Journal of Banking and Finance 103 (2019) 113–129

Contents lists available at ScienceDirect

Journal of Banking and Finance

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

Bulk volume classification and information detection[☆]

Marios A. Panayides[a], Thomas D. Shohfi[b,*], Jared D. Smith[c]

[a] University of Cyprus, Nicosia, Cyprus
[b] Rensselaer Polytechnic Institute, Troy, NY, United States
[c] North Carolina State University, Raleigh, NC, United States

ARTICLE INFO

Article history:
Received 13 February 2017
Accepted 5 April 2019
Available online 13 April 2019

JEL classification:
C55
G11
G12
G14
G18

Keywords:
Classification algorithms
Bulk volume
Informed trading strategies

ABSTRACT

Using European stock data from two different venues and time periods for which we can identify each trade's aggressor, we test the performance of the bulk volume classification (Easley et al. (2016); BVC) algorithm. BVC is data efficient, but may identify trade aggressors less accurately than "bulk" versions of traditional trade-level algorithms. BVC-estimated trade flow is the only algorithm related to proxies of informed trading, however. This is because traditional algorithms are designed to find individual trade aggressors, but we find that trade aggressor no longer captures information. Finally, we find that after calibrating BVC to trading characteristics in out-of-sample data, it is better able to detect information and to identify trade aggressors. In the new era of fast trading, sophisticated investors, and smart order execution, BVC appears to be the most versatile algorithm.

1. Introduction

Markets across the world have experienced dramatic change in the past fifteen years. The proliferation of high-speed computers and the declining cost of trading in electronic limit order markets have produced an explosion in trading volume and speed (Jain, 2005; Hendershott and Moulton, 2011). The rapid growth of algorithmic low-latency trading, including high frequency trading (HFT), has called into question the efficacy and the relevance of traditional methods to identify the aggressor side of each trade (Holden and Jacobsen, 2014). O'Hara (2015) notes that common features in today's markets—for example, smart execution algorithms that use limit orders, microsecond trading frequencies, and quote volatility—impair traditional individual trade-level classification algorithms from uncovering the trading intentions underlying the orders. Detecting information from trades is important since such trading underpins much of the theoretical work on trading and price formation (see Kyle, 1985; Glosten and Milgrom, 1985) and impacts empirical work investigating toxic order flow (Pöppe et al., 2016; Easley et al., 2012). It also helps researchers and regulators understand and prevent extreme volatility events like the "Flash Crash" (see Kirilenko et al., 2017; Easley et al., 2011). Though information detection is of primary importance, aggressor-signing can be useful as well, for characterizing investor clientele behavior and assessing trading costs, and an ideal algorithm should do both.

This paper helps address these issues by examining the newly developed bulk volume classification algorithm (Easley et al., 2016); hereafter BVC) in modern, low-latency equity markets. BVC uses total volume and price changes within a block of trades to classify order flow into buying and selling volume. We assess BVC performance in terms of the accuracy in finding trade aggressors and ability to capture informative trade flow. To help us calibrate the algorithm, we compare BVC's performance to bulk versions of traditional trade-level algorithms, bulk tick test (Smidt, 1985; Holthausen et al., 1987), and the Lee and Ready (1991) algorithm (hereafter LR). We find that BVC can identify trade aggressors as

[☆] We thank the editor (Geert Bekaert) and three anonymous referees for helpful comments and feedback. We also thank Akiko Watanabe, Steve Mann, Ryan Davies, members of Citadel's Global Equities group, members of BlackRock's Scientific Active Equities group, along with participants at the 2014 Northern Finance Association, 2014 Southern Finance Association, and 2015 Financial Management Association annual meetings for providing useful comments. We thank the Don Shohfi Financial Research Fund for computing resources. Earlier version of this manuscript was circulated under the title, "Comparing Trade Flow Classification Algorithms in the Electronic Era: The Good, the Bad, and the Uninformative."

[*] Corresponding author.

E-mail addresses: panayides.marios@ucy.ac.cy (M.A. Panayides), shohft@rpi.edu (T.D. Shohfi), jared_smith@ncsu.edu (J.D. Smith).

https://doi.org/10.1016/j.jbankfin.2019.04.001
0378-4266/© 2019 Elsevier B.V. All rights reserved.

well as traditional algorithms, and its signed order flow is the only measure that is reliably related to different illiquidity measures shown to capture the trading intensions of informed traders (e.g., Easley et al., 2016).

We use equities data from NYSE Euronext for 2007 and 2008 and from the London Stock Exchange for 2017 to perform our analyses. These datasets have several advantages that make them ideal for our study. First, we have two distinct periods to test the effects of low-latency trading on the performance of the different classification algorithms. The early data contain second-level timestamps, while the later data use microseconds, allowing us to examine the benefit of more granular data. In addition, European markets did not fragment as rapidly as U.S. equity markets (see Fig. 1 of Menkveld, 2013). The low level of fragmentation, combined with rich datasets that allow us to identify the aggressor side of 97.9% of the trades in our samples, means that our study uses nearly all trading activity in the dominant trading venues to conduct our analyses. This is particularly important for testing which classification algorithms can capture the trading intentions of informed traders.

We begin our analysis by investigating how well BVC classifies trade aggressors in our samples. BVC involves putting trades into blocks, or bars, by either volume or time. A percentage of the block is then classified as buys (the remainder as sells) based upon the movement of prices around the bars. By construction, the BVC algorithm is highly data efficient as it uses aggregate bar-size trading volume and prices, which translates to less than 1% of the trade data points. The Euronext sample results on finding trade aggressors comport with those in Chakrabarty et al. (2015) and Easley et al. (2016); BVC is not as accurate as bulk versions of traditional trade-level algorithms. This reverses, however, in our more recent 2017 sample. Indeed, even though all three algorithms perform worse in the LSE sample, BVC is the most accurate. This suggests that more granular data does not save traditional algorithms from broad shifts in speed, volume, and market structure. According to Nanex, peak intra-second quote rates have increased 100 times since 2007, a growth rate far exceeding that of computing power.[1] Several studies suggest this increased quote activity might not be informative to trade-level algorithms. For example, Baruch and Glosten (2016) find quote randomization optimally manages risk from predatory trading; Hasbrouck (2018) finds that HFT-driven quote volatility degrades information within quotes.

In our next set of analyses, we examine whether BVC can effectively uncover underlying trading intentions in order flow. In today's markets, researchers and practitioners are increasingly interested in identifying buying or selling pressure that can be destabilizing and/or toxic. Information-related order flow will unavoidably disadvantage other traders (retail traders and some institutional traders; O'Hara, 2015). To test information in BVC order flow, we run spread regressions and an event study analysis focusing on return predictability from order flow in the pre-event period. We measure spreads in two ways, using the Corwin and Schultz (2012) high-low spread and calculating intraday effective spreads. Easley et al. (2016); hereafter ELO use the first approach, which they argue is best for BVC, because it removes underlying asset volatility. We then regress our spread measures on the buy/sell order imbalance with stock and month fixed effects. If an algorithm successfully estimates the underlying informed order flow, then a larger algorithm-estimated order imbalance should be directly associated with a larger spread in a given bar.

We find that BVC-estimated order imbalance is positively related to the spread measures for nearly all bar sizes in various samples. When estimated in a pre-event window, it is also directly related to event period returns in a sample of various corporate events (e.g., earnings releases, buyback and M&A announcements). Bulk tick test order imbalance is negatively related to the spread measures, opposite to what standard microstructure theory would suggest, and it is unrelated to event returns. Thus, while BVC imbalance relates to the spread in a direction we expect, bulk tick imbalance does not.

This is not a failure of the tick test, however. When we run our spread regressions using the actual trade aggressor, known from our order data in both samples, aggressor imbalance is often negatively related to the spread measures and it is unrelated to event returns in a multivariate analysis. These results indicate that trade aggressor identification does not convey underlying information in today's fast markets, where sophisticated traders use smart algorithmic trading to hide their trading intentions and minimize market impact. For example, informed traders are increasingly relying on passive orders, i.e., limit orders, to disguise themselves in the market (Bouchaud et al., 2009; Menkhoff et al., 2010; Zhang, 2013). Therefore, any traditional trade-level algorithm designed to find a trade's aggressor will not tell us much about information, no matter how accurate. By contrast, BVC order imbalance is defined in a way that captures the resulting (equilibrium) price impact of order flow at the end of a bulk period, which seems better able to capture information.

Lastly, we explore how changes to BVC's implementation affect its performance. Though most often researchers focus on detecting information, trade aggressors can also be important for describing the trading behaviors of various investor groups (such as individual investors, or short-sellers), for measuring the impact of maker-taker fees (Battalio et al., 2016), and for assessing trading costs associated with market anomalies (Novy-Marx and Velikov, 2016). To begin our BVC calibration, we examine heterogeneity in bar size "fill rates," showing how differently identical bar sizes will function across stocks. We also find a systematic bias when trades are truncated to make volume bar sizes exact, a bias which can be as large as 44%. By using flexible, minimum volume bar sizes, rather than exact ones, this bias is eliminated. In our suggested calibration approach, we iterate across a set of bar sizes to select a bar size large enough to limit excess kurtosis (and better fit the BVC's assumed t-distribution of price changes), but small enough to produce enough data points in any sample-month (and thus have meaningful variation in price). Relative to a set of randomly chosen BVC bar sizes, our calibrated bars better find aggressors and successfully detect information in both times series regression and event study settings.

This paper contributes to the nascent literature on trade classification algorithms (including BVC) and low-latency trading, as well as adding to the long list of papers investigating the performance of the LR and tick test algorithms. While ELO use futures data to investigate BVC, our study examines equities markets using recent, rich data from two prominent and relatively concentrated venues. Using equities entails an additional challenge, which requires security level calibration, because "the optimal interval [bar size] is unlikely to be uniform across stocks with disparate trading activity" (ELO (2016) pg. 35). This is important because of institutional differences, clientele effects, and differences in investors' trading behavior (especially for informed traders) that exist between futures and equity markets. For example, there is evidence that block purchases and sales have a differential price impact in equity markets but not the futures market (see, for example, Chan and Lakonishok, 1993; Berkman et al., 2005).

Chakrabarty et al. (2015) also examine BVC in an equities sample, using NASDAQ data from 2011. They find that BVC cannot classify trade aggressors as accurately as bulk versions of tick and LR. Further, they find that BVC does worse in measuring the order flow

[1] See "HFT is Out of Control" at http://www.nanex.net/research/MsgRates/EquityMessageRates.html.

