN in Eq. (3), we can define VPIN as follows:

$$\text{VPIN} = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} \approx \frac{E\left[\left|V_\tau^B - V_\tau^S\right|\right]}{E\left[V_\tau^B + V_\tau^S\right]}. \tag{6}$$

In effect, all trades are partitioned into τ buckets with a constant amount of volume V (but with different time spans) in order to "mimic the arrival to the market of news of comparable relevance" (Easley et al., 2012b, p. 1465). The authors illustrate the VPIN computation by setting the average daily number of buckets to 50. This corresponds to V equaling one-fiftieth of average daily volume. Then, with the rolling window length = 50, VPIN can be written as:

$$\text{VPIN} = \frac{\sum_{\tau=1}^{n} \left|V_\tau^B - V_\tau^S\right|}{nV}. \tag{7}$$

The order imbalances $\sum_{\tau=1}^{n} |V_\tau^B - V_\tau^S|$ are determined by filling each bucket τ with the respective buy and sell volume.

Third, he heuristic estimation of buy and sell volume is as described in Section 2.2. As long as the aggregated volume in a time or volume bar is insufficient to fill a bucket, the volume of the subsequent bar will be used. Once a bucket is full, any remaining volume will be moved to the subsequent bucket, thereby splitting the volume across adjacent buckets. Since the very last bucket is always incomplete or empty, it is ultimately neglected in VPIN calculations.

Note further that VPIN is computed in a rolling process, i.e. it is the moving average of the order imbalance over the last n preceding buckets. With the parameters chosen by Easley et al. (2012b), Eq. (7) corresponds to a daily VPIN. The advantage of using volume time is especially apparent in the process of updating VPIN, where the last bucket is dropped and the next bucket is added. This allows VPIN to be updated at a speed similar to the arrival of information, assuming that each bucket holds a homogenous amount of information.

3.2. Evaluation of the sensitivity of VPIN to the choice of trade classification algorithm

The primary purpose of this paper is to evaluate if, how, and under what circumstances the choice of trade classification algorithm influences the computation and application of VPIN. Although the intentions of both approaches are equivalent, they may not show identical results, because classification differences can influence higher-order, aggregate metrics. The question is to what degree the differences detected in raw trade classifications create biases in more sophisticated, aggregate measurements such as VPIN, and whether those differences vanish when the data is aggregated. Our approach follows the idea to run a step-by-step analysis, in a bottom-up manner, beginning with trade classification, proceeding to order imbalance, then actual VPIN calculations, and, finally, to toxic periods, the main application of VPIN.

3.2.1. Comparison raw trading data classifications

To begin a comparison of trade-by-trade classification and BVC, we first let all algorithms classify the 34 million trades of our XETRA data for 2012. The quote rule is omitted, as it leaves midpoint trades unclassified. Note that the BVC and trade-by-trade algorithms cannot be directly compared given the different granularity of their results. We therefore sign the volume of each trade as buyer- or seller-initiated based on the results of the trade-by-trade classification, and then sum the buyer- and seller-signed volume over 1 min time bars (hereafter, Bulk$_{\text{Trad}}$). This allows us to compute results on the same level of granularity as BVC. We can compare the share of volume classified as buy (or sell) and then calculate the correlation of the buy classification rate between BVC and the trade-by-trade algorithms to determine to what degree the results align.

3.2.2. Application 1: order flow imbalance

Order flow imbalance is the first step up the complexity hierarchy of measures that require trade-initiator signed data as input. Order imbalance has been shown to be correlated with, or to be a reliable predictor of informed trading (Easley et al., 2012a), and it is the main input for VPIN. We calculate two different metrics (see also Chakrabarty et al., 2012) to compare the order imbalances based on trade-by-trade and bulk classification algorithms. Both use the common definition of order imbalance, i.e., the signed difference between buy and sell volume:

$$OI_{ji} = B_{ji} - S_{ji} \tag{8}$$

We index stocks using the letter j and time bars using the letter i. B_{ji} is thus the buy volume for stock j in time bar i, and S_{ji} is the corresponding sell volume. The volume-adjusted order imbalance match over k time bars is then defined as:

$$VaOIM_j = 1 - \frac{1}{2k}\sum_{i=1}^{k} \frac{\left|Trad\left(OI_{ji}\right) - BVC\left(OI_{ji}\right)\right|}{V_{ji}} \tag{9}$$

where V_{ji} is the volume of bar i, $Trad(OI_{ji})$ is the order imbalance in bar i (calculated using the trade volume signed by the trade-by-trade algorithms), and $BVC(OI_{ji})$ is the order imbalance calculated using BVC. Multiplying the sum by $\frac{1}{2}$ lets the term on the right-hand side take a maximum of 1 if the classification results in each time bar are exactly opposite. In turn, the variable $VaOIM_j$ has a minimum of 0 and a maximum of 1 if order imbalances are the same for both algorithms.

We define the directional difference in order imbalance stock j as:

$$DirOIM_j = \frac{1}{k}\sum_{i=1}^{k} \frac{\left|sgn\left(Trad\left(OI_{ji}\right)\right) + sgn\left(BVC\left(OI_{ji}\right)\right)\right|}{2} \tag{10}$$

The directional match in order imbalance gives the percentage of bars where based on trade-by-trade algorithms matches that based on trade volume signed by bulk classification. While the definitions above are given for order imbalances based on time bars i, both metrics are also calculated by using order imbalances calculated over buckets τ. Because buckets span several time bars, errors are more likely to offset each other, and hence we expect to find greater consensus between the two approaches.

3.2.3. Application 2: VPIN estimation

The third level up from classified trading data is the VPIN variable itself. VPIN is a specific aggregation of order imbalances over several buckets of volume. The bucket size needs to be set to a fraction of the average daily trading volume. We use the exact specification of VPIN that was used and advocated for by the original authors in the key papers on VPIN (Easley et al., 2011, 2012b). That is, we use 1/50 of average daily trading volume as the bucket size, with VPIN calculated over 50 buckets, which, on a day with average trading volume, corresponds to one VPIN calculation spanning one trading day.[3] Andersen and Bondarenko (2014a, 2014b, 2015) also use a bucket size of 1/50 and a VPIN window length of 50. Hence, we study the most relevant choice of parameter settings.

Another design choice to be made concerns the timespan used for updating average bucket size. We calculate VPIN with both a constant bucket size for the entire year and, as a robustness check, the bucket size for each month based on the monthly average trading volume. The results are nearly identical. Therefore, the discussion in the remainder of this paper is based on “yearly data”, where the bucket size is fixed for the entire observation period.

3.2.4. Application 3: toxic periods

The VPIN metric was developed to detect the degree of toxic order flow that results from asymmetric levels of information among traders. The particular value of VPIN that should alert traders is inevitably somewhat arbitrary. We follow the common definition used in recent publications regarding VPIN: “A toxic period begins when the empirical cumulative distribution function (CDF) of VPIN reaches or crosses the 0.9 percentile and ends when the CDF falls below the 0.8 percentile” (Chakrabarty et al., 2012, p. 25). Accordingly, a toxic period begins with a bucket where the VPIN calculated with the current and last 49 buckets crosses the 0.9 percentile. The 49 buckets before the current bucket all contribute to the toxicity measured at bucket 50, but we only refer to bucket 50 as “toxic”, because VPIN reaches its critical level only with the inclusion of that bucket. All consecutive buckets where the VPIN estimated from the rolling last 50 buckets does not fall below the 0.8 percentile form the “toxic period”.

The toxic periods are the highest-order metric in terms of aggregation and complexity that we use in this paper. We conduct two tests in order to compare the toxic periods calculated from BVC to toxic periods based on trade-by-trade classification results. We identify all toxic periods with VPIN(BVC), and test how many toxic buckets within this period overlap with those identified by VPIN(Bulk$_{\text{Trad}}$). The first test, toxic *period* match, requires an overlap of only one bucket to assess the two methods as equal. The second test, toxic *bucket* match, denotes the percentage of toxic buckets in a period of VPIN(BVC) that is also regarded as toxic by VPIN (BulkTrad).

3.2.5. Determinants of VPIN differences

Note that, as we mentioned earlier, the computation of VPIN using two different trade classification strategies may not yield the exact same results. To resolve this issue, we need to determine under what circumstances either classification scheme is more applicable, or when the bias induced by the choice of algorithm is especially large. Therefore, we design a regression model to test the influence of trading and stock characteristics on the differences in VPIN levels. We run the model in volume time. The dependent variable $\Delta VPIN_{j\tau}$ denotes the difference between every VPIN value calculated with BVC minus VPIN calculated with Bulk$_{\text{Trad}}$.Index j numbers the companies, and index τ numbers the VPIN estimations per company (or buckets, beginning from bucket 50 for the first VPIN calculation, and proceeding onward).

We hypothesize different liquidity levels and return volatility will affect the differences, because trade-by-trade algorithms can run into more errors during high frequency trading and BVC is defined based on the basis of returns. In doing so, we aim to distinguish a “size effect”, in terms of both trading volume and returns, from a “volatility effect”. In other words, do the methods differ because of different levels of trading activity or do they differ during periods of abnormal trading activity—or both?

We intend to measure the size effect using the variables $Trades_{j\tau}$ and $ReturnVola_{j\tau}$, which represent the total number of trades and the volatility of the returns over the time bars within each VPIN. Whether current trading conditions are abnormal is measured by the variable $VolumeVola_{j\tau}$, which represents the volatility of traded volume over time bars. The stock return, split into two variables $posReturn_{j\tau}$ and $negReturn_{j\tau}$, captures rising and falling market conditions, respectively.. The variable $Timebars_{j\tau,}$ which contains the number of time bars during each VPIN estimation as a standardized measure of abnormal volume (more time bars required to fill the VPIN buckets means less than average trading volume during that period). In other words, the number of time bars indicates how far a VPIN estimation is spread across clock time.