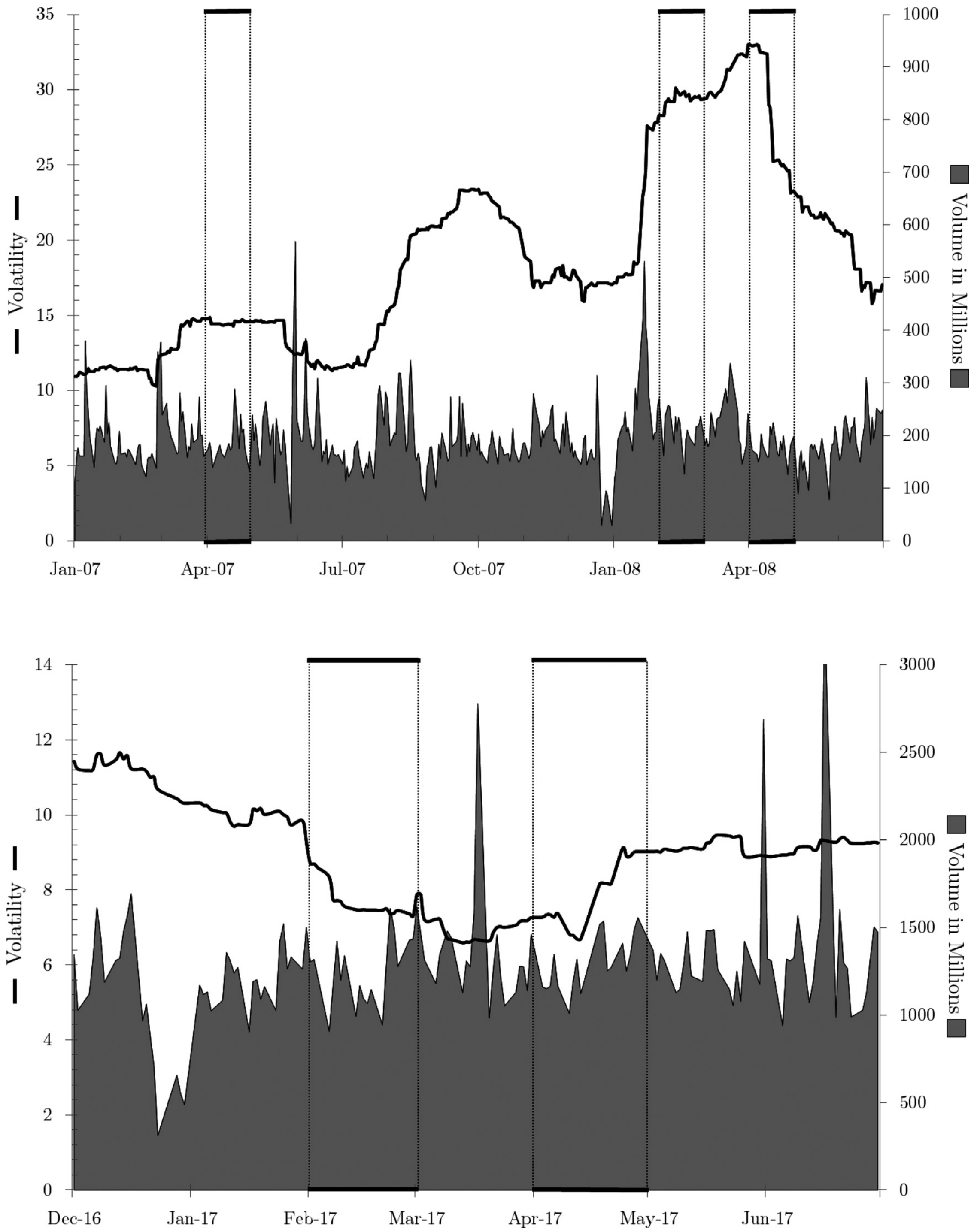


Fig. 1. Graph of volatility and volume of the CAC-250 Index (FTSE-350 Index) from January 2007 through June 2008 (December 2016 through June 2017) in the first (second) row. 60 day average volatility of the index is represented by the black line and measured by the left vertical axis. Volume of the CAC-250 (FTSE-350) index components, in millions, is represented by the gray area at the bottom of the chart and measured by the right axis. The time periods of our sample, April 2007, February 2008, and April 2008 (February 2017 and April 2017), are highlighted.

of informed traders, where they assume that aggressive order imbalance is informative. In contrast, we explore not only how features of BVC's implementation affect the results generated by it, but also whether aggressive order flow is actually informative. In our data from less fragmented, European equity markets, we find that an out-of-sample calibrated BVC successfully captures the aggressor side of trades and information, and it does so without requiring costly, real-time analysis of low latency individual trade and quote data. This suggests that researchers can use BVC to both classify aggressors and measure information, while capturing the data efficiency gains inherent in BVC.

The remainder of the paper proceeds as follows. Section 2 describes our data in detail. Section 3 describes the methodology used to detect trade aggressors, followed by a review of trade-level and bulk volume classification algorithms. Section 4 presents the main empirical results, including a discussion of calibration and other methodology refinements for improving the accuracy of BVC. Section 5 specifically examines the ability of each algorithm to detect underlying information in trade flow. Finally, Section 6 provides a brief conclusion.

2. Data

2.1. Data sets and samples

We use two complementary data sets in our analysis. The first set comes from NYSE Euronext "NextHistory" files. These data contain all trades and quotes, and nearly all orders submitted (the iceberg orders that did not participate in trades are not included). The data are time-stamped at the second-level. The second set, from the London Stock Exchange (LSE), is built from the "Tick Data" and "Rebuild Order Book" data. The former includes price and volume information, while the latter provides daily order book activity (trades, orders, deletions), which we need to determine individual trade aggressors. These data are time-stamped at the microsecond-level.[2] In both data sets, we focus exclusively on continuously traded equities (i.e., we drop non-equity instruments and equities traded in daily call auctions). Both data sets can be purchased from the respective exchanges.

In total, our Euronext data files span a period of 19 months (Jan 2007-Jul 2008) and cover all stocks traded on Euronext Paris. LSE makes data available back to 1996. To grapple with the size (and cost) of the data we create two tractable samples from these data sets. First, we choose sample periods. From Euronext, we use April 2007, February 2008, and April 2008, because these months represent different periods of volatility, stable-low, stable-high, and dropping periods of volatility, respectively. From LSE, we choose a much more recent sample, February and April 2017. These data are characterized by lower, more stable volatility and much greater trade volume. This is seen clearly in the volatility and volume graphs of the CAC-250 and FTSE-350 indices contained in Fig. 1. Though testing across market-wide volatility conditions is useful, even more importantly, these two samples have different periods of algorithmic trading. The Euronext data represent a nascent period of algorithmic trading and the LSE data a more mature period (Hendershott and Riordan, 2013; Menkveld, 2013; Mahmoodzadeh and Gençay, 2017).

Our sample periods are useful because of the relatively greater consolidation of listed firms' equity trading on Euronext and the LSE. We capture as much as 82.9% of all trading volume on Euronext, and 74.3% of all volume on the LSE for equities listed on each exchange. Comparatively, in 2007, 63.7% of NYSE listed equities volume occurred on the NYSE exchange. This fell even further, so that in 2017 only 30.6% of trading occurred on the NYSE.[3] This also compares favorably to the TotalView-ITCH equities data available from NASDAQ for 2005 and 2011, which has roughly 16% to 26% of total volume. The higher volume share of Euronext and LSE listed stocks make these data ideal for our tests, especially regarding informed order flow.

In addition to choosing sample periods, we choose a sample of stocks to focus on. For Euronext, we take the 469 continuously-traded French stocks common to all three time periods and then form a random, representative sample of 100 stocks. Our Euronext sample is comprised of thirty-four small-cap stocks, 33 mid-cap, and 33 large-cap, which we define as those companies less than €1 billion, more than €1 billion but less than €10 billion, and those more than €10 billion, respectively. For the LSE we construct a slightly larger sample of 125 U.K.-based listed equities across varying size groups (using 2017 euros converted from British pounds). Summary statistics shown in Table 1 indicate that the samples include stocks of varying liquidity and volatility levels among capitalization groups. The list of included companies is provided in the internet appendix.

Consistent with prior literature, our analysis excludes trades in the first 15 min of the daily trading period to avoid opening call auctions in our continuously traded sample (Odders-White, 2000). Since Euronext stocks have closing call auctions, we also exclude trades executed during the last 5 min of the daily trading period. Therefore, our Euronext sample includes only trades executed between 09:15:00 and 17:25:00 CEST.[4] Similarly for LSE, we include only trades executed between continuous trading hours of 08:00:00 and 16:30:00 GMT and ignore trades conducted during the 2-minute midday auction occurring at 12:00:00. We also impose standard trade and quote filters on the data, such as positive price, volume, and quote size, and the bid must be weakly lower than the ask. These filters result in approximately 210 and 335 gigabytes of trade, quote, and order data for Euronext and LSE respectively. In contrast, our constructed time and volume bar data across all bar sizes and both samples use only 13.6 gigabytes, a 97.5% reduction in storage that demonstrates BVC's data efficiency.

2.2. Aggressor side of trades

To identify whether each trade in our sample is a buy or a sell, we follow the definition of trade aggressor/initiator used in Odders-White (2000), which Ellis et al. (2000) note is preferred when a researcher has access to the order book. She defines the trade initiator based on chronological order arrival, that is, the order that arrives second is the order that actually "initiates" the trade. For example, if a market buy order comes in at 11:15AM and hits a limit sell order that had been standing in the book since 11:00AM, that trade would be classified as a buy for our purposes. To determine the trade aggressor in our sample, we first classify fully-executed orders into active and passive categories. An active order is executed at the same date and time as it is submitted to the marketplace, and is, essentially, a market order. A passive order is a non-market order whose execution time is always later than its submission time. In this case, the initiator of a trade will be the *opposite* buy or sell direction of a matching passive order. Active orders account for approximately 98% of trade aggressors identified across our sample. We construct a seven-stage procedure in signing trades, the details of which are available in the internet appendix. Overall, untabulated results suggest that our procedure performs very well at identifying the trade aggressor (97.9% of total trades).

3. Methodology

3.1. The bulk volume classification (BVC) algorithm

3.1.1. Overview

The bulk volume classification procedure was developed in ELO for use in the Easley et al. (2012) volume-synchronized probability of informed trading (VPIN) calculation. It is designed to classify bars of trades (i.e., trades put in blocks either by time or

[2] Technically, to order trades within a millisecond, we use a microsecond-accurate ordering variable provided by LSE.

[3] NYSE listed on-exchange trading as a percentage of overall volume is taken directly from NYSE Euronext 2nd quarter 2007 operating data within the exchange's reported financial results or obtained from the NYSE Market Data website at http://www.nyxdata.com/Data-Products/NYSE-Volume-Summary.

[4] According to Euronext rules, from 07:15 until 09:00, orders accumulate in the order book, at 09:00 orders in the central book are matched and an opening price is set. Stocks are then to trade continuously starting at 09:01 so we are being conservative in deleting the first 15 minutes. This process occurs at the end of the day as well, with orders accumulating in the book starting at 17:25.

Table 1
Sample Summary Statistics.
Panel A displays the mean, standard deviation, median, minimum, and maximum market capitalization for each capitalization group, small, medium, and large, which we define as companies worth less than €1 billion, between €1 billion and €10 billion, and above €10 billion, respectively. All market capitalization numbers are in millions of 2017 euros. These summary statistics are taken over all months of our Euronext (April 2007 and 2008 and February 2008) and LSE (February 2017 and April 2017) subsamples. Panels B, C, and D display analgous statistics for daily traded volume, volume per second (Euronext), and volume per microsecond (LSE) for each capitalization group.

Panel A: Market capitalization (mm 2017 EUR)

	N	Mean	Std Dev	Median	Min	Max
Small Cap	72	244.52	215.89	165.42	6.44	990.90
Mid Cap	84	5,123.78	2,887.29	5,724.16	1,130.34	9,939.69
Large Cap	69	35,482.59	33,384.80	22,496.23	10,788.75	154,698.53
Total	225	12,872.45	23,931.01	5,337.50	6.44	154,698.53

Panel B: Daily share volume

	N	Mean	Std Dev	Median	Min	Max
Small Cap	72	225,472	469,586	33,153	150	2,719,815
Mid Cap	84	2,268,175	3,387,343	923,525	233	19,027,110
Large Cap	69	9,849,100	26,468,510	2,978,475	900	207,374,215
Total	225	3,939,327	15,272,101	750,925	150	207,374,215

Panel C: Volume per second (Euronext subsample)

	N	Mean	Std Dev	Median	Mode	Mode / Total
Small Cap	34	649.40	1,680.61	211	100	5.239%
Mid Cap	33	672.77	1,709.15	250	100	4.757%
Large Cap	33	626.68	1,266.65	265	100	3.465%
Total	100	634.18	1,353.92	261	100	1.852%

Panel D: Volume per microsecond (LSE subsample)

	N	Mean	Std Dev	Median	Mode	Mode / Total
Small Cap	38	11,477.63	24,102.57	885	5,000	18.155%
Mid Cap	51	1,639.33	1,503.97	559	100	3.390%
Large Cap	36	2,624.98	3,183.19	679	100	3.562%
Total	125	4,914.04	14,065.70	635	100	3.406%

volume)[5] as a percentage of buys and sells, rather than classifying each individual trade. It was implemented this way to find large order imbalances, which would point to "flow toxicity." It represents an attractive alternative to traditional classification methods, particularly in situations where a researcher need only know the percentage of buys and sells in the data (rather than the direction of individual trades) such as the calculation of VPIN introduced by Easley et al. (2012). By putting the trades into volume blocks, the algorithm mitigates any impact from order splitting and economizes on the number of data points used for classification. For instance, using time or volume bars in our analysis, the best overall accuracy is achieved using only 0.21% of the individual trade data points. This represents an incredible difference in computing storage resources. Whereas the tick test may take days to run, BVC with a large, appropriately chosen block size can be implemented in a matter of minutes (even the bulk tick test method, discussed later, must run the standard tick rule prior to aggregation).

3.1.2. Implementation

We apply the BVC algorithm using the Perl programming language, directly adapted from the example Python code provided by ELO.[6] The implementation in the Euronext and LSE data is similar, but for LSE the following is at the microsecond, rather than the second, level. First, we aggregate trade data to the second. Bars are filled with consecutive trade seconds until the specified bar size is met or exceeded,[7] then the working bar data is stored in a relational database and construction of the next bar begins if additional trade second data is available.[8] Next, for each stock-month combination, we calculate the volume-weighted standard deviation of price changes between consecutive bars as shown in formula (1).

$$\sigma_{\Delta P_i} = \sqrt{\frac{\sum_{\tau=1}^{n} V_{i,\tau}\left(\Delta P_{i,\tau} - \overline{\Delta P_i}\right)^2}{\sum_{\tau=1}^{n} V_{i,\tau}}} \tag{1}$$

where $V_{i,\ \tau}$ is the actual volume of shares traded of stock-month i during the time or volume bar τ which is decomposed into the buy ($\hat{V}^{Buy}_{i,\tau}$) and sell ($\hat{V}^{Sell}_{i,\tau}$) volume estimate components. $\Delta P_{i,\tau} = P_{i,\tau} - P_{i,\tau-1}$ is the price change between two consecutive bars. With these available data points, we can then use formula (2) of ELO to calculate BVC's buy volume for each bar:

$$\hat{V}^{Buy}_{i,\tau} = V_{i,\tau} \cdot t\left(\frac{P_{i,\tau} - P_{i,\tau-1}}{\sigma_{\Delta P_i}},\ df\right)$$

$$\hat{V}^{Sell}_{i,\tau} = V_{i,\ \tau} - \hat{V}^{Buy}_{i,\tau} = V_{i,\ \tau} \cdot \left[1 - t\left(\frac{P_{i,\tau} - P_{i,\tau-1}}{\sigma_{\Delta P_i}}, df\right)\right] \tag{2}$$

[5] ELO use trade bars in addition to volume and time bars. We do not include trade bars since trade size distributions in the equities market are not concentrated and discrete as in the futures market.

[6] We thank David Easley, Marcos Lopez de Prado, and Maureen O'Hara for making this code available.

[7] Volume bar size can be exceeded if the final added trade second contains more volume than the specified bar size. The benefits of this volume bar construction methodology are further examined in Section 5. Time bars, on the contrary, can never exceed their specified bar size. Only the final volume (time) bar in a stock-month may have lower volume (duration) than the specified bar size since the stock-month may terminate before the final bar is completely filled.

[8] We create bars continuously throughout a stock-month, meaning that if a volume bar is unfilled at the end of continuous trading on 15 April it will continue to fill with trades when continuous trading resumes on 16 April. This does not apply to time bars, which are truncated at the end of the trading day. Further, unlike Chakrabarty et al. (2015), our implementation does not use "fixed" time bar beginning and ending timestamps. Muravyev and Picard (2016) find that algorithmic trader synchronization occurs at round start times. To address this periodicity, we use "dynamic" timestamps for a time bar begin when the first trade occurs and ends with the last trade within the specified bar size. Time bar construction is discussed in greater detail in Section 5.

The price associated with each bar, $P_{i,\tau}$, is the price of the last trade within that particular bar and t is simply the cumulative density function of Student's t-distribution with degrees of freedom. Following ELO, we perform our baseline analysis using 0.25 degrees of freedom.

3.2. Tick test and LR algorithms

3.2.1. Overview

The Lee and Ready (1991) trade classification algorithm is widely used in market microstructure. The algorithm uses the quote rule when trades are not at the midpoint, any trade price above (below) the midquote is a buy (sell), and at the midquote it uses the tick rule. The tick rule compares the current trade price to the previous price. When the price is higher (lower) than the previous price, the trade is classified as a buy (sell). Because it is also common to use the tick rule as a standalone algorithm (which we will refer to as the tick test method), we use the results from both LR and the tick test as comparisons for BVC. Because BVC puts trades into bars and offsets misclassifications, we compute bulk versions of LR and the tick test to better compare across algorithms (Chakrabarty et al., (2015); ELO, 2016).

3.2.2. Implementation

Since the Euronext data do not provide sub-second timestamps, we collapse trades at the second level using volume-weighted average price (VWAP; similar to Boehmer and Kelly, 2009) when implementing trade-level LR and tick.[9] We do this to simplify the trade flow because we cannot observe the exact ordering of the trades within a second in our Euronext data. Although one potential criticism of our Euronext trade level algorithm implementations is this lack of millisecond timestamps (as in daily TAQ), we believe it emphasizes the classification issues facing researchers. Our LSE data exhibit the same clustering, with many trades that occur within the same microsecond, though each trade contains a unique identifier that allows us to sequence the data correctly. This inadequacy of more granular data is also observable in other recent equity data (e.g., Muravyev and Picard, 2016; Conrad et al., 2015). O'Hara (2015) argues that continually seeking a perfect data set is not a solution since the underlying trading intentions of an order can be masked in an environment of fast-paced, fragmented markets.

We also have to establish a prevailing quote to be in force in a given second when there are multiple quotes in that second for the Euronext data. We treat quotes in the same second with the same bid and ask prices but different sizes (approximately 49% of sample quotes) as one quote. For any multiple-quote seconds that remain (approximately 28% of sample quotes), we take the best bid and offer (BBO) for each stock-second. In the rare cases (0.4% of sample quotes, or just less than 130,000 quotes for the Euronext data) in which that process creates a crossed quote (i.e., the offer is less than the bid), we just take a midpoint and set both the bid and ask equal to it.[10] This process establishes a single prevailing BBO quote for each firm-second in the sample, which allows us to sign trades using the LR classification algorithm. For LSE data, we use order data to reconstruct the limit order book at the microsecond timestamp of each trade to obtain the BBO.

Finally, trades are matched to quotes and signed according to the quote and tick rules to implement LR and the tick rule only for the tick test. These LR and tick test classified trades are then matched to sample trades for which a trade aggressor could be established (97.9% of full sample trades). Then, we aggregate our trades into volume and time bars mirroring BVC.

4. Performance

4.1. Accuracy in detecting trade aggressors

First, we examine how well BVC, bulk tick test, and bulk LR do in matching aggressor trade classification. Table 2 displays accuracy rates for volume and time bars in Panels A and B respectively. For both panels, columns 1–3 present results for the Euronext sample, and columns 4–6 for the LSE sample. The BVC algorithm uses Student's t-distribution with 0.25 degrees of freedom.

We first note that the accuracy rates are lower than those reported in ELO, who use futures data; their accuracy rates top 94% versus 89% in our analysis. We expect lower accuracy in equities because of greater heterogeneity in stock trading characteristics as well as because blocks of buys and sells have disparate price impact in equities markets (Chan and Lakonishok, 1993; Berkman et al., 2005). This difference in price impact should lower the overall accuracy rates of BVC because the algorithm uses symmetric distributions (e.g. Student's t-distribution) to estimate buy volume in a bar. For example, consider a 50,000 share volume bar composed of a buy and sell, both of 25,000 shares. Asymmetric price impact suggests there will be a price change, which means BVC will not classify buy volume as 50%, and will thus be expected to have lower accuracy.

Despite this challenge, BVC performs well in our two samples. Regarding the Euronext sample (columns 1–3 of Table 2) BVC accuracy ranges from 62.62% to 87.90% in Panel A with volume bars, and from 58.66% to 89.68% in Panel B with time bars. These peak accuracy rates exceed the rates reported in Chakrabarty et al. (2015) NASDAQ sample. We do find, however, that just like ELO and Chakrabarty et al. (2015), BVC accuracy rates are lower than bulk tick test and bulk LR at all comparable bar sizes. These algorithms' accuracy ratios range from 78.14% to 95.50%, and though their accuracy is similar, we find that bulk tick test is more accurate than LR in every bar size. Given that LR requires quote data, it is dominated in the Euronext sample by the tick test, which is more accurate and more data efficient.

In the LSE sample (columns 4–6) we see some important differences from the Euronext results. First, accuracy rates are lower for BVC, as well as bulk tick test and bulk LR. This change in accuracy is likely a consequence of increasing speed and volume in equities markets highlighted in Table 1, which shows much more volume in the LSE sample, though with a more granular timestamp (microsecond vs. second). In Panel A, BVC accuracy ranges from 53.22% to 75.68% using volume bars; in Panel B, using time bars, the accuracy is much better, ranging from 60.30% to 87.14%. The accuracy of bulk tick test and bulk LR ranges from 39.65% to 83.75% (across both volume and time bars), lower than BVC in all cases. This stark reversal in relative accuracy suggests that increasing market speed makes it increasingly difficult for trade-level algorithms (even when assembled into bars).[11]

[9] Collapsing at the price-second level results in a slight increase in accuracy, but the advantage of VWAP is it does not assume an order for same-second trades that occur at different prices, at the cost of reduced granularity. Unfortunately, because we do not know trade or quote ordering within a second for our Euronext data, we cannot implement the interpolated time method from Holden and Jacobsen (2014) and are forced to aggregate to the second-level.

[10] Dropping trades matched to crossed quotes from the analysis does not significantly impact bulk LR accuracy rates.

[11] BVC accuracy is particularly impressive given that Euronext (LSE) trade level data are compressed by 68.36% to 99.87% (76.83% to 99.89%). Running LR requires quote data, adding almost 19 million Euronext quote and 369 million LSE order observations, vastly increasing the rate of compression BVC offers. Chakrabarty et al. (2015) find that algorithm CPU time between the tick test and BVC are comparable. However, the data efficiency advantage of BVC has benefits in practical implementation. For example, several market data providers can transmit

Table 2
Algorithm Accuracy Comparison.
This table displays the accuracy results from bulk volume classification (BVC), bulk tick test (Bulk Tick), and bulk Lee & Ready algorithms (Bulk LR) in classifying the aggressor side of the trade. Each algorithm is implemented so that the unit of observation is monthly trade data. Panel A displays results for volume bar aggregation and Panel B shows the results for time bar aggregation. Results for the Euronext and London Stock Exchange (LSE) subsamples are shown separately, indicated by column header.