Variations on stock level, which are likely present, are taken into account using fixed effects per stock. Trading volume itself cannot be chosen as an explanatory variable because it is, by construct, constant per stock and VPIN estimation. This results in the following model:

$$\begin{aligned}\Delta VPIN_{j\tau} = {} & \beta_0 + \beta_1 Trades_{j\tau} + \beta_2 ReturnVola_{j\tau} + \beta_3 VolumeVola_{j\tau} \\ & + \beta_4 posReturn_{j\tau} + \beta_5 negReturn_{j\tau} + \beta_6 Timebars_{j\tau} \\ & + \alpha_j + u_{j\tau} \qquad (11)\end{aligned}$$

The independent variables relate to the last 50 buckets over which the two VPINs for the dependent variable are calculated. Hence, on a day with roughly average trading volume, all variables span roughly one trading day. Using the signed difference as dependent variable allows us to conclude which factors lead to divergent VPIN measures, while also allowing certain relationships to offset each other, if present.

We differentiate between two versions of the specified model to account for VPIN's calculation over a rolling window of 50 buckets. In one model, we only use only every 50th VPIN observation in order to avoid introducingautocorrelation by design. In a second approach, we use every VPIN observation but correct for autocorrelation using Newey–West standard errors for 49 lags.

3.2.6. Predictive power of VPIN

The analysis in the preceding sections tells us whether VPIN is sensitive to the choice of trade classification, and in what trading environments this sensitivity is most pronounced. From Section 2.3, we also know that the trade-by-trade classification algorithms perform exceptionally well at classifying single trades. Ultimately, however, Easley et al. (2012b) argue that VPIN calculated with BVC captures something beside just precise “theoretical” order imbalance. The fast order entries and immediate cancellations pursued by high frequency traders, it is argued, render the identification of the actual aggressor side somehwat ambiguous. Thus, BVC is considered to be better at capturing *actual* informed trading, or toxicity of order flow.

To address this valid concern, we extend our analysis in two ways. First, in the spirit of the “flash crash” analyzed in the initial VPIN studies, we evaluate VPIN's ability to detect a single-security severe crash. Second, we test VPIN's predictive power for future

[3] Easley et al. (2012b) experiment with bucket and window sizes ranging from 10 to 250 but run their main analysis with results based on a bucket size of 1/50th and a VPIN window size of 50.

volatility. Both Easley et al. (2012b) and Andersen and Bondarenko (2015) evaluate VPIN's ability to predict future volatility or future price movements. We choose an approach similar to Andersen and Bondarenko (2015) in order to test VPIN's *incremental* predictive power for 1-day volatility using the following fixed effects regression model:

$$RV_{jt} = \beta_0 + \beta_1 RV_{jt-lag} + \beta_2 VPIN^{Algo}_{jt-lag} + \alpha_j + u_{jt} \tag{12}$$

where index t denotes intervals of 10 min length; *lag* is fixed at 51 to lag observations by exactly one trading day (one trading day spans 8.5 h, i.e. 51 10-min intervals). RV_{jt} is realized volatility for stock j at interval t, calculated as the average of the absolute log returns over the last trading day's 10-min-intervals, i.e. 51 lags. $VPIN^{Algo}_{jt-lag}$ is the latest complete VPIN calculation available at time t for stock j, based on either BVC or one of the trade-by-trade algorithms. We run regressions for 5 and 15 min intervals, and respective lags of 102 and 34 as robustness checks.

4. Data

This study uses trade and quote data from XETRA, the electronic trading system of Deutsche Boerse, which is by far the leading stock exchange in Germany.[4] XETRA offers an anonymous order-driven market model with an electronic limit order book. Orders are executed by price/time priority. Trading begins with an opening auction, followed by continuous trading, which is interrupted by an intraday auction at 1.00 pm and ends with a closing auction. We choose approximately one year of trading data for our analysis, that is, the 234 trading days from January 2, 2012 through November 27, 2012. We limit the analysis to stocks listed on the DAX, which is the leading stock index in Germany and includes the 30 largest and most traded stocks.[5] The raw data contains roughly 50 million quotes and 34 million trades, representing 29 billion shares traded.

Equity trading in the DAX on XETRA is on par, if not even more frequent than trading at the other major stock exchange in Europe, the London Stock Exchange (LSE). The average daily traded volume of the DAX 30 in the first quarter of 2014 ranges from 3.2–3.8 billion EUR, which is roughly 80% of the daily traded volume of the 100 stocks of the FTSE 100 traded on the LSE.[6] Moreover, the total turnover of DAX 30 stocks alone in 2012 was 803 billion EUR, which equals 55% of the traded volume of all stocks on the LSE in the same year.

Out dataset as the unique advantage of being much less fragmented other major marketplaces. Deutsche Boerse captures 70% of global trading in the DAX 30. In contrast, the leading U.S. exchange, the NYSE, only captures 35% and 29%, respectively, of trading in the Dow Jones and S&P 500 stocks. The concentration of stocks on the NASDAQ, the data used in Chakrabarty et al. (2012), is below 50%. Trading on the FTSE 100 is slightly less fragmented, but is still lower, with only 62% of traded volume executed at the LSE.[7] Thus, the dataset used here is both high frequency and comes closest to covering the full order stream in the observed security. Hence, it provides the perfect conditions under which to properly assess an order flow toxicity metric like VPIN.

Our data comes from ThomsonReuters TickHistory. Trades, quotes and auctions are reported in milliseconds, and we exclude all trades and quotes resulting from opening, closing, and intraday auctions. Holden and Jacobsen (2014) demonstrate how a lack of precision and diligence in the processing of intraday data can introduce a heavy bias to the results. We draw on their excellent list for raw trading data, all of which we validate on our own data: And we find no negative trade prices or quotes, orcrossed or blocked spreads, at any time.

Table 4
Distribution of trades relative to spread. This table reports the location of trades relative to the spread, and the share of those trades classified as a buy, for each of the traditional trade classification algorithms. The sample contains a total of 34,335,259 trades.

Location	Share of trades	Classified as buy (%)					
		Tick	Re. Tick	Quote	LR	EMO	CLNV
Above Ask	6.7	92.8	51.8	100	100	92.8	92.8
At Ask	34.8	79.1	48.6	100	100	100	100
Below Ask	6.4	65.4	47.0	100	100	65.4	76.0
Midpoint	4.3	50.0	49.7	0	50.0	50.0	50.0
Above Bid	6.3	34.2	53.0	0	0	34.2	23.7
At Bid	34.5	21.1	51.4	0	0	0	0
Below Bid	7.0	7.3	48.8	0	0	7.3	7.3
Total	100	50.0	50.0	47.9	50.1	50.1	50.1

The first column of Table 4 shows the distribution of trades based on their position relative to the spread: 69.3% are at the spread, 13.7% are outside the spread, and 17% are inside the spread. The relatively large share of trades inside the spread is surprising at first. A certain percentage of these trades, as per experts at Deutsche Boerse, is due to three types of special order executions in the XETRA trading model that allow trades to be cleared inside the spread: hidden orders, “Xetra Midpoint”, and “Xetra BEST”. The validation of our data with proprietary data retrieved directly from Deutsche Boerse, as described in Section 2.3 shows that we can match 100% of trades from Deutsche Boerse to our sample. This is very strong evidence that any bias in the results does not stem from incorrectly processed data.

5. Results

This section discusses our empirical from raw trade classification up to toxic periods (ascending in complexity and aggregation). The last subsections use results from regression analyses to identify market conditions where VPIN is most sensitive to the choice of classification algorithm, and also to evaluate VPIN's predictive power for future volatility.

5.1. Trade classification with trade-by-trade algorithms and true initiator flag

Our first level of comparison is the pure classification of trades as buyer- or seller-initiated, after calibration of the trade-by-trade algorithms as described in the section on pre-processing. The buy classification rate of all trade-by-trade algorithms is shown in Table 4, where results are divided by the trade's location relative to the spread. As expected, we find that the buy classification rate is close to 50% on an aggregated yearly level. Differences between the algorithms become more apparent when we compare the classification rates by each trade's location relative to the quotes. The buy classification rate of the tick rule reaches its highest value above the ask and declines until below the bid. Approximately 20% of trades at the ask are classified as sells, which indicates a high number of quote movements contrary to the trade direction. The quote rule classifies all trades above the midpoint of the spread as buys, and LR, EMO, and CLNV yield combinations of the results of the tick and quote rule. Analyzing the classification results on a single-stock level shows no relevant differences or patterns based on stock characteristics.

[4] See Deutsche Boerse's XETRA website: http://xetra.com/xetra/dispatch/en/kir/navigation/xetra/300_trading_clearing/ 100_trading_platforms/100_xetra.

[5] The stocks Lanxess and Continental are dropped from the sample as they joined the DAX only in late 2012.

[6] http://www.londonstockexchange.com/exchange/statistics/share-of-trading/lit-figures/FTSE100.html.

[7] Fragmentation numbers from http://fragmentation.fidessa.com/.

Table 5
Correlation of trade classification results. This table reports the correlations of the time bar series of the share of volume classified as buy between BVC and the four trade-by-trade algorithms. The values are averaged over all stocks by applying Fisher's z transformation on the stock values, using the number of time bars as sample lengths.

	BVC	$\text{Bulk}_{\text{Tick}}$	Bulk_{LR}	Bulk_{EMO}	$\text{Bulk}_{\text{CLNV}}$
BVC	1	.643	.484	.530	.519
$\text{Bulk}_{\text{Tick}}$		1	.635	.720	.697
Bulk_{LR}			1	.931	.953
Bulk_{EMO}				1	.985
$\text{Bulk}_{\text{CLNV}}$					1

5.2. Comparison of trade-by-trade and bulk classification

Results for bulk classification cannot be presented in a partition of trades relative to their location in the spread. On a total aggregate, as well as a single stock level (not shown), the buy classification rates of the trade-by-trade algorithms and BVC are all close to 50%.[8] BVC yields a slightly lower buy percentage rate of 49.55%, differing from $\text{Bulk}_{\text{Tick}}$ by 0.45 percentage points. The other three algorithms' buy classification rates range from 50.22% to 50.36%.

But more revealing than comparing aggregated buy classification rates are the correlations of the time bar series. Table 5 shows the results for every pair of algorithms. The correlation values among the sophisticated trade-by-trade algorithms themselves can serve as benchmarks for comparison with BVC—and the values for LR, EMO, and CLNV are well above 90%. However, the correlation of trade-by-trade algorithms and BVC reaches 64% for the tick rule, 48% for LR, 53% for EMO and 52% for CLNV. These levels are surprisingly low, given that these algorithms can achieve accuracies of up to 90%, as shown in the literature review and the performance evaluation in this paper. So how can the bulk classification computation provide such different results? The trade-by-trade algorithm deemed the least accurate of the four algorithms, the tick rule, has a substantially higher correlation with BVC than the other, more accurate algorithms, EMO and CLNV. This is likely due to the intrinsic design of the tick rule. In rising markets, it tends to classify more trades as buys because every trade going up in price is classified as a buy. This is not the case for EMO and CLNV algorithms, however, which also capture quote data. Consequently, the bulk classification evaluates trades substantially differently. Given that the purpose of these algorithms is the same, the low correlations of the time bar series support our concern that the heuristic approach of bulk classification is measuring something different than what it is intended. The following sections analyze how these differences translate into different results for the calculations of order imbalance, VPIN, and toxic periods.

5.3. Order flow imbalances in time bars and volume buckets

The next level up from raw trading data is the computation of order imbalance; results are shown in Table 6. The volume-adjusted order imbalance match between trade-by-trade algorithms and BVC is around 65% for the small bars of 1-min. It increases closer to 90% for the larger buckets. A higher value for the buckets is expected given that they aggregate more volume. The direction of the imbalance is equal in approximately 70% of cases, except for the tick rule, where bulk and trade-by-trade agree in 75% of the cases.

These two more complex metrics for comparing order imbalances again show that the tick rule is slightly closer to the bulk classification than the other algorithms. The difference is small regarding the volume-adjusted order imbalance match, and larger regarding the direction imbalance match. Overall, however, the ranges of the absolute values suggest that trade-by-trade and bulk classification do not measure the exact same phenomena. On the other hand, the higher volume-adjusted order imbalance match in buckets, and the much higher level of agreement compared to raw trade classification, would indicate that differences diminish once we continue to aggregate further to higher-order metrics.

Table 6
Order imbalance match. This table compares order imbalances calculated from BVC results with those based on classification results of the trade-by-trade algorithms. We use two metrics. Columns 2 and 3 report the volume-adjusted order imbalance match *VaOIM*. The right-hand side of the table reports the directional order imbalance match *DirOIM*. For both metrics, the results are presented based on time bar and bucket granularity. All values are cross-sectional averages.

	Volume-adjusted order imbalance match *VaOIM*		Directional order imbalance match *DirOIM*	
	% Time bars	% Buckets	% Time bars	% Buckets
$\text{Bulk}_{\text{Tick}}$	.647	.893	.771	.742
Bulk_{LR}	.649	.872	.698	.673
Bulk_{EMO}	.652	.879	.718	.692
$\text{Bulk}_{\text{CLNV}}$	.651	.878	.712	.689

Table 7
Summary of VPIN results. This table gives the results of the VPIN computations based on input from BVC and each of the four trade-by-trade algorithms. The first row contains the cross-sectional averages. The next three rows (minimum, median, and maximum) are based on VPIN values averaged on a single stock basis.

Stock	BVC	$\text{Bulk}_{\text{Tick}}$	Bulk_{LR}	Bulk_{EMO}	$\text{Bulk}_{\text{CLNV}}$
Average	25.8	26.1	25.2	24.8	24.9
Minimum	23.1	18.3	18.0	18.0	18.0
Median	25.6	25.6	25.2	24.7	24.7
Maximum	28.1	34.5	33.0	32.2	32.4

5.4. VPIN estimation

We next discuss cross-sectional averages and the time-series of VPINs to check the correlations computed using each trade classification algorithm. The cross-sectional averages for the five different algorithms and the ranges of the averages per stock, i.e., the minimum, median, and maximum, are in Table 7. The highest cross-sectional average value is measured for VPIN($\text{Bulk}_{\text{Tick}}$) with 26.09%, while VPIN(BVC) is slightly lower at 25.76%, and VPIN(Bulk_{EMO}) yields the lowest average at 24.81%. These values are in the ranges of those reported by Abad and Yagüe (2012), but exceed the average VPINs of futures given by Easley et al. (2012b) by roughly four percentage points.

Note that the VPIN values averaged over the cross-section do not appear to differ significantly, but the range of VPIN values on a stock level do, as shown in the last three rows of Table 7 and illustrated in the scatter plot in Fig. 2: VPIN(BVC) is plotted on the horizontal axis and VPIN($\text{Bulk}_{\text{Tick}}$), as representative of VPIN($\text{Bulk}_{\text{Trad}}$), is on the vertical axis. We make three observations.

First, VPIN(BVC) and VPIN($\text{Bulk}_{\text{Tick}}$) do not yield consistently similar results. Second, the deviations does not appear to be random. Instead, for smaller than average VPIN(BVC) values, the difference from VPIN($\text{Bulk}_{\text{Trad}}$) is negative; for values larger than average VPIN(BVC) values, the difference from VPIN($\text{Bulk}_{\text{Trad}}$) is positive. In other words, the slope in the graph is much steeper than the 45° line we would expect to see if both approaches yielded roughly the same results. The results of VPIN($\text{Bulk}_{\text{Tick}}$) ultimately spread across an interval that is twice as large as the min–max spread for values of VPIN(BVC). For example, the VPIN values for Munich Re (MUV) differ by 7.69 percentage points between VPIN($\text{Bulk}_{\text{Tick}}$) and VPIN(BVC). The standard deviation for VPIN(BVC) is $\sigma = 1.64$, and for VPIN($\text{Bulk}_{\text{Trad}}$) it is $\sigma = 4.2$. Traditional trade classification

[8] We use only 1-min time bars, as in the original formulation.

Table 8
Correlation of VPIN results. This table reports the correlations of the time bar series between VPIN(BVC) and the VPIN results based on trading data that is classified using the four trade-by-trade algorithms. The total values are averaged by applying Fisher's z-transformation with the number of VPIN observations as sample lengths.

	VPIN(BVC)	VPIN($\mathrm{Bulk}_{\mathrm{Tick}}$)	VPIN($\mathrm{Bulk}_{\mathrm{LR}}$)	VPIN($\mathrm{Bulk}_{\mathrm{EMO}}$)	VPIN ($\mathrm{Bulk}_{\mathrm{CLNV}}$)
VPIN(BVC)	1	.553	.542	.531	.532
VPIN($\mathrm{Bulk}_{\mathrm{Tick}}$)		1	.790	.830	.820
VPIN($\mathrm{Bulk}_{\mathrm{LR}}$)			1	.957	.966
VPIN($\mathrm{Bulk}_{\mathrm{EMO}}$)				1	.994
VPIN($\mathrm{Bulk}_{\mathrm{CLNV}}$)					1

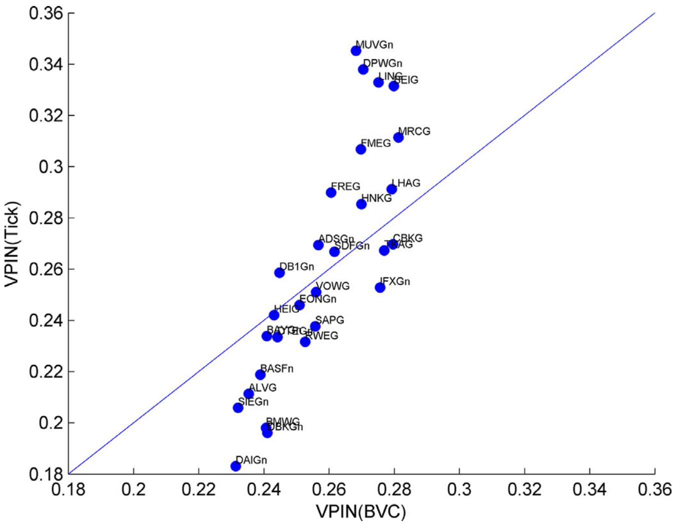

Fig. 2. Scatter plot of VPIN(BVC) and VPIN(BulkTick). This figure plots the volatility of VPIN(BVC) on the horizontal axis against the volatility of VPIN(BulkTick) on the vertical axis, on a single-stock basis. Average volatilities are $\sigma_{BVC} = 1.642$ and $\sigma_{Tick} = 4.535$.

algorithms seem to induce a higher cross-sectional variance in VPIN estimates.

This observation changes on a single-stock level if we examine the time-series of VPINs throughout the year. Fig. 3 illustrates the series of VPIN($\mathrm{Bulk}_{\mathrm{Trad}}$) and VPIN(BVC) for the stocks with the highest and lowest VPIN averages, that is, Daimler (DAI) with 18.02% and Munich Re (MUV) with 34.52%. The differences between the two series on a stock level are clearly visible. The VPIN($\mathrm{Bulk}_{\mathrm{CLNV}}$) for Daimler is generally lower and does not exhibit such high peaks as the results for BVC do. In contrast, the VPIN($\mathrm{Bulk}_{\mathrm{Tick}}$) of Munich Re is higher than that for VPIN(BVC), but is a poor match for the movements of VPIN(BVC). Other stocks show similar characteristics of visibly diverging graphs for VPIN calculations based on BVC and trade-by-trade algorithms.