	Euronext (2007–2008) Sample			LSE (2017) Sample		
	BVC	Bulk Tick	Bulk LR	BVC	Bulk Tick	Bulk LR
Panel A: Volume bars						
1,000	62.62%	79.85%	78.58%	53.22%	39.65%	46.16%
2,500	68.86%	81.17%	80.26%	56.82%	41.17%	46.47%
5,000	73.97%	82.80%	82.17%	60.30%	43.94%	47.41%
10,000	78.75%	84.93%	84.58%	64.03%	47.66%	48.69%
15,000	81.16%	86.31%	86.09%	66.28%	50.37%	49.69%
20,000	82.71%	87.29%	87.15%	67.81%	52.72%	50.78%
25,000	83.73%	88.07%	87.98%	69.11%	54.35%	51.25%
30,000	84.52%	88.69%	88.61%	70.02%	55.96%	52.14%
40,000	85.62%	89.66%	89.59%	71.48%	58.43%	53.20%
50,000	86.32%	90.35%	90.34%	72.55%	60.28%	54.12%
75,000	87.37%	91.57%	91.53%	74.44%	63.57%	55.84%
100,000	87.90%	92.36%	92.27%	75.68%	65.88%	57.17%
Panel B: Time bars						
2 s	58.66%	79.32%	78.14%	60.30%	39.90%	46.03%
5	62.51%	79.94%	79.06%	62.90%	43.01%	47.05%
10	65.73%	80.60%	79.93%	65.59%	46.56%	48.27%
30	71.78%	82.32%	81.96%	70.81%	54.35%	51.18%
60	75.87%	83.92%	83.72%	74.14%	60.05%	53.62%
120	79.80%	85.87%	85.79%	77.18%	65.59%	56.46%
300	84.16%	88.65%	88.66%	80.45%	71.70%	60.68%
600	86.57%	90.66%	90.64%	82.37%	75.33%	64.21%
1,200	88.27%	92.41%	92.28%	83.94%	78.24%	68.02%
1,800	88.90%	93.24%	93.07%	84.78%	79.87%	70.40%
3,600	89.65%	94.49%	94.15%	86.03%	81.94%	74.27%
7,200	89.68%	95.50%	94.92%	87.14%	83.75%	77.81%

4.2. *Can the algorithms detect informative flow?*

4.2.1. *Overall spread results*

Although accuracy in classifying the aggressor side of trades has been our focus thus far, it is also important to examine underlying information. Informed traders are increasingly relying on passive orders, i.e., limit orders (which can be exacerbated by order splitting), to disguise themselves in the market (see, for example, Bouchaud et al., 2009; Menkhoff et al., 2010; Zhang, 2013 estimates the probability of informed liquidity provision to be 85% post-decimalization). These changes in informed trading raise two issues, which we explore in this section. First, can BVC detect informed trading? Second, is the identification of individual trade aggressors still important in measuring informative order flow? To examine these issues, we run regressions like those in ELO, regressing various spread estimates on estimated absolute order imbalance and lagged spread estimates,

$$Spread_\tau = \alpha_0 + \alpha_1[Spread_{\tau-1}] + \gamma\left|\widehat{OI}_\tau\right| + \varepsilon_\tau, \tag{3}$$

where the estimated absolute order imbalance (OI) is defined as

$$\left|\widehat{OI}_\tau\right| = \left|\frac{\hat{V}^B_\tau - \hat{V}^S_\tau}{V_\tau}\right| = \left|2\frac{\hat{V}^B_\tau}{V_\tau} - 1\right| \tag{4}$$

ELO argue that if an algorithm is capturing underlying information, then absolute order imbalance should be positively related to spread in bar τ. Theory in market microstructure predicts that with more informed trading, liquidity is affected as market makers widen spreads to protect themselves.

An issue here is how to measure our spread variable. Because BVC uses volatility as an input in its estimation, any possible relation of volatility to our spread measure could give us noisy or even spurious correlations. Therefore, we choose the Corwin and Schultz (2012) estimated spread measure, which specifically removes underlying asset volatility, and only extracts the illiquidity feature of the high-low price range for each bar. This measure is also used by ELO. In addition, we calculate intraday effective spreads as an alternative to the Corwin-Schultz measure. Effective spreads have been widely used in the literature to capture illiquidity, and they also seem unrelated to volatility (Chakrabarty et al., 2015; Table 6). We calculate effective spreads at the second (microsecond) timestamp level for the Euronext (LSE) sample and then take volume-weighted average over the bar period. For order imbalance, we use the ones based on BVC and bulk tick test estimates, respectively, as well as order imbalance based on actual trade aggressors, which are known in our data sets. Table 3 contains the results of these regressions for a selection of different volume bar sizes.[12] Note that from this table forward we do not report results for LR because we find that it is dominated by the tick test in our samples (in terms of data efficiency, accuracy, and information detection).

Every regression includes stock and sample-month fixed effects, and standard errors are clustered by stock. Using stock-fixed effects is econometrically important since it can capture stock-specific

bar, as opposed to trade, level data. This will reduce network bandwidth utilization. Similarly, necessary algorithmic back-testing of BVC requires less data storage capacity.

[12] Unreported results for time bars are similar.

Table 3

Spread Regressions on Order Imbalance.

This table displays results from the regression, $Spread_\tau = \alpha_0 + \alpha_1[Spread_{\tau-1}] + \gamma|\widehat{OI}_\tau| + \varepsilon_\tau$, with the addition of firm and month fixed effects. Please see the text for variable definitions. We measure within-bar order imbalance using BVC, bulk tick test (Tick OI), and trade aggressor (Aggressor OI), which is based on each trade's initiator, known in our data. Panel A displays results using the Corwin and Schultz (2012) estimator and Panel B uses volume-weighted effective spread within each bar. We display the order imbalance coefficient and its *t*-statistic for each spread measure and each imbalance measure. Standard errors are clustered by firm. ** and * denote statistical significance at the 1% and 5% levels.

Bar size	BVC OI estimate		Tick OI estimate		Aggressor OI		Observations
	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)	
Panel A: Corwin-Schultz spread estimator							
1,000	0.015**	3.04	0.004**	4.50	0.003**	4.76	6,285,058
5,000	0.028**	3.53	0.004**	4.36	0.003**	4.23	2,536,332
10,000	0.031**	3.38	0.003**	3.51	0.002**	3.56	1,529,693
30,000	0.028**	3.24	0.001	0.91	0.000	-0.83	611,580
50,000	0.029**	3.38	-0.001	-1.34	-0.001	-1.93	385,250
100,000	0.024**	3.23	-0.002	-1.63	-0.003*	-2.01	200,876
Mean effect	0.023**	3.52	0.002*	1.97	0.000	1.16	
Panel B: Effective spread							
1,000	0.260*	2.52	-0.013*	-2.63	-0.025**	-3.63	6,314,070
5,000	0.747	1.58	-0.020*	-2.15	-0.016**	-2.65	2,548,550
10,000	0.460*	2.42	-0.044	-1.70	-0.038	-1.55	1,537,894
30,000	0.131*	2.28	-0.018	-1.83	-0.018	-1.93	614,859
50,000	0.301	1.89	-0.018	-1.54	-0.018	-1.89	387,041
100,000	0.107*	2.35	-0.016	-1.56	-0.023	-1.54	201,267
Mean effect	0.151	1.44	-0.016	-1.41	-0.020	-1.71	

characteristics that have been shown to affect spreads, such as market capitalization and volatility. Sample-month fixed effects could capture overall market characteristics that affect spreads (e.g., upturn or downturn markets, periods of high and low market volatility, etc.). In the table, our three order imbalance measures are shown column-wise, with the imbalance coefficient (γ) and its *t*-statistic reported for each measure. Each panel of the table uses a different spread estimate.

Panel A contains the regression results using the Corwin and Schultz spread measure. Consistent with ELO, the γ coefficients are significantly positive across all bar sizes for BVC. In contrast, although bulk tick test and aggressor order imbalances have positive and significant coefficients over the first three sizes, the coefficients are insignificant for the size of 30,000, and then turn negative and significant.

These patterns are exaggerated in Panel B using volume-weighted effective spread. The γ coefficients for BVC are all significantly positive with larger magnitudes, but every coefficient is either insignificant or negative for the other two imbalance estimates. These results indicate that when BVC order imbalance is large the spread measures are also large. Since spreads reflect the possible price impact due to adverse selection, this is what theory would predict for the presence of informed order flow. Thus, the negative coefficients for bulk tick test and aggressor imbalances reveal two important takeaways. First, bulk tick test does not capture underlying informative flow, while BVC does match high order imbalances with higher spreads. Second, and more importantly, knowing the aggressor imbalance is not sufficient in modern markets to detect informed order flow. This suggests algorithms designed to flag aggressors cannot capture information. We explore this further below.

4.2.2. Spread results in subsamples

In Table 4, we focus on separate subsamples of small and large absolute returns. We expect the large absolute return subsample, which we define as bars with returns in the first or fourth quartile of non-absolute returns (for each bar size), to contain more informed trading due to larger price movements. Just as in Table 3, Panel A of Table 4 shows regression results for each OI estimate using the Corwin and Schultz (2012) spread measure, and Panel B uses volume-weighted effective spreads for each bar.

Since the BVC algorithm uses returns in its calculation, it is not surprising to find that BVC OI estimates relate more strongly to the spread measure for the large absolute return subsample. The difference is in the magnitude as well as the sign: the mean coefficient is 0.086 for the subsample of large absolute return bars versus -0.001 for small absolute return bars.[13] Interestingly, the tick and aggressor OI measures have similar results. In the large returns subsample, the coefficients are small and positive for small bars and negative for larger bars. The results in Panel B using effective spread are along the same lines, but with starker differences. For BVC, the coefficients in the small return subsample are close to zero, and only one is significant. All coefficients using the tick or aggressor OI are negative, and most are significant.

To test whether it is informed, passive orders that render bulk tick test unable to detect underlying information, we run the spread regressions in subsamples that are likely to contain informed trading using limit orders. Baruch et al. (2016) find that when borrowing costs are high for investors (when firms are not index members or when there is no options market) informed traders tend to use passive orders more often. If simply correctly classifying the aggressor side of a trade is no longer sufficient in detecting the underlying trading information, then bulk tick test's inverse relation between estimated order imbalance and spread measures should be exacerbated here.

In Table 5, we run the regressions separately for firms that do and do not have an active options market. We define active options market as *any* options volume in the Bloomberg Terminal for that stock month. Again Panels A and B use the Corwin and Schultz (2012) spread and effective spread as measures of illiquidity. The γ coefficients are positive for all regressions using BVC, with roughly equal coefficients across subsamples. Bulk tick test and aggressor results show similar patterns to those in Table 3 in

[13] In Tables 4 and 5, we estimate the mean effect sizes by weighting each regression coefficient by the reciprocal of the squared standard error. We estimate cross-bar size dependence using the ratio of the true to estimated standard errors found in Chordia et al. (2000, 2005) ($[1+2(N-1)\rho]^{1/2}$).

Table 4
Spread Regressions on Order Imbalance - Return Subsamples.
This table displays results from the regression, $Spread_\tau = \alpha_0 + \alpha_1[Spread_{\tau-1}] + \gamma|\widehat{OI}_\tau| + \varepsilon_\tau$, with the addition of firm and month fixed effects. Please see the text for variable definitions. We measure within-bar order imbalance using BVC, bulk tick test (Tick OI), and trade aggressor (Aggressor OI), which is known in our data. Panel A displays results using the Corwin and Schultz (2012) estimator and Panel B uses the volume-weighted effective spread in the bar. We display the order imbalance coefficient and its *t*-statistic for each spread measure and each imbalance measure. The regressions are further split by magnitude of returns, using large (first or fourth quartile of returns) and small (second or third quartile of returns). The distribution of returns that defines the quartiles is estimated for each bar. Standard errors are clustered by firm. The aggregated coefficients are weighted by the reciprocal of the squared standard error for each regression specification and the aggregate standard errors are corrected using the method from Chordia et al. (2000). ** and * denote statistical significance at the 1% and 5% levels.