The results in Table 8 generalize the observations from the line charts by showing the correlations of VPIN computations. As the first row of Table 8 shows, the time-series of VPIN($\mathrm{Bulk}_{\mathrm{Trad}}$) and VPIN(BVC) are not the same. Overall, VPIN($\mathrm{Bulk}_{\mathrm{Tick}}$) exhibits the highest correlation with VPIN(BVC) at 55.3% and VPIN($\mathrm{Bulk}_{\mathrm{EMO}}$) the lowest correlation, at 53.1%. To provide some further examples on a single-stock level, the correlations between VPIN(BVC) and VPIN($\mathrm{Bulk}_{\mathrm{Tick}}$) are only 40.7% for Munich Re and 51.7% for Daimler; the highest correlation values are slightly above 70% for Lufthansa. Compared to the metrics in the previous sections, the VPIN correlations computed with different trade classification algorithms are on similarly low levels as the correlations of the raw classification results in Table 5. The low variance between VPIN($\mathrm{Bulk}_{\mathrm{Trad}}$) and VPIN(BVC) for the calculation of the order imbalance reported in Table 6 does not diminish further when calculating VPIN. Instead, surprisingly, the variance increases.

The correlations among the trade-by-trade algorithms themselves, as shown in rows two to five of Table 8, can serve as benchmarks for the expected correlation levels. The VPIN($\mathrm{Bulk}_{\mathrm{Trad}}$) series correlate among themselves to a much higher degrees with rates from 79% to 99%. Similarly to the correlations of the time bar series, the sophisticated trade-by-trade algorithms show correlations of well above 90%. Once again, this implies that our results indicate that the bulk classification is measuring something other than what it should be.

The average correlations provide rough indications of the consensus between the VPIN series, given the high variations in stock levels. However, differences in VPIN computations, such as the reported 7.69 percentage points for Munich Re and the single-stock correlations of around 30%, are a cause for concern for the application of the VPIN methods. The next section tests how these large differences affect the detection of toxic periods, one of the proposed applications of VPIN.

5.5. Toxic periods

The results for toxic periods are in Table 9 and illustrated in Figs. 4 and 5. The number of toxic periods across stocks in our sample varies from a maximum of 47 (Deutsche Boerse) to a minimum of 10 toxic periods within one year (SAP). Fig. 4 illustrates the VPIN(BVC) series of these stocks for the entire year, together with the closest matching VPIN($\mathrm{Bulk}_{\mathrm{Trad}}$) series. The graphs show that a large number of toxic periods is not equivalent, with high volatility in VPIN. Instead, the VPIN series of SAP displays only one-fifth the number of toxic periods, but the peaks are higher and last longer than the 47 peaks of Deutsche Boerse shown in the top chart.

Of the 784 toxic periods identified by VPIN based on BVC, $\mathrm{Bulk}_{\mathrm{Tick}}$, $\mathrm{Bulk}_{\mathrm{EMO}}$, and $\mathrm{Bulk}_{\mathrm{CLNV}}$ identify about 55% of those periods as well. On a single-stock level, the share of consistent alerts varies across stocks and trade-by-trade algorithms. The highest overlap of equally detected toxic periods is 81.8%, with $\mathrm{Bulk}_{\mathrm{LR}}$ and $\mathrm{Bulk}_{\mathrm{CLNV}}$ for Adidas (ADS). The lowest overlap is 32.5%, with $\mathrm{Bulk}_{\mathrm{Tick}}$ for Bayer (BAY). Note, however, that we classify toxic periods as overlapping if at least one toxic bucket overlaps, but not necessarily for the entire toxic period. Therefore, we demand the absolute minimum in order to count a toxic period as a “match” between the two approaches. Nevertheless, the numbers remain quite low.

The right-hand side of Table 9 gives the results for the second comparison, where we measure the percentage of toxic buckets that overlap, instead of the toxic periods of several buckets' lengths. $\mathrm{Bulk}_{\mathrm{LR}}$, which scores lowest at detecting toxic periods, achieves the highest score at detecting single toxic buckets, with a 48.2% overlap with the 35,553 buckets detected by BVC. A high accuracy on the period level does not seem to imply great accuracy on the single bucket level, however. The other three algorithms achieve matches to the VPIN bulk classification of 47.5%. Again,

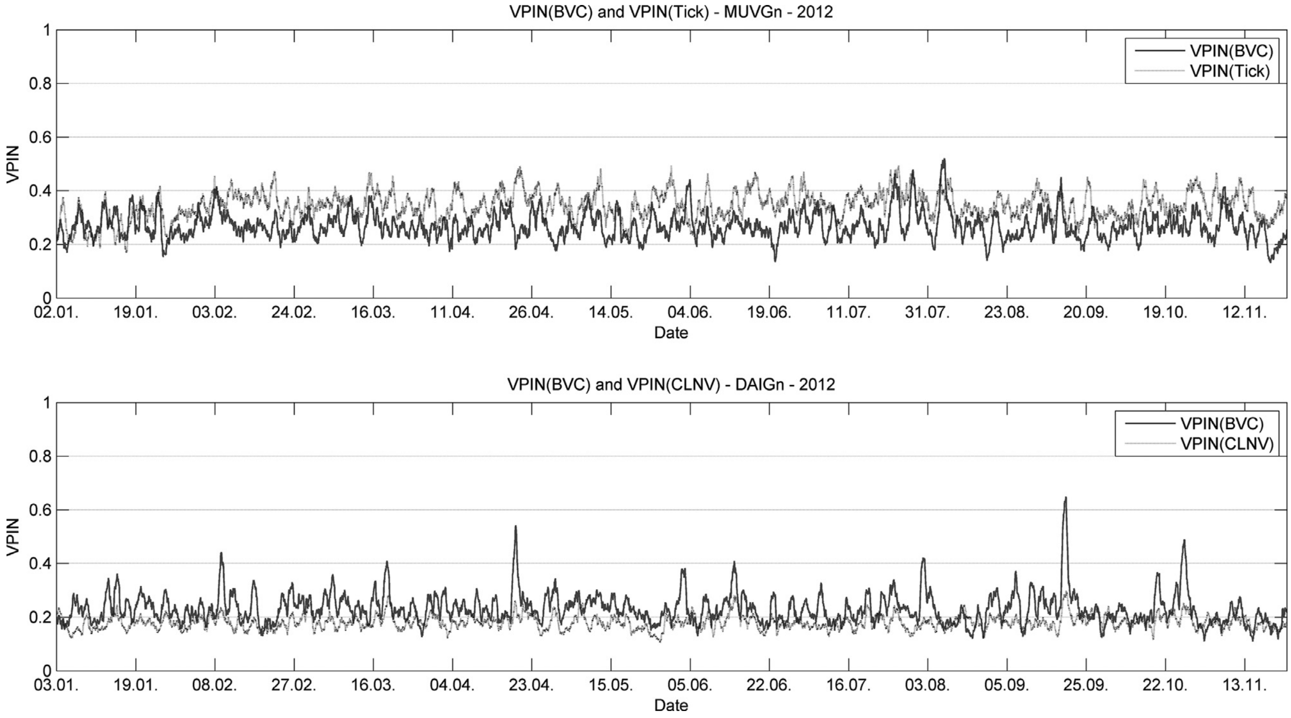

Fig. 3. Time-series of VPIN for selected stocks, in volume time. This figure illustrates the VPIN time-series based on both trade-by-trade algorithms and BVC. The first graph shows the stock with the highest VPIN average (Munich Re—MUVGn), the second shows the stock with the lowest VPIN average (Daimler—DAIGn). The VPIN(BVC) time-series are in black. The horizontal axis is in volume time scale.

Table 9
Toxic periods. This table shows the number of toxic periods and toxic buckets per stock. The numbers in columns 2, 7 and 8 are shown for calculations based on BVC. Columns 3–6 report the share of toxic periods identified by both VPIN(BVC) and VPIN(BVC_{Trad}), named toxic period match. The last four columns report the share of toxic buckets equally identified by VBPIN(BVC) and VPIN(BVC_{Trad}). The "total" row shows the cross-sectional averages.

Stock	# Toxic periods	Toxic *period* match				# Toxic VPINs	# VPINs	Toxic *bucket* match			
		$Bulk_{Tick}$	$Bulk_{LR}$	$Bulk_{EMO}$	$Bulk_{CLNV}$			$Bulk_{Tick}$	$Bulk_{LR}$	$Bulk_{EMO}$	$Bulk_{CLNV}$
Total	784	.55	.53	.55	.56	35,553	326,228	.48	.48	.47	.48
ADSGn	22	.77	.82	.73	.82	1216	11,651	.51	.55	.52	.55
ALVG	40	.40	.35	.48	.45	1473	11,651	.26	.26	.30	.27
BASFn	47	.55	.38	.43	.45	1787	11,651	.37	.31	.32	.33
BAYGn	40	.33	.48	.43	.48	1353	11,651	.29	.37	.36	.34
BEIG	16	.75	.81	.81	.81	1171	11,651	.65	.78	.72	.75
BMWG	37	.51	.49	.54	.51	1405	11,651	.45	.47	.43	.46
CBKG	27	.56	.63	.56	.59	1233	11,651	.35	.46	.42	.43
DAIGn	32	.41	.47	.41	.44	1101	11,651	.43	.42	.44	.44
DB1Gn	47	.40	.45	.47	.49	1673	11,651	.31	.23	.25	.25
DBKGn	26	.50	.50	.42	.42	1162	11,651	.37	.37	.31	.31
DPWGn	22	.64	.46	.46	.50	1207	11,651	.51	.43	.46	.46
DTEGn	27	.67	.67	.70	.70	1362	11,651	.55	.51	.52	.52
EONGn	14	.64	.64	.57	.57	1133	11,651	.64	.68	.71	.70
FMEG	26	.58	.62	.73	.69	1196	11,651	.56	.60	.63	.64
FREG	22	.59	.55	.64	.64	1039	11,651	.57	.56	.58	.59
HEIG	34	.65	.65	.68	.71	1389	11,651	.47	.49	.48	.48
HNKG	28	.64	.50	.50	.46	1339	11,651	.53	.41	.43	.41
IFXGn	16	.69	.81	.81	.81	1057	11,651	.61	.77	.72	.74
LHAG	22	.64	.55	.59	.59	1143	11,651	.65	.59	.50	.52
LING	24	.58	.38	.42	.42	1095	11,651	.60	.50	.50	.50
MRCG	24	.63	.75	.75	.75	1452	11,651	.56	.64	.59	.58
MUVGn	46	.46	.41	.44	.44	1523	11,651	.36	.32	.34	.34
RWEG	21	.76	.57	.52	.62	1235	11,651	.54	.49	.46	.47
SAPG	10	.70	.70	.80	.70	866	11,651	.67	.73	.69	.67
SDFGn	26	.73	.58	.62	.65	1118	11,651	.62	.61	.60	.62
SIEGn	29	.41	.45	.45	.41	1329	11,651	.37	.40	.41	.39
TKAG	22	.50	.68	.73	.68	1211	11,651	.39	.50	.50	.49
VOWG	37	.51	.46	.51	.51	1285	11,651	.49	.44	.42	.42