	BVC OI Estimate				Tick OI Estimate				Aggressor OI			
	|Large ret|		|Small ret|		|Large ret|		|Small ret|		|Large ret|		|Small ret|	
	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)
Panel A: Corwin-Schultz estimator												
1,000	0.076**	3.79	-0.001**	-6.31	0.003**	3.79	0.004**	4.59	0.003**	4.39	0.002**	4.92
5,000	0.111**	4.45	-0.001**	-4.14	0.001**	2.60	0.005**	4.11	0.002**	4.11	0.002**	4.08
10,000	0.114**	4.20	-0.001	-1.64	-0.001	-1.25	0.005**	3.52	0.001*	2.60	0.002**	3.42
50,000	0.085**	3.96	-0.002	-1.10	-0.005**	-3.15	0.002**	2.18	-0.004**	-3.07	0.000	0.86
100,000	0.070**	3.85	-0.007	-1.22	-0.007**	-3.15	0.001	0.94	-0.008**	-3.29	0.000	0.07
Mean effect	0.086**	4.33	-0.001**	-3.83	0.001	0.99	0.003**	3.30	0.001*	2.35	0.001**	3.09
Panel B: Effective spread												
1,000	1.721**	2.74	0.000	-0.09	-0.022**	-2.93	-0.007**	-3.05	-0.034**	-3.38	-0.001	-0.65
5,000	2.350	1.64	0.000**	7.00	-0.028	-1.88	-0.005**	-3.56	-0.022*	-2.33	-0.002**	-2.75
10,000	1.370**	2.71	0.000	0.02	-0.077	-1.60	-0.006**	-2.28	-0.070	-1.45	-0.004*	-2.51
50,000	0.786	1.95	-0.002	-1.47	-0.034	-1.66	-0.002	-0.99	-0.038*	-2.19	-0.002	-1.32
100,000	0.241*	2.45	-0.007	-1.50	-0.026	-1.63	-0.005	-1.21	-0.044	-1.60	-0.004	-1.48
Mean effect	0.349	1.28	0.000*	2.50	-0.025	-1.46	-0.005	-1.84	-0.030	-1.67	-0.002	-1.42

Table 5
Spread Regressions on Order Imbalance – Options Trading Subsamples.
This table displays results from the regression, $Spread_\tau = \alpha_0 + \alpha_1[Spread_{\tau-1}] + \gamma|\widehat{OI}_\tau| + \varepsilon_\tau$, with the addition of firm and month fixed effects. Please see the text for variable definitions. We measure within-bar order imbalance using BVC, bulk tick test (Tick OI), and trade aggressor (Aggressor OI), which is known in our data. Panel A displays results using the Corwin and Schultz (2012) estimator and Panel B uses the volume-weighted effective spread in the bar. We display the order imbalance coefficient and its *t*-statistic for each spread measure and each imbalance measure. The regressions are further split by whether the stock had any options trading data in Bloomberg in our sample windows. Results using index membership are included in the appendix (Table 8A). Standard errors are clustered by firm. The aggregated coefficients are weighted by the reciprocal of the squared standard error for each regression specification and the aggregate standard errors are corrected using the method from Chordia et al. (2000). ** and * denote statistical significance at the 1% and 5% levels.

	BVC OI Estimate				Tick OI Estimate				Aggressor OI			
	Options market		No options		Options market		No options		Options market		No options	
	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)	Coef(γ)	t(γ)
Panel A: Corwin-Schultz estimator												
1,000	0.016**	2.73	0.011	1.44	0.004**	3.88	0.005**	2.67	0.003**	4.12	0.003**	2.77
5,000	0.028**	3.00	0.028*	2.25	0.004**	3.82	0.004**	2.90	0.002**	3.62	0.003**	3.38
10,000	0.033**	2.89	0.027*	2.15	0.003**	3.20	0.002*	2.09	0.002**	3.15	0.002**	3.27
50,000	0.030**	2.86	0.025*	2.40	-0.001	-1.13	-0.002	-0.89	-0.001	-1.86	-0.001	-0.44
100,000	0.025**	2.74	0.022	1.94	-0.002	-1.47	-0.003	-0.90	-0.004	-1.89	-0.001	-0.47
Mean effect	0.023**	2.96	0.020*	2.16	0.002	1.87	0.002	1.69	0.001	1.85	0.002*	2.46
Panel B: Effective spread												
1,000	0.079**	2.97	1.028*	2.02	-0.008**	-3.13	-0.040	-1.40	-0.008**	-2.67	-0.132**	-2.69
5,000	0.061**	3.24	3.350	1.44	-0.006**	-3.31	-0.108	-1.49	-0.003*	-2.09	-0.106*	-2.12
10,000	0.055**	3.10	1.949*	2.18	-0.005**	-3.03	-0.332	-1.49	-0.002*	-2.15	-0.307	-1.42
50,000	0.032**	3.07	1.299	1.78	-0.001	-1.64	-0.166	-1.51	-0.001	-1.25	-0.176	-1.94
100,000	0.043	1.95	0.379	1.86	0.000	0.22	-0.164	-1.69	0.000	0.24	-0.227	-1.73
Mean effect	0.045*	2.29	0.595	1.12	-0.002	-1.25	-0.066	-0.90	-0.002	-1.13	-0.136	-1.46

both subsamples. The coefficients are positive and significant in small bars, but are insignificant in the larger bars. The signs and magnitudes are such that the mean effects are insignificant, except for the aggressor OI measure in the subsample without options trading. In Panel B, the γ coefficients estimated using BVC are all positive, though with higher variation. The no options mean effect is insignificantly larger using effective spread. The coefficients using bulk tick test and aggressor OI are mostly insignificant in Panel B (the point estimates have negative signs). The mean effects in the no options markets subsamples tend to be more strongly negative than those in the subsample with an options market. The results of Table 5 suggest that when informed traders trade passively, trade level algorithms appear less capable of capturing informed order flow (i.e., a higher absolute OI is associated with a narrower spread).[14]

Overall, the results in the last three tables suggest that BVC-based order imbalances do well at detecting underlying information in comparison to order imbalances derived from identifying individual trades as buys or sells. If knowledge of the true aggressor side of individual trades no longer adequately captures infor-

[14] We find similar results in Table 5A of the internet appendix when we look separately for firms that belong and do not belong to a major stock index.

mative buying and selling activity, the utility of using tick test and LR algorithms to classify trades is diminished. This is exacerbated in subsamples likely to contain informed trading (large returns) and subsamples in which informed traders are likely to use passive orders.

5. Can BVC's ability to find traditional buy/sell pressure improve?

Results in Section 4 suggest BVC is the better choice when researchers are interested in detecting information, but that traditional algorithms are successful at identifying trade aggressors. Information is generally of utmost importance, but, as mentioned above, in some situations researchers want to detect traditional buying and selling pressure based on trade initiators, such as when calculating maker-taker fees, measuring trading costs, or identifying investor clienteles. Therefore, we are interested in capitalizing on BVC's data and information advantage while making its aggressor accuracy comparable to the other methods. While tick and LR cannot be calibrated because they are inherently trade-level, the heterogeneity of equity trade distributions, in terms of both size and arrival time, suggests that BVC should not be applied uniformly across different stocks. We examine several ways of calibrating BVC in this section: (1) choice of bar size and price change t-distribution degrees of freedom (*df*) parameter (2) time spacing and weighting considerations, and (3) exact versus minimum volume bar sizes. We then turn to how this calibration relates to its ability to detect information.

5.1. Calibrating BVC

5.1.1. Netting and bar size/distribution calibration

The choice of bar size affects BVC and bulk tick test/LR in very different ways. Tick and LR are trade-level; we use them to sign individual trades, and then aggregate those trades into volume or time bars. As noted in ELO and Chakrabarty et al. (2015), aggregation will offset misclassified trades, thereby increasing accuracy in the bar. Indeed, one can see in Table 2 that accuracy monotonically increases with bar size for these algorithms. On the other hand, it is less clear how bar size choices will affect BVC, because bar size influences the distribution of price changes across bars. In other words, as bar size changes, the numerator (price changes), the denominator (weighted standard deviation of price changes), and the *df* for the CDF in formula (2) will change.

To help us consider how *df* might be calibrated to certain securities, we visually examine excess kurtosis of price change distributions and how they vary with time bar size and firm size in Fig. 2. There is substantial variation in distribution shape across both bar and firm size dimensions. In particular, excess kurtosis tends to be lower for larger bars and for smaller capitalization stocks.

To adjust the shape of the distribution to kurtosis heterogeneity, we introduce an additional distribution parameter calibration, motivated by the return distribution analysis of Bakshi et al. (2003), which lowers the degrees of freedom in the t-distribution from 0.25 to 0.05 (0.1) for large (mid) cap stocks.[15]

5.1.2. Temporal spacing and weighting

To further investigate the effect of bar size choice on BVC accuracy, we consider scenarios in which bar size can either be "too small" or "too large," given the distribution of trade sizes. In particular, with respect to time bars, if the bar size is too small, the bar will not contain enough trades to benefit from netting misclassified trades. This reduced netting will impact both bulk tick test/LR and BVC algorithms. At the same time, however, a greater number of smaller bars will lead to a smaller standard deviation of price changes which will only impact BVC and *not* bulk trade level algorithms. To better see this effect, we present the mean volume within time bars in Panel A of Table 6.

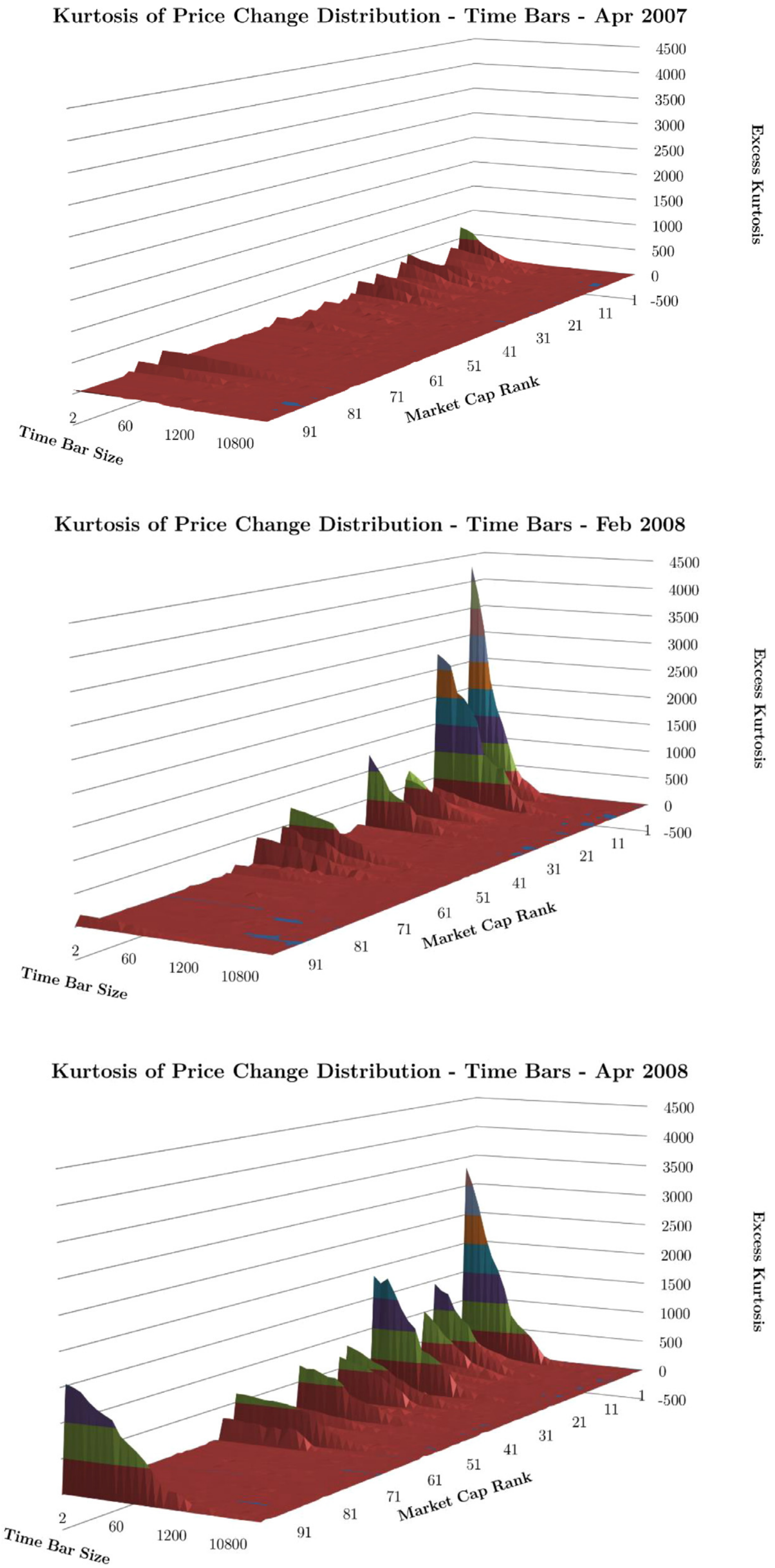
For small capitalization stocks, average bar volume does not increase with time bar size as much as it does for large capitalization stocks. For example, mean volume in five second bars is similar for small and large capitalization stocks at 3943 and 4168 shares, respectively. However, in a 10 min bar the mean volume is 13,326 and 128,734 shares for small and large capitalization stocks, respectively. Because of this small increase in mean volume for small cap stocks, a greater proportion of bars will likely have small price changes and thus bars will be more evenly weighted between buys and sells using BVC. Indeed, we believe that this is why BVC small bar accuracy results are low in Table 2.

On the other hand, if bars are "too large" and thus consecutive bars are too far apart in time, the price of the previous bar can become "stale" and much of the useful price variation within the current bar can be lost. This will affect BVC more than bulk tick test since a fundamental difference between the two algorithms is the duration of time that elapses between consecutive data points. In the case of the tick test, this time period is determined only by the arrival distribution of trades. Duration between time points for BVC, however, is influenced by both the distribution of trades and the choice of bar size. Slower trade arrival will result in longer time periods between volume bars since they will take more time to fill completely. Holding trade arrival constant, increasing volume bar size will have a similar effect. For time bars, since the duration of the bar is predefined, this temporal effect is driven by the choice of bar size as well as the presence of empty (zero volume) time bars, which increase the temporal distance between bars that generate price differences. Since less liquid stocks see longer and more frequent periods of trading inactivity, they are more likely to have time bars that are unequally spaced and contain very low volume.

To assess the potential impact of "staleness" in BVC implementation, we present the time elapsed between trades and volume/time bars for our sample in Panels B and C of Table 6. In Panel B, time elapsed between consecutive volume bars is much longer for small capitalization stocks, due in part to lower liquidity and trading frequency. For a volume bar size of 50,000, small capitalization stocks see an average of 7651.69 s between consecutive bars. Considering that there are 29,400 (30,480) seconds within each trading day in our Euronext (LSE) sample, this represents only 3.84 (3.98) volume bars per day. A standard deviation for 50,000 bar size of 10,021.6 s suggests that some small cap stocks may have as few as two volume bars in a given week. This reemphasizes the need for an appropriate choice of bar size.

Addressing the potentially negative impact of "staleness" on BVC's accuracy requires a separate investigation that we leave to future research. Of prime consideration is the estimation window to calculate the volume weighted standard deviation of price changes ($\sigma_{\Delta P_i}$ in formula (1)). Given that $\sigma_{\Delta P_i}$ is volume weighted, it does not seem reasonable that a large trade 11 months before a bar should have the same impact in the calculation of $\sigma_{\Delta P_i}$as a similar large trade 11 h prior. A simple way to address this is to let large, more liquid stocks have shorter windows (e.g., weekly or monthly), while using longer windows (e.g., quarterly) for small, less liquid stocks. To more rigorously approach this issue, one could use multi-dimensional weighted standard deviation of price changes, weighting on both time elapsed and volume.

[15] Bakshi et al. (2003) identify that the returns distribution kurtosis across stocks increases with market capitalization so, we follow their return distribution analysis by reducing the degrees of freedom in the large and mid-cap group Student's t-distributions to 0.05 and 0.1, respectively.


Kurtosis of Price Change Distribution - Time Bars - Apr 2007

Kurtosis of Price Change Distribution - Time Bars - Feb 2008

Kurtosis of Price Change Distribution - Time Bars - Apr 2008

Fig. 2. The surface graphs above show variation in excess kurtosis of the distribution of price changes with the choice of time bar size and market capitalization. Excess kurtosis is on the Y-axis, market capitalization rank in the sample on the X-, and time bar size on the Z-axis. The first three graphs show the Euronext sample months in 2007–2008 and last two graphs show LSE sample months in 2017.

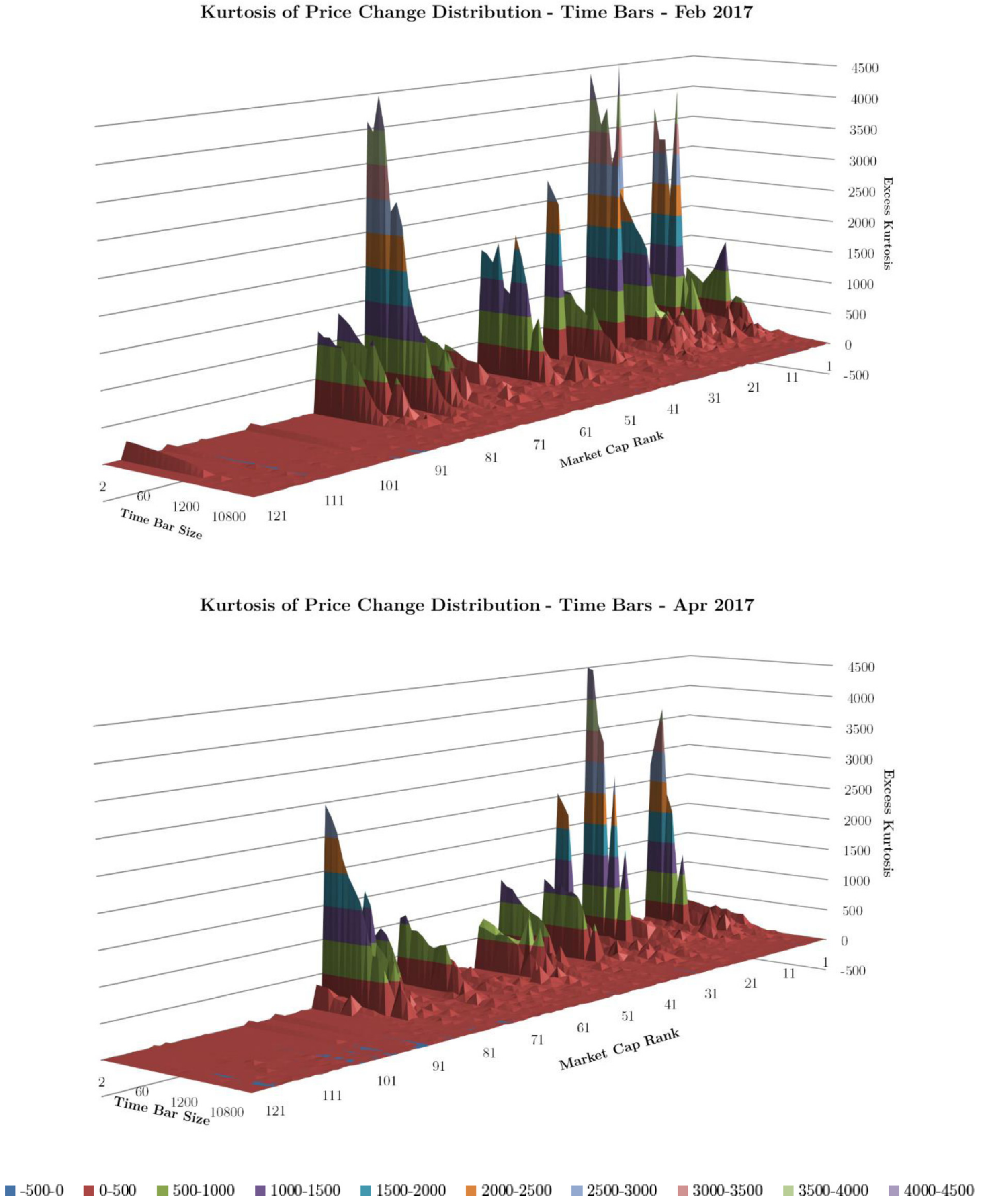


Fig. 2. Continued

5.1.3. *Bias in truncated versus minimum volume bar sizes*

A potential source of bias in the volume bar BVC applied to equities arises not from the choice of bar size but how that bar size is applied to the data. BVC proposed in ELO does not specify whether volume bars should contain volume equal to bar size or if that size is a minimum amount of volume for each bar. In the former case, if the last trade in the bar causes the volume in the bar to be greater than the specified size then the trade will be truncated and the remainder applied to the next bar. Suppose that $Trade_L$ is a large true buy trade executed at a price higher than that of both the previous trade and volume bar ($P_1 > P_0$). First, note that bulk trade level algorithms do not suffer from this bias because they classify trades before they are aggregated into bars. Whether the volume of this trade is then inserted within one volume bar or many, each part of the trade will be correctly signed in this case.

We display the bias that would have occurred in our results if we had used truncated, rather than minimum, volume bars in Table 7. This bias is estimated by summing the volume in each

Table 6
Time Bar Volume / Market Time Between Filled Bars.
Panel A displays the mean and standard deviation of the volume in a time bar. Panel B shows the mean and standard deviation of the time elapsed (in seconds) between volume bars. The first column in each panel shows the overall mean by stock-month, while the last three show the stock-month averages split by market capitalization. The small, medium, and large capitalization groups (firm size columns) are defined in Table 1. Our sample is the trading of 100 (125) firms listed on Euronext (LSE) for April 2007 and 2008 and February 2008 (February 2017 and April 2017). Our Euronext (LSE) data includes an order book time-stamped to the second (microsecond).