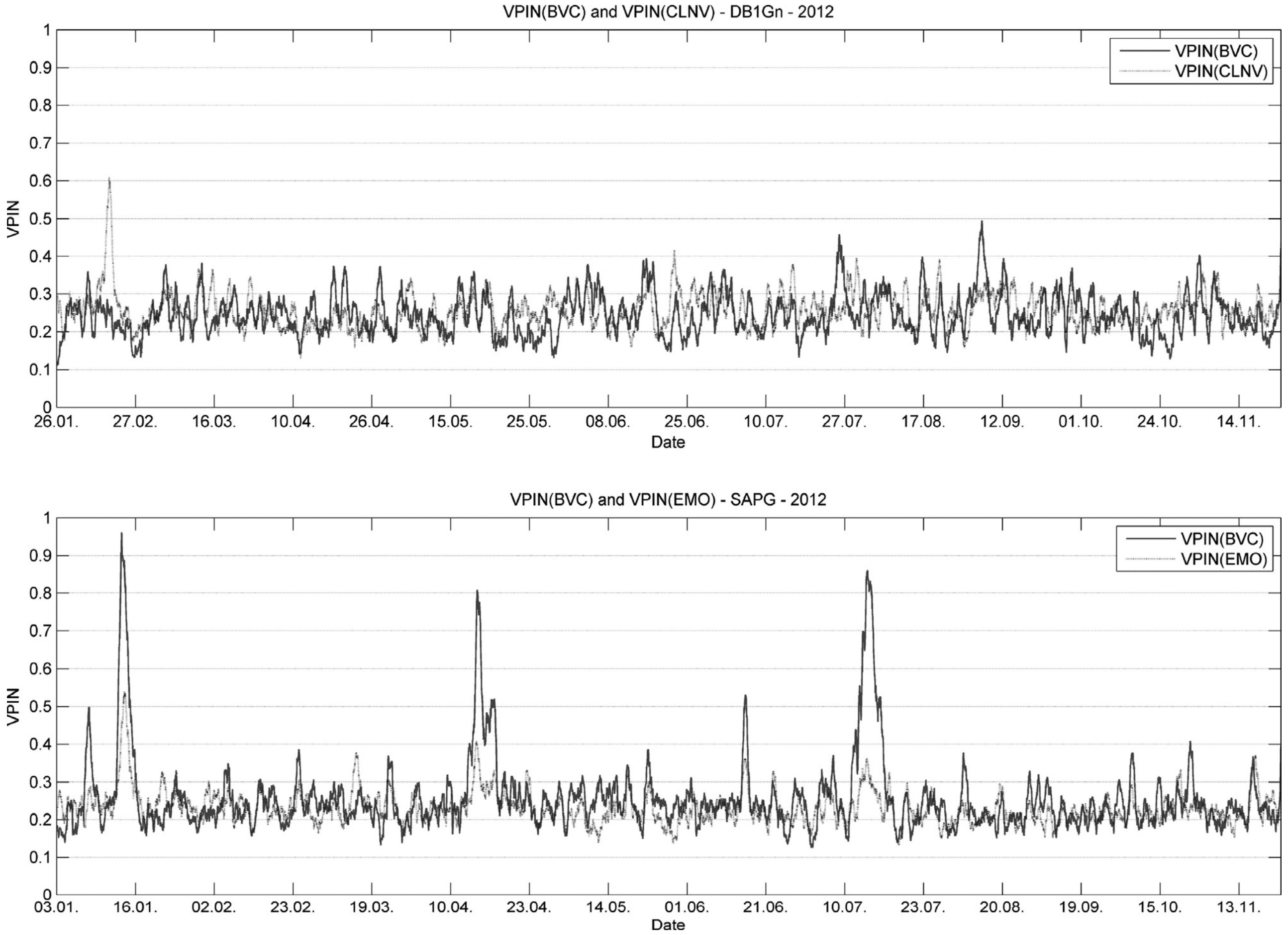

Fig. 4. Toxic period examples Deutsche Boerse and SAP. This figure illustrates the VPIN(BVC) series of the stock with the highest number of toxic periods (Deutsche Boerse—DB1Gn) and the stock with the lowest number of toxic periods (SAP—SAPG). Each chart also displays the corresponding VPIN($\text{Bulk}_{\text{Trad}}$) series of the algorithm with the highest toxic period match. The horizontal axis is in volume time scale.

the results differ significantly between stocks and algorithms. The highest score per stock is observable for Beiersdorf (BEI) in combination with Bulk_{LR} (77.5%); the lowest is found for Deutsche Boerse with Bulk_{LR} (22.5%).

How do the results of the toxic period overlap relate to our previous results of correlation between trade classifications? Fig. 5 shows two VPIN time-series for 1 month, one where 100% of toxic periods and toxic buckets match, and a second where 0% match. In the first graph, the VPIN of Infineon (IFC) in February 2013, all toxic periods within 1 month are identified equally by VPIN based on BVC and trade-by-trade algorithms. The second graph shows the VPIN results for Daimler (DAI) in August 2013. In this case, no toxic periods and no toxic buckets match. The corresponding correlations of the monthly VPIN time-series are 87.3% for Infineon and 22.79% for Daimler. The high correlations between the two approaches used to calculate VPINs would seem to imply correlations on the next higher-order level of toxic periods. Table 10 supports this supposition for the entire sample. The correlation of the VPIN time-series based on different classification algorithms correlates at about an 85% rate, with equal detection of toxic buckets.

One of VPIN's intended practical uses is to operate as an early warning system for toxic order flow. This application motivates our final analyses. For all toxic periods where both approaches overlap, and thus "agree" that there is actually a toxic period, we check which approach is first to detect the rise of toxicity in the order flow. Table 11 presents results by single stocks. Averaged over all stocks, the result is close to a random draw. In 48–51.4% of the cases, depending on which trade classification algorithm is chosen, we find that VPIN calculated with BVC rises first. Also on the single-stock level, this share deviates around a 50–50 split without a clear tendency toward one approach.

Table 10
Correlation between VPIN robustness and toxic period match. This table reports the correlations between the robustness of VPIN to calculations using different algorithms, i.e. Corr(VPIN(BVC), VPIN(BulkTrad)) as in Table 8 on a single stock level, and the numbers from the toxic period and toxic bucket match, respectively, from Table 9.

	$\text{Bulk}_{\text{Tick}}$	Bulk_{LR}	Bulk_{EMO}	$\text{Bulk}_{\text{CLNV}}$
Toxic period match	.699	.683	.704	.676
Toxic bucket match	.858	.873	.829	.842

In summary, the detection of toxic order flow from VPIN($\text{Bulk}_{\text{Trad}}$) and VPIN(BVC) is relatively inconsistent. The choice of classification algorithm influences the detection of toxic periods. For our sample, the different methods do not achieve more than a 60% agreement on either the level of periods identified or on the level of toxic buckets do the different methods achieve an agreement more often than in 60% of the cases in our sample. The choice of a particular trade-by-trade algorithm does not impact these results. Rather, it appears to be a systemic difference between bulk classification and trade-by-trade classification.

We believe, the substantial difference between $\text{Bulk}_{\text{Trad}}$ and BVC in the detection of toxic periods is worrysome for the application

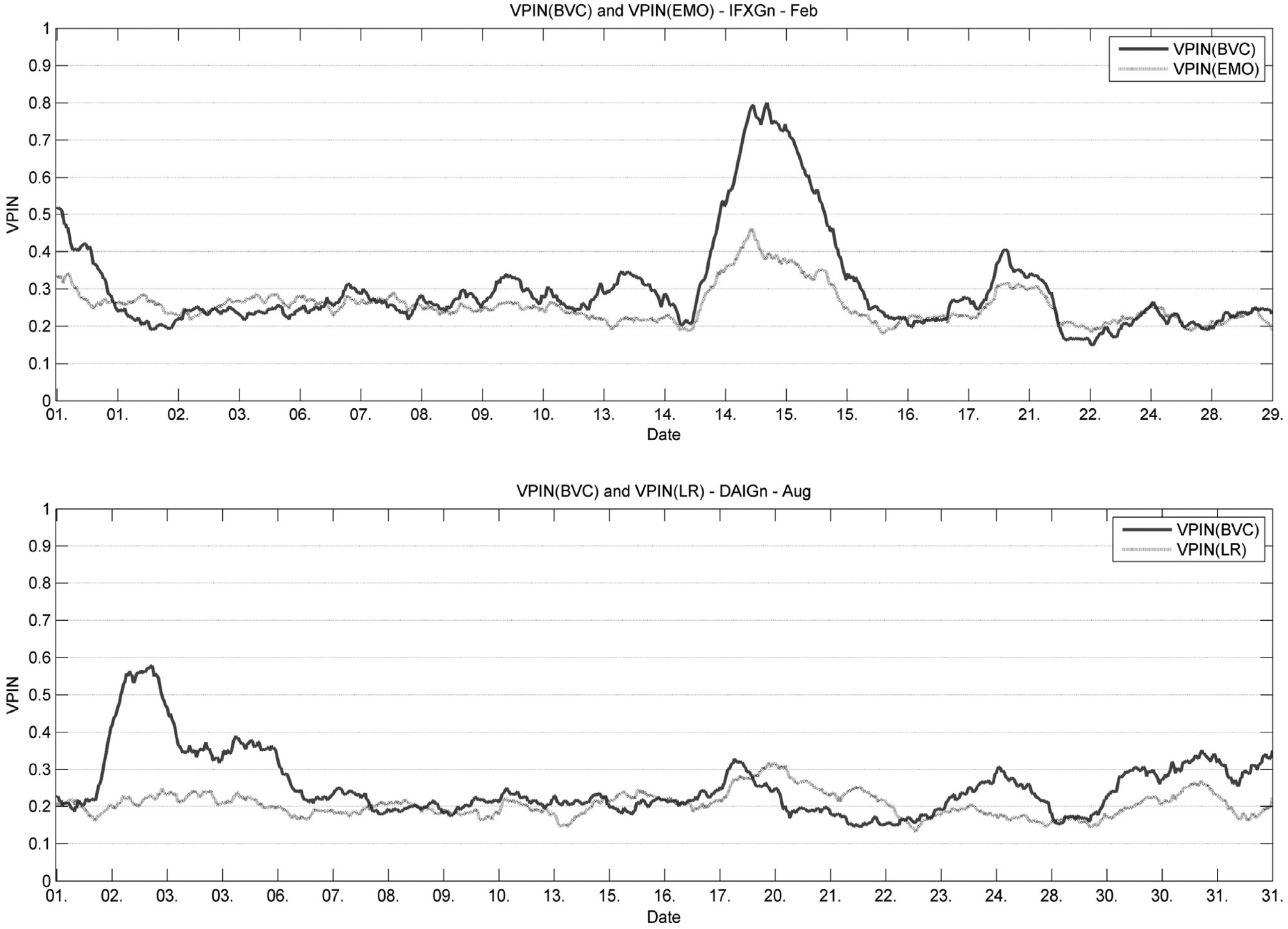

Fig. 5. Extreme toxic period match examples: 100% match for Infineon, 0% for Daimler. This figure illustrates the monthly VPIN series of a 100% toxic period and bucket match—Infineon (IFXGn) with VPIN(Bulk_{EMO}), and a 0% match—Daimler (DAIGn) with VPIN(Bulk_{LR}). For Infineon, two of two toxic periods and 110 of 110 toxic buckets are detected equally. For Daimler, zero of two toxic periods and zero of 137 toxic buckets are detected equally. The horizontal axis is in volume time scale.

of these methods in empirical research. Toward the intended application of VPIN, the BVC approach is also not superior at detecting toxic periods earlier than VPIN calculated with the traditional approach.[9] To clarify where this error comes from, and what parameters that mitigate or amplify the inconsistencies we explore the determinants of VPIN differences in the next section.

5.6. Determinants of VPIN differences

This section uses regression analysis to investigate the reasons for the sensitivity of VPIN to the choice of trade classification algorithm. The dependent variable is the difference in VPIN values computed with bulk classification minus VPIN calculated with trade-by-trade algorithms.

In Table 12, we note first that, apart from the *posReturn* variables, all coefficients are statistically significant in every regression, most at the 1% level. In combination with an R^2 (within) of around 30% we conclude that the hypothesized variables that capture intensity in trading and pricing actually affect differences in VPIN computation to a high degree. Each statistically significant coefficient across both panels points in the same direction. This confirms our prior results that the structural differences between the trade-by-trade approach and BVC are what drives the different VPIN results, not minor variations in the algorithms.

The coefficients' direction and relative influence provide further insight. The coefficient of *Trades* is significantly positive. Given that the total volumes per stock and bucket are fixed, a higher number of trades per VPIN observation, which spans 50 buckets, means that average trade size is smaller, if at least part of the difference in trading activity between stocks is captured by the fixed effect components of the model. Smaller trade sizes may stem from the order slicing and baiting by high frequency traders, hint at uncertainty, given that no one risks larger trades, or indicate the arrival of (uninformed) retail traders. All three possibilities are relevant for the degree of toxicity of order flow, albeit in different directions.

The variable that captures positive returns is insignificant in all models. Large negative returns, however, tend to increase the difference between VPIN computations, although the coefficient is negligible compared to the other variables' standardized coefficients.

Moreover, return volatility and volume volatility strongly increase the differences, as the significant, positive, and large standardized coefficients of *VolumeVola* and especially *ReturnVola* indicate consistently in both panels. In fact, *ReturnVola* exhibits the largest standardized coefficient in every model, implying not only a statistically significant influence, but also the greatest influence of all factors investigated. Similarly, Andersen and Bondarenko (2015) find that VPIN's correlation with future volatility changes its

[9] In another attempt to evaluate VPIN's sensitivity and applicability on a simulated "true positive"-sample, we test whether toxic periods are able to signal the 10 largest 30- or 90-min returns per stock at the start of each of those periods of high volatility. Regardless of the choice of trade classification procedure, only 10–15% of these periods can be identified via a toxic period, with VPIN(BVC) achieving the lowest performance.

Table 11
First detection of toxic periods. This table extends the analysis of the toxic periods match in Table 9 to check which VPIN calculation picks up toxic periods first. Only toxic periods where BVC and Bulk$_{\text{Trad}}$ can detect a toxic period equally are included. For each trade-by-trade algorithm, the share of periods where the rise to a toxic level of VPIN is detected first is reported in comparison to BVC. For each stock and trade-by-trade algorithm, the remaining difference to 100% is made up from the share of simultaneous starts of toxic periods. The totals for 1,714 equally identified toxic periods are 50% first detection for BVC, 46.4% for Bulk$_{\text{Trad}}$ and 3.6% simultaneously detected toxic periods.

	Tick Rule			Lee–Ready			EMO			CLNV		
	# over-lap tox. per.	% First detected		# over-lap tox. per.	% First detected		# over-lap tox. per.	% First detected		# over-lap tox. per.	% First detected	
		Bulk Tick	BVC		Bulk LR	BVC		Bulk EMO	BVC		Bulk CLNV	BVC
Total	431	.478	.480	417	.453	.504	430	.460	.502	436	.461	.514
ADSGn	17	.294	.706	18	.333	.611	16	.188	.750	18	.278	.611
ALVG	16	.500	.500	14	.357	.571	19	.474	.474	18	.444	.556
BASFn	26	.538	.423	18	.278	.556	20	.300	.650	21	.286	.714
BAYGn	13	.538	.308	19	.474	.526	17	.353	.471	19	.316	.632
BEIG	12	.500	.417	13	.692	.308	13	.692	.308	13	.538	.462
BMWG	19	.474	.421	18	.278	.667	20	.300	.700	19	.316	.684
CBKG	15	.533	.467	17	.647	.294	15	.533	.400	16	.563	.375
DAIGn	13	.462	.538	15	.333	.600	13	.308	.615	14	.429	.571
DB1Gn	19	.368	.632	21	.524	.429	22	.545	.455	23	.522	.435
DBKGn	13	.308	.538	13	.385	.538	11	.545	.455	11	.545	.455
DPWGn	14	.286	.643	10	.300	.700	10	.400	.600	11	.364	.636
DTEGn	18	.222	.722	18	.222	.778	19	.368	.632	19	.368	.579
EONGn	9	.444	.556	9	.333	.667	8	.500	.500	8	.500	.500
FMEG	15	.533	.467	16	.563	.375	19	.632	.368	18	.611	.389
FREG	13	.385	.615	12	.583	.417	14	.429	.500	14	.500	.500
HEIG	22	.455	.455	22	.500	.455	23	.478	.391	24	.458	.542
HNKG	18	.444	.444	14	.143	.857	14	.143	.786	13	.154	.769
IFXGn	11	.636	.364	13	.615	.308	13	.615	.308	13	.692	.231
LHAG	14	.429	.571	12	.417	.583	13	.308	.692	13	.308	.692
LING	14	.786	.214	9	.778	.222	10	.700	.300	10	.700	.300
MRCG	15	.400	.533	18	.389	.556	18	.389	.611	18	.389	.611
MUVGn	21	.714	.286	19	.526	.474	20	.500	.500	20	.500	.500
RWEG	16	.375	.625	12	.500	.417	11	.636	.364	13	.538	.462
SAPG	7	.714	.286	7	.714	.286	8	.625	.375	7	.714	.286
SDFGn	19	.579	.316	15	.733	.267	16	.563	.438	17	.588	.353
SIEGn	12	.667	.333	13	.462	.538	13	.462	.538	12	.500	.500
TKAG	11	.273	.636	15	.267	.533	16	.438	.438	15	.400	.467
VOWG	19	.579	.421	17	.588	.412	19	.684	.316	19	.684	.316

Table 12
Determinants of differences in VPIN results. This table reports the standardized coefficients and absolute t-statistics in parentheses from the fixed effects regression model of Eq. (11). The dependent variable is the signed difference between VPINs calculated with BVC and with Bulk$_{\text{Trad}}$. Panel A uses every 50th VPIN observation, while panel B uses all observations, and then identifies statistical significance with Newey–West standard errors with 49 lags. The explanatory variables relate to the time frame of each included VPIN calculation, i.e. 50 buckets. *Trades* is the total number of trades, and *ReturnVola* is the volatility of all timebar-to-timebar returns within each VPIN estimation. *VolumeVola* represents the volatility of traded volume, *posReturn* and *negReturn* are the stock returns split into two variables to capture rising and falling market conditions separately. The variable *Timebars* contains the number of time bars during each VPIN estimation as a standardized (inverted) measure of abnormal volume. Asterisks $**$ denotes a 1% significance level and $*$ denotes a 5% significance level.