	Overall		Firm size					
			Small		Medium		Large	
Panel A: Time bars - Mean (Standard deviation) Volume								
Sub-second	1,055	(39,235)	2,766	(60,840)	725	(19,106)	1,173	(44,957)
1 s	2,380	(64,017)	3,491	(72,456)	2,008	(35,371)	2,476	(70,851)
5	3,738	(80,713)	3,943	(78,417)	2,694	(41,083)	4,168	(92,469)
10	4,792	(91,885)	4,241	(81,801)	3,177	(44,863)	5,587	(107,638)
30	8,161	(123,427)	5,077	(93,676)	4,689	(55,122)	10,490	(151,936)
180	27,271	(252,582)	8,189	(137,471)	13,455	(100,920)	42,673	(343,074)
300	40,721	(327,978)	9,875	(151,998)	19,820	(124,985)	67,494	(459,946)
600	71,918	(498,739)	13,326	(184,361)	35,039	(174,906)	128,734	(730,617)
900	101,172	(649,519)	16,296	(208,758)	49,792	(215,746)	188,970	(974,794)
1,800	182,857	(1,057,158)	24,040	(263,566)	92,318	(320,942)	366,031	(1,648,747)
3,600	327,136	(1,722,257)	37,298	(377,805)	170,930	(498,278)	686,838	(2,749,794)
7,200	571,948	(2,897,165)	60,271	(515,464)	306,830	(771,127)	1,232,013	(4,698,815)
Panel B: Volume bars - Mean (Standard deviation) market time in seconds elapsed between bars								
Trade Level	23.99	(619.1)	711.49	(3,541.5)	29.56	(668.4)	11.20	(406.2)
1,000	47.41	(408.5)	1,056.40	(2,900.7)	66.30	(357.4)	21.69	(92.9)
2,500	79.16	(578.9)	1,572.36	(3,875.4)	115.01	(529.4)	36.77	(129.9)
5,000	110.77	(722.8)	2,205.88	(4,720.1)	166.60	(723.1)	52.38	(173.3)
10,000	176.43	(981.2)	3,206.40	(5,925.5)	281.76	(1,068.2)	85.18	(278.8)
15,000	245.74	(1,204.9)	4,064.15	(6,736.2)	400.31	(1,367.9)	121.10	(375.5)
20,000	304.10	(1,383.7)	4,757.88	(7,412.3)	506.63	(1,598.0)	151.77	(466.8)
25,000	360.76	(1,559.0)	5,413.69	(8,032.5)	613.00	(1,849.3)	181.60	(549.9)
30,000	405.62	(1,670.7)	5,913.25	(8,439.8)	706.58	(2,013.1)	205.95	(608.9)
40,000	521.28	(1,971.0)	7,040.45	(9,382.2)	912.76	(2,376.2)	269.14	(767.8)
50,000	610.11	(2,162.5)	7,651.69	(10,021.6)	1,093.53	(2,635.9)	321.41	(888.2)
75,000	858.92	(2,697.9)	9,451.88	(11,559.6)	1,555.34	(3,283.3)	466.80	(1,207.7)
100,000	1,070.26	(3,073.1)	10,482.05	(12,642.9)	1,971.44	(3,772.2)	600.59	(1,473.1)

Table 7
Bias from Exact Volume Bar Implementation of BVC.
This table documents the bias from implementing the bulk volume classification (BVC) algorithm using exact volume bar sizes. It also shows the mean and standard deviation of trade size, as this will affect the size of the potential bias. The bias is shown for a range of bar sizes as well as on the cross-sectional cuts of sample month and firm capitalization. Please see the text and Figure 3 of the internet appendix for a detailed discussion of how the bias arises.

	Overall	Subsample period (Euronext)			Subsample period (LSE)		Firm size		
		April 2007	Feb 2008	April 2008	Feb 2017	April 2017	Small	Medium	Large
Mean trade size	1,054.8	650.0	648.2	609.5	1,481.1	1,571.9	2,765.8	725.4	1,173.2
Trade size std. dev.	39,234.7	1,523.1	1,322.1	1,252.8	56,936.3	69,059.7	60,839.5	19,106.1	44,956.7
Bias from exact volume bars									
1,000	33.38%	26.42%	25.01%	24.07%	35.22%	35.36%	43.88%	28.70%	34.18%
2,500	26.02%	17.76%	16.18%	15.40%	28.04%	28.11%	40.00%	21.40%	26.77%
5,000	21.38%	11.83%	10.52%	9.96%	23.71%	23.16%	37.34%	17.26%	21.96%
10,000	17.63%	7.39%	6.34%	6.01%	19.89%	19.35%	34.47%	14.43%	17.93%
15,000	15.90%	5.45%	4.63%	4.34%	17.97%	17.65%	32.58%	13.15%	16.07%
20,000	14.91%	4.28%	3.59%	3.41%	17.01%	16.57%	31.25%	12.36%	15.04%
25,000	13.96%	3.55%	2.98%	2.81%	15.97%	15.57%	30.05%	11.71%	14.02%
30,000	13.48%	3.10%	2.52%	2.42%	15.52%	14.99%	29.45%	11.40%	13.48%
40,000	12.74%	2.40%	1.94%	1.82%	14.79%	14.16%	28.03%	10.68%	12.77%
50,000	12.16%	1.93%	1.58%	1.50%	14.14%	13.55%	27.01%	10.37%	12.13%
75,000	11.22%	1.34%	1.06%	1.02%	13.15%	12.51%	24.79%	9.63%	11.18%
100,000	10.59%	0.99%	0.80%	0.77%	12.51%	11.77%	23.39%	9.14%	10.55%

volume bar over the specified size ("excess" volume) and multiplying the sum by one-half.[16] The bias can be large, especially in the LSE sample. With a bar size of 1000, the bias would have been 35.36% in April 2017. The bias monotonically declines with volume bar size, especially in the Euronext sample, where it is only 0.77% for bar sizes of 100,000. In the more recent LSE sample however, the minimum bias is 11.77%. Not surprisingly, the bias is larger when average trade size is large relative to volume bar size. These results suggest that minimum volume bar sizes should be used rather than forcing all volume bars to contain exact volume.

5.1.4. Calibration procedure

Following the analysis and results in Sections 5.1.1–5.1.3 above, we now move to calibrate BVC to stock trading characteristics. In

[16] When a large trade in a truncated volume bar fully fills the next bar, the price change is zero and the next bar will incorrectly be identified by BVC as half buy and half sell volume (see Figure 3 of the internet appendix).

finding an optimal bar size for each stock, we want to balance selecting a bar size that is too small, resulting in high excess kurtosis and little data compression, with a bar size that is too large, resulting in too few bars and masking meaningful price changes. To strike this balance, our calibration procedure searches for the smallest bar size s in the set of time (or volume) bar sizes S_{Time} (S_{Volume})—ordered from smallest to largest as shown in Table 2—subject to two constraints. First, we introduce a maximum excess kurtosis constraint (K_i) for each market capitalization group i, to better fit BVC's assumed t-distribution for the underlying price change distribution. Second, the volume (time) bar size must produce a minimum number of bars N_i (minimum volume per bar V_i) because we need "enough" data points (for example, imagine the extreme case in which there was only a single time bar in a month). For each firm j at each bar size s, we determine the excess kurtosis ($k_{s,\,j}$) and, either the average volume per bar ($v_{s,\,j}$) for time bars, or number of bars ($n_{s,\,j}$) for volume bars, and then search for the smallest bar size that satisfies the following:

$$Time\ Bars:\ For\ each\ firm\ j,\ \min_{s \in S_{Time}} s\ s.t.\ k_{s,j} \le K_i,\ v_{s,j} \ge V_i \tag{5}$$

$$Volume\ Bars:\ For\ each\ firm\ j,\ \min_{s \in S_{Volume}} s\ s.t.\ k_{s,j} \le K_i,\ n_{s,j} \ge N_i \tag{6}$$

We need to choose a kurtosis parameter K_i along with volume and quantity parameters V_i, and N_i for time and volume bar implementations, respectively. First, we note that in Fig. 2, excess kurtosis tends to rise with market capitalization. Liquidity also tends to rise with market capitalization, which we see in Table 6 in the greater volume per time bar (panel A) and less time from the end of one volume bar to the beginning of the next bar (panel B). Therefore, we choose separate parameters for small, mid, and large capitalization stocks. Looking at Fig. 2, we see that the excess kurtosis tends to smooth in the middle of each 3-D plot (across capitalizations), which roughly corresponds to a time bar size of 300 s. This time bar size can be translated into share volume using Table 6. For calibration minimums, we choose slightly less than the average volume in 300 s bars, using minimum average volume within a calibrated time bar of 5000, 10,000, and 50,000 shares for small, mid, and large capitalization firms. Though these minimums help limit kurtosis, we also directly apply excess kurtosis maximums of 100, 200, and 300 for small, mid, and large capitalization stocks based on the data underlying Fig. 2. In a similar fashion, we require a minimum number of volume bars (20, 40, and 60 bars per sample-month) for our market capitalization groups. If no bar size meets the requirements of formula (5) (formula (6)), we select the smallest time (volume) bar size that produces at least 10,000, 50,000, 100,000 average volume per bar (100, 200, or 300 bars).[17] We believe this procedure is generalizable to any equities market, but it will require parameter adjustment based on the distribution of price changes (driven by excess kurtosis) and volume in any set of trading data.

As part of the calibration, we also use flexible minimum volume bars to avoid the bias noted above, and we adjust the degrees of freedom for the underlying t-distribution based on market capitalization (0.25, 0.1, and 0.05). To run this procedure, we find the largest bar as described above using the data from the first month of each sample (i.e., April 2007 and February 2017) to find one bar size per stock. Then we use the bar selected from this first month in the remaining months of each sample to produce all the subsequent calibrated BVC results.

5.1.5. Calibrated BVC and aggressor accuracy results

Table 8 compares the aggressor accuracy results for calibrated BVC bars and randomly selected bars. Both use bulk tick as a reference point, so the percentages in the table are the result of taking BVC's accuracy minus bulk tick test's, and a positive number means BVC outperformed bulk tick test. The table is split by sample, but because the calibrated results (the second column in each sample) use bar sizes selected from the prior sample month's data according to the procedure in Section 5.1.4, only three total sample months are included here (i.e., February and April 2008 and April 2017). The results in Panels A and B use volume and time bars, respectively.

We start with the same basic implementation as Table 2, but use a market capitalization-adjusted *df* t-distribution (see Section 5.1.1) rather than a static df of 0.25. In the first column of the Euronext February 2008 sample, using randomly selected bars, the rate at which bulk tick test outperforms BVC ranges from 8.45% to 14.28% across market capitalization groups using volume bars (and from 7.80% to 15.50% for time bars). Calibration slightly reduces the accuracy of BVC across nearly all market cap groups in the Euronext February 2008 sample. Volume weighted accuracy across all firms decreases 2.15% (3.12%) for calibrated volume (time) bars relative to random bar sizes. With the exceptions of an improvement in accuracy for small caps and total calibrated time bar sizes, Euronext April 2008 results are similar.

We repeat these same comparisons in the two columns for April 2017 of the LSE sample. The differences are all positive relative to bulk tick test, highlighting again how BVC outperforms bulk tick (and LR) in the LSE sample. Interestingly, the calibrated bar sizes exhibit slightly lower accuracy relative to random bar sizes across all but one market capitalization group (small cap volume bars). Aggregate accuracy in the LSE sample decreases 3.02% and 0.45% for volume and time bars respectively.

Our evidence in this section shows that BVC's ability to classify the aggressor side of trades can be improved or maintain superiority to bulk tick test through calibration of BVC. We find that estimating an appropriate bar size through an iterative, parameterized procedure and using that in the subsequent sample period is an effective approach. Next, using both pooled spread regressions and an event study analysis, we examine whether calibrated BVC still detects informative order flow. Further, we also test whether bulk tick test and trade aggressor imbalances are themselves useful proxies for information.

5.2. Detecting information with calibrated bars

5.2.1. OI regressions with calibrated bars

It is possible that better aggressor accuracy changes how the OI estimated with calibrated BVC relates to the spread. For this test we re-run the OI regressions using only calibrated bars, that is, finding appropriate bar size in the first sample-month, then using this for the remaining months of each sample. These regressions pool the data together and add stock, bar size, and month fixed effects.

Table 9 displays the results of the spread regressions for volume bars using the Corwin and Schultz (2012) and effective spread measures. There are four different OI measures used across the eight models: We use calibrated BVC in the first two columns, BVC with a random bar size in the next two, bulk tick test in the next two, and finally the true aggressor OI in the last two. Comparing results between calibrated (models 1 and 2) and random (models 3 and 4) bar sizes, calibrated BVC order imbalance is positively related to both spread measures, but the random bar OI is only related to the effective spread. Bulk tick OI continues to perform poorly, with a negative and significant coefficient in model 6.