	Tick	LR	EMO	CLNV
Panel A: signed VPIN delta, spaced regression; n = 6552				
Trades	.439 (11.5)**	.311 (7.9)**	.284 (7.2)**	.282 (7.2)**
ReturnVola	.515 (28.2)**	.456 (24.2)**	.486 (25.7)**	.476 (25.3)**
VolumeVola	.171 (12.9)**	.147 (10.8)**	.153 (11.2)**	.153 (11.2)**
posReturn	−.006 (0.6)	.006 (0.5)	.000 (0.0)	.002 (0.1)
negReturn	−.010 (0.9)	−.030 (2.5)*	−.019 (1.5)	−.021 (1.7)
Timebars	−.314 (28.0)**	−.315 (27.3)**	−.326 (28.1)**	−.324 (28.0)**
R^2 (within)	.318	.288	.305	.301
Panel B: signed VPIN delta, overlapping regression; n = 326,228				
Trades	.436 (11.4)**	.298 (8.1)**	.273 (7.3)**	.269 (7.3)**
ReturnVola	.458 (8.8)**	.404 (8.3)**	.431 (8.4)**	.423 (8.4)**
VolumeVola	.149 (2.3)*	.129 (2.1)*	.133 (2.1)*	.132 (2.1)*
posReturn	.011 (1.1)	.018 (1.7)	.011 (1.0)	.012 (1.2)
negReturn	−.024 (2.5)*	−.041 (4.3)**	−.030 (3.1)**	−.032 (3.4)**
Timebars	−.328 (20.2)**	−.326 (21.2)**	−.339 (21.3)**	−.337 (21.3)**
R^2 (within)	.311	.281	.298	.295

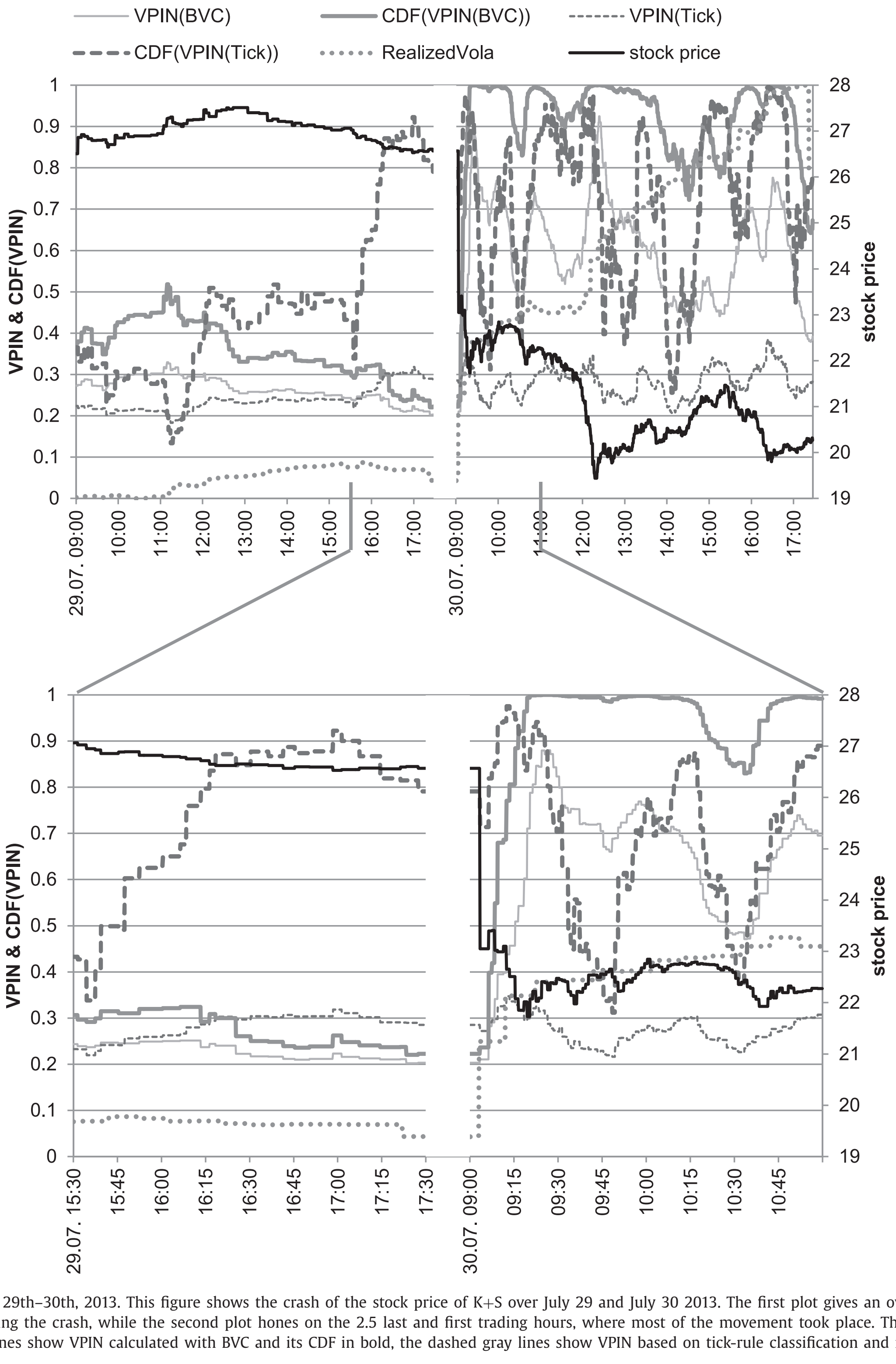

Fig. 6. K+S crash July 29th–30th, 2013. This figure shows the crash of the stock price of K+S over July 29 and July 30 2013. The first plot gives an overview of the two full trading days surrounding the crash, while the second plot hones on the 2.5 last and first trading hours, where most of the movement took place. The black line shows the stock price, the gray lines show VPIN calculated with BVC and its CDF in bold, the dashed gray lines show VPIN based on tick-rule classification and its CDF in bold.

sign depending on the trade classification scheme.[10] However, especially during highly volatile and potentially toxic periods, VPIN should yield reliable results, and not depend on the underlying algorithm. The variable *Timebars* confirms the observations from *ReturnVola* and *VolumeVola*. The more time bars a single VPIN computation spans, the less intense trading is at the time, because it takes more time bars to fill the bucket with the required volume. Hence, a large and negative coefficient indicates that VPIN estimations diverge less and are more robust during times of abnormally low volume, although rather the opposite market conditions require a robust estimation of VPIN.

Ultimately, an increasingly positive difference means that VPIN(BVC) increases faster, or exhibits more volatility, than

[10] "In contrast, the VPIN metric based on the actual order imbalance is negatively correlated with future volatility" (Andersen and Bondarenko, 2015).

Table 13

Forecast regression. This table shows the coefficients and t-statistics in parentheses of the fixed effects panel regressions. The dependent variable is realized volatility (rv), measured in 10-min intervals as the average of absolute 10-min returns over 51 intervals, i.e. the time span of one trading day. The independent variables, listed in columns, are the latest VPIN observations available at each 10-min interval, based on the different trade classification algorithms, and realized volatility. All independent variables are lagged by 51 intervals, i.e., the length of one trading day. The last column shows the within-R^2. Significance is based on Newey–West standard errors with 51 lags. $***$, $**$, and $*$ denote a 0.1%, 1% and 5% significance levels, respectively.

	Lagged independent variables							
Reg	BVC	Tick	LR	EMO	CLNV	rv	Const	R^2
(1)	0.00137*** (15.4)						0.0010*** (33.9)	.03
(2)			0.00023* (2.1)				0.0013*** (35.5)	.00
(3)			0.00004 (0.3)				0.0014*** (37.3)	.00
(4)				−0.00001 (−0.1)			0.0014*** (37.6)	.00
(5)					−0.00002 (−0.2)		0.0014*** (37.7)	.00
(6)	−0.00059*** (−8.4)					0.522*** (37.4)	0.0008*** (30.7)	.24
(7)		−0.00049*** (−5.5)				0.492*** (38.2)	0.0008*** (25.7)	.24
(8)			−0.00058*** (−6.7)			0.492*** (38.6)	0.0009*** (25.9)	.24
(9)				−0.00063*** (−7.1)		0.492*** (38.6)	0.0009*** (26.4)	.24
(10)					−0.00064*** (−7.2)	0.492*** (38.6)	0.0009*** (26.4)	.24
(11)						0.464*** (34.1)	0.0007*** (28.5)	.23

VPIN(Trad)This is assuming that both react in roughly the same direction, which the correlation values support. Such a characteristic of VPIN(BVC) is welcome, but only if a spike denotes no false positive alert. The next two sub sections provide evidence in this direction.

5.7. Predictive power of different VPIN implementations

We could interpret the first five models in Table 13 as a weak indication that VPIN calculated with BVC captures something besides pure trade classification. The regression on realized volatility with lagged VPIN(BVC) is the only of those first five models where VPIN's coefficient is significant and R^2 is not zero. We use the within-R^2 from the fixed regression estimation, which basically detrends the involved time series as a way to correct for the explanatory power that stems solely from the fixed effect dummy variables. However, the only way to fully explore whether VPIN(BVC) has any incremental explanatory power over realized volatility itself, with which it is highly correlated by the design of BVC using monotone transformation price changes, is to add lagged realized volatility to the regression.

This completely upends the previous evidence. VPIN calculated with BVC does not add explanatory power: The R^2 of VPIN(BVC) in regression (6) is only one percentage point larger than that of model (11), which exhibits only lagged realized volatility as an explanatory variable. The regressions using VPIN(Trad) as explanatory variables besides lagged realized volatility achieve an R^2 identical to VPIN(BVC). Furthermore, the relationship between VPIN(BVC) and future volatility now turns significantly negative, as Andersen and Bondarenko (2015) also identify.

But this paper is motivated by the sensitivity and robustness of VPIN to different choices of trade classification, not VPIN's explanatory power. The preceding analysis supports concerns around VPIN's predictive power for equity markets as well, which so far had only been established on futures markets (Andersen and Bondarenko, 2014b, 2015).

5.8. The crash of K+S

On July 30, 2013, Germany experienced a rare instance of a significant crash, roughly similar in magnitude to the "flash crash" analyzed in Easley et al. (2012b). The stock of K+S, one of the DAX 30, crashed from 26.54 EUR at closing on July 29, to 20.24 EUR at closing on July 30,—a 24% loss in less than one day. The majority of the crash occurred in the first minutes and hours of trading.

Fig. 6 illustrates the 2 days surrounding the crash in clock time. The stock price (solid black line) begins to drop slowly during the afternoon on July 29, and then crashes during the first 10–15 mins of trading the next morning. Another smaller crash happens right after noon on July 30. The thin light gray lines illustrate that VPIN calculated with BVC (solid) and with the tick-rule (dashed) both shot up steeply as soon as the price dropped on July 30, but no obvious warning reaction was present before the crash.

However, this changes when we examine the CDFs. The CDF(VPIN(tick)) calculated with trade-by-trade classification (dark gray dashed line) had already begun rising dramatically during the late afternoon of July 29. Note that, from around 3:30 pm, it climbed from 0.35 upward, passing the "critical level" of 0.9 (Easley et al., 2011) and turning "toxic" at 4:58 pm. Thus, traders could have been warned the day before the actual crash. The CDF of VPIN(BVC), the solid gray line, on the other hand, decreases during the last hours of trading on July 29, as did realized volatility, the dotted dark gray line, which was added as a control in the previous section.[11]

Hence, a VPIN calculated using traditional trade-by-trade algorithms would have signaled the crash, while the VPIN calculated with bulk volume classification would not have. This result is exactly opposite from what Easley et al. (2012b) find. In their analysis of the S&P 500 flash crash VPIN computed with BVC signalled the crash, but VPIN calculated with trade-by-trade

[11] For better display in the graph, realized volatility is standardized to an interval between 0 and 1.

algorithms did not. Admittedly, this is only a single case, but this is by far the most severe crash of a German blue-chip stock in either 2012 or 2013. We believe this evidence cannot dispel doubts about the robustness of the VPIN metric.

6. Conclusion

This paper aims to understand the sensitivity of VPIN to the choice of trade classification algorithm. In light of high frequency market environments, Easley et al. (2012b) propose a heuristic approach to trade classification for calculating VPIN. Traditional trade-by-trade classification algorithms, however, have been evaluated on most financial markets of interest and literature reviews and a new evaluation of proprietary signed trading data shows they perform reasonably well, with accuracy rates of up to 90%.

We first examine whether a simple heuristic can beat a 90% accurate algorithmic approach. Second, if both approaches yield results with similar performance, we explore whether VPIN calculations provide similar results and are robust to the choice of (proper) trade classification. To address these questions, we analyze comparisons of deterministic and heuristic approaches in the application of VPINs to detect toxic order.

We find that VPIN is not robust to the choice of classification algorithm and that the bias increases with the level of trading volatility. On every level ascending in complexity and aggregation, from raw trade classification, to order imbalance, to VPINs and then toxic periods, the choice of trade classification algorithm induces substantial differences in the results. The gap widens for higher aggregate metrics, instead of the more likely result of diminishing differences, once we compute aggregates. And we find that the tick rule produces VPIN results closest to BVC, despite being the least accurate algorithm on a trade-by-trade basis. In the detection of toxic periods, the major proposed application of VPIN, both approache gives consistent results more than 60% of the time for our sample. Moreover, neither approach is consistently faster or earlier at detecting toxic periods.

Regression analysis provides hints as to what trading environments VPIN should be used in with the most caution. Unfortunately, they are the times of highest return and volume volatility, which are exactly the conditions under which the application of VPIN is intended to be most useful. One promising way to extend this analysis would be to evaluate the detection of true positives, as well as false positives, by means of company-specific or market-wide information events. We provide an example in this direction by analyzing one of the most extreme crashes of a blue-chip stock that occurred in Germany within the last 2 years—the 24% crash of K+S on July 30, 2013. In this case, VPIN calculated using the tick rule was able to signal the crash, but VPIN(BVC) was not. This is only one suggestive episode, of course, so it cannot fully justify the given conclusion, but it indicates the agenda to follow for exploiting a VPIN-style metric as a tool for identifiying toxic events. Overall, our results suggest that VPIN is not superior to existing measures, and not reliable enough for its proposed application.

References

Abad, D., Yagüe, J., 2012. From PIN to VPIN: An introduction to order flow toxicity. The Spanish Review of Financial Economics.

Aitken, M., Frino, A., 1996. The accuracy of the tick test: Evidence from the Australian stock exchange. Journal of Banking & Finance 20 (10) 17151729.

Andersen, T.G., Bondarenko, O., 2013. Comments on “Testing VPIN on Big Data– response to reflecting on the VPIN dispute". Available at SSRN *2331106*.

Andersen, T.G., Bondarenko, O., 2014a. Reflecting on the VPIN dispute. Journal of Financial Markets 17, 53–64.

Andersen, T.G., Bondarenko, O., 2014b. VPIN and the flash crash. Journal of Financial Markets 17, 1–46.

Andersen, T.G., Bondarenko, O., 2015. Assessing Measures of Order Flow Toxicity and Early Warning Signals for Market Turbulence. Review of Finance 19 (1), 1–54.

Chakrabarty, B., Li, B., Nguyen, V., Van Ness, R.A, 2007. Trade classification algorithms for electronic communications network trades. Journal of Banking & Finance 31 (12), 3806–3821.

Chakrabarty, B., Pascual Gascó, R, Shkilko, A, 2012. Trade classification algorithms: a horse race between the bulk-based and the tick-based rules. Available at SSRN 2182819.

Easley, D., de Prado, M.L., O'Hara, M., 2016. Discerning information from trade data. Journal of Financial Economics 120 (2), 269–285.

Easley, D., Engle, R.F., O'Hara, M., Wu, L., 2008. Time-Varying Arrival Rates of Informed and Uninformed Trades. Journal of Financial Econometrics 6 (2), 171–207.

Easley, D., Kiefer, N.M., O'Hara, M., Paperman, J.B., 1996. Liquidity, information, and infrequently traded stocks. The Journal of Finance 51 (4), 1405–1436.

Easley, D., Lopez de Prado, M., O'Hara, M., 2011. 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., Lopez de Prado, M., O'Hara, M., 2012a. Bulk classification of trading activity. Johnson School Research Paper Series (8–2012).

Easley, D., Lopez de Prado, M., O'Hara, M., 2012b. Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies 25 (5), 1457–1493.

Easley, D., Lopez de Prado, M., O'Hara, M., 2014. VPIN and the Flash Crash: A rejoinder. Journal of Financial Markets 17, 47–52.

Easley, D., O'Hara, M., 1987. Price, trade size, and information in securities markets. Journal of Financial Economics 19 (1), 69–90.

Ellis, K., Michaely, R., O'Hara, M., 2000. The accuracy of trade classification rules: Evidence from Nasdaq. Journal of Financial and Quantitative Analysis 35 (4), 529–551.

Finucane, T.J., 2000. A direct test of methods for inferring trade direction from intra-day data. Journal of Financial and Quantitative Analysis 35 (4), 553–576.

Harris, L., 1989. A day-end transaction price anomaly. Journal of Financial and Quantitative Analysis 24 (1), 29–45.

Holden, C.W., Jacobsen, S., 2014. Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions. The Journal of Finance 69 (4), 1747–1785.

Lee, C., Radhakrishna, B., 2000. Inferring investor behavior: Evidence from TORQ data. Journal of Financial Markets 3, 83–111.

Lee, C., Ready, M.J., 1991. Inferring trade direction from intraday data. The Journal of Finance 46 (2), 733–746.

Odders-White, E.R., 2000. On the occurrence and consequences of inaccurate trade classification. Journal of Financial Markets 3 (3), 259–286.

Peterson, M., Sirri, E., 2003. Evaluation of the biases in execution cost estimation using trade and quote data. Journal of Financial Markets 6 (3), 259–280.

Savickas, R., Wilson, A.J., 2003. On inferring the direction of option trades. Journal of Financial and Quantitative Analysis 38 (4), 881–902.

Tanggaard, C., 2004. Errors in trade classification: Consequences and remedies. Available at SSRN 420680.

Theissen, E., 2001. A test of the accuracy of the Lee/Ready trade classification algorithm. Journal of International Financial Markets, Institutions & Money 11 (2), 147–165.

Wu, K., Bethel, E.W., Gu, M., Leinweber, D., Ruebel, O., 2013a. A Big Data Approach to Analyzing Market Volatility. Available at SSRN *2274991*.

Wu, K., Bethel, E.W., Gu, M., Leinweber, D., Ruebel, O., 2013b. Testing VPIN on Big Data – response to “Reflecting on the VPIN dispute”. Available at SSRN *2318259*.