[17] This "catch all" rule is used infrequently in our three sample-months, and uses larger parameters to ensure we limit excess kurtosis, which is the most binding constraint in our calibration requirements.

Table 8
Comparison of BVC Calibrations and Bulk Tick Test.
This table displays accuracy comparisons of bulk volume classification (BVC) and bulk tick test for finding trade aggressor. Each percentage is the accuracy differential between algorithms, where we take BVC accuracy minus tick test accuracy, so a positive number means the BVC outperformed bulk tick test. For each sample, the first column displays differences using random bar sizes. The second column displays differences using a calibrated, Student's *t*-distribution implementation, where we calibrate each firm adjusting bar size and degrees of freedom by market capitalization (see Section 5.1.4 for calibration methodology details). Panels A and B display results for a selection of volume bars and time bars.

	Euronext Feb 2008		Euronext April 2008		LSE April 2017		All Samples	
	Random	*Calibrated*	*Random*	*Calibrated*	*Random*	*Calibrated*	*Random*	*Calibrated*
Panel A: Volume bars								
Small Cap	-14.28%	-7.02%	-13.86%	-7.59%	5.00%	4.92%	1.25%	2.45%
Mid Cap	-13.35%	-15.92%	-8.67%	-11.86%	14.07%	10.32%	9.99%	6.70%
Large Cap	-8.45%	-10.75%	-9.05%	-13.25%	14.27%	11.36%	8.66%	6.08%
Total	-9.31%	-11.46%	-9.09%	-12.89%	13.99%	10.97%	8.78%	6.14%
Panel B: Time bars								
Small Cap	-15.50%	-18.60%	-13.76%	-7.97%	7.52%	10.34%	3.29%	5.89%
Mid Cap	-12.94%	-21.93%	-18.95%	-21.24%	15.49%	9.19%	10.86%	4.66%
Large Cap	-7.80%	-9.84%	-8.34%	-2.47%	12.06%	13.18%	8.09%	9.29%
Total	-8.74%	-11.86%	-10.24%	-5.77%	12.71%	12.26%	8.50%	8.26%

Table 9
Regressions using Calibrated Volume Bars.
This table displays results for the regression, $Spread_\tau = \alpha_0 + \alpha_1[Spread_{\tau-1}] + \gamma|\widehat{OI}_\tau| + \varepsilon_\tau$ where columns (1), (3), and (5) display results using the Corwin and Schultz (2012) estimator and columns (2), (4), and (6) use the volume-weighted effective spread in the bar. Please see the text for variable definitions. Each model has stock, bar size, and month fixed effects. In this table, the Euronext (February 2008 and April 2008) and LSE (April 2017) data are pooled together. The first two models use the order imbalance estimated using calibrated BVC bar sizes, the next two use random BVC sizes, the next two use bulk tick test, and the last two the true trade aggressor, defined by the buy/sell initiator known in our data. In all regressions, we select the stock-month-bar size for BVC calibrated for each firm by bar size and degrees of freedom (see Section 5.1.4 for calibration methodology details). T-statistics derived from standard errors clustered by firm are below each coefficient estimate in parentheses. ** and * denote statistical significance at the 1% and 5% levels.

	Calibrated BVC OI		Random Bars BVC OI		Bulk Tick Test OI		Aggressor OI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Order Imbalance Estimate	0.0643**	0.2662*	0.0468**	0.1692	-0.0033**	-0.0034**	-0.0025	-0.0067
	(3.0810)	(2.3876)	(2.8190)	(1.7124)	(-2.7650)	(-2.6695)	(-0.8546)	(-1.3394)
Corwin-Schultz Estimator $_{t-1}$	0.6689**		0.5256**		0.6699**		0.6699**	
	(13.9812)		(14.3406)		(14.0012)		(14.0023)	
Effective Spread$_{t-1}$		0.6712**		0.6304**		0.6713**		0.6713**
		(34.9769)		(5.1082)		(35.0118)		(35.0155)
Bar Volatility	-0.0021	0.0174	-0.0010*	0.0224	-0.0018	0.0173	-0.0018	0.0168
	(-1.5916)	(0.9384)	(-2.3208)	(1.4840)	(-1.4658)	(0.9213)	(-1.4707)	(0.9237)
Zero Tick Volume	0.0001	0.0001	0.0000	0.0001	0.0000	-0.0000	0.0000	-0.0000
	(1.1232)	(0.4423)	(1.1075)	(0.8952)	(1.1141)	(-0.5980)	(1.1291)	(-0.5145)
Constant	-0.0101**	-0.0521	-0.0061**	-0.0266	0.0003	-0.0017*	0.0002	0.0002
	(-2.8211)	(-1.4390)	(-3.1635)	(-1.5833)	(0.8911)	(-2.5722)	(0.6450)	(0.1198)
Stock, Bar Size, Month FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	1,024,622	960,873	1,091,337	931,340	1,024,623	960,873	1,024,623	960,873
Adj. R-Squared	0.5015	0.4678	0.3657	0.4130	0.5004	0.4676	0.5003	0.4676

Interestingly, the coefficients using actual aggressor OI are insignificant, with both point estimates having a negative sign.

For robustness, we include additional control variables that might explain differences in information detection between BVC and tick. For example, one possible explanation for BVC's positive relation to the spread is volatility. Andersen and Bondarenko (2015) argue that BVC only detects order flow through a correlation with volatility. Although ELO argue that the Corwin and Schultz (2012) estimator addresses this concern, we further include actual within-bar price volatility and zero tick volume, the proportion of trades in a bar that have no price change from the prior trade. Zero returns are more likely to occur in the absence of information (Lesmond et al., 1999) and may systematically make bulk tick test performance worse. Including these variables (separately or together) does not change the results. Indeed, the inclusion of neither of these variables influences our results. Overall, Table 9 shows that calibrated BVC is still positively related to the spread, and this is not driven by in-bar volatility. Researchers can calibrate BVC to find aggressors without impairing how BVC's underlying order flow relates to the spread.

5.2.2. *Calibrated event study*

As an alternative test of the calibrated BVC's ability to detect information, we conduct an event study using event dates obtained from S&P Capital IQ. These dates include both scheduled (e.g., earnings announcements) and unexpected events (dividend changes, buyback announcements, merger and acquisition developments). We construct two-day event window cumulative abnormal returns (CARs) on [0, 1] adjusted by the CAC-250 (Euronext) or FTSE-350 (LSE) index returns. Across the 11 different event types, the mean (median) -0.11% (-0.05%) CAR is not significantly different from zero. The interquartile range is -11.94% to 8.13%, indicating that these event CARs capture both negative and positive information releases.

In Table 10, we present results for the 199 events with calibrated volume bars between the event date and two days prior (i.e. [-2,-1]).[18] Prior to the event, we sort firms on BVC order imbalance (volume-weighted average signed OI) using calibrated bar

[18] Unreported time bar results are similar.

Table 10
Event Study using Calibrated Volume Bars.
This table shows univariate and multivariate event study results for February and April 2008 using Euronext data and April 2017 using LSE data. Ten event types are obtained from Capital IQ and include both scheduled (earnings announcements) and unexpected (M&A announcements, buyback announcements, dividend increases/decreases, etc.) event dates. We calculate volume weighted signed order imbalance in the period between the event date and two trading days prior (i.e. days [-2,-1]). We select the stock-month-bar size for BVC calibrated for each firm by bar size and degrees of freedom (see Section 5.1.4 for calibration methodology details). For comparison, we also include bulk tick test and aggressor OI results. Mean abnormal CARs are adjusted by the CAC-250 (Euronext) and FTSE-350 (LSE) index returns respectively. T-statistics are displayed in parentheses. ** and * denote statistical significance at the 1% and 5% levels.

Panel A: Univariate quintiles

Order imbalance quintile	Calibrated BVC OI		Bulk tick test OI		Aggressor OI	
	N	CAR [0,1]	N	CAR [0,1]	N	CAR [0,1]
1	40	-0.024**	40	-0.012	40	-0.017*
		(-3.191)		(-1.451)		(-2.438)
2	40	-0.002	40	-0.005	40	0.006
		(-0.373)		(-0.855)		(1.167)
3	40	-0.002	40	-0.003	40	-0.002
		(-0.595)		(-0.801)		(-0.405)
4	40	0.004	40	0.007	40	0.003
		(0.931)		(1.464)		(0.494)
5	39	0.014*	39	0.002	39	-0.000
		(2.061)		(0.259)		(-0.079)
5–1	79	0.037**	79	0.013	79	0.016*
		(3.750)		(1.261)		(2.437)

Panel B: Multivariate

	Calibrated BVC OI	Bulk tick test OI	Aggressor OI
Order Imbalance	0.207**	0.026	0.045
	(4.309)	(0.898)	(1.329)
Constant	-0.022*	-0.018*	-0.016*
	(-2.542)	(-2.176)	(-1.972)
Month Fixed Effects	Yes	Yes	Yes
Event Type Fixed Effects	Yes	Yes	Yes
Number of Observations	199	199	199
Adj. R-Squared	0.286	0.143	0.156

sizes. We then form long-short portfolios, buying the highest quintile of order imbalance (i.e., the group with the most buying pressure) and selling the lowest. Panel A displays univariate CARs for each quintile of order imbalance and the long-short CAR between the highest and lowest quintiles. For comparison to the BVC OI, the table displays results using bulk tick test OI and the actual aggressor OI. Using the BVC, the long-short CAR is 3.7% and statistically significant at the 1% level. On the other hand, bulk tick test long-short CAR is insignificant, and while the CAR using aggressor is statistically significant, it is 210 basis points lower than the BVC portfolio.

In the multivariate specifications in Panel B, we regress event CARs [0,1] on pre-event order imbalance [-2,-1] and add event month and type fixed effects, which soak up heterogeneity within time and within event types. The coefficient on calibrated bulk volume classification order imbalance is positive and statistically significant. The positive coefficient indicates that when there is more pre-event buying (selling) pressure captured, there are larger (smaller) event CARs, as indicated in Panel A. Both bulk tick test and aggressor order imbalance coefficients are statistically insignificant, however. Thus, only calibrated BVC-estimated order flow leads returns across a wide range events, suggesting that this order imbalance estimate is picking up informative order flow.

6. Conclusion

The ability to correctly identify the aggressor side of each trade without order or even quote data has been a critical part of much of the market microstructure literature. Researchers commonly use the Lee and Ready trade classification algorithm if quotes are available or the tick test if not. The recently introduced bulk volume classification (BVC) has an alternative design that makes it much more data efficient. We test the performance of BVC in both classifying the aggressor side of trades and identifying informed trading. We also examine the tick test and LR for comparison.

To run our tests, we use a detailed data sample of Euronext (2007–2008) and LSE (2017) trades and quotes for which we can identify the aggressor of each trade. The uniqueness of our data plays an important part in the contribution of our paper. European markets have been slower to fragment than U.S. markets; therefore a large scale test of LR, the tick test, and BVC algorithms is much more feasible than attempting to aggregate all trades for a given listing across many exchanges that are executing them. Moreover, the data are from periods with different levels of algorithmic trading and they have different levels of granularity.

Our initial results of aggressor classification accuracy in the Euronext sample are similar to those in ELO and Chakrabarty et al. (2015) that BVC cannot classify trade aggressors as well as the tick test and LR. Using the newer LSE sample with more granular time stamps, however, we find that BVC is more accurate than traditional algorithms. Because BVC is a very new method, however, calibrating it for firm-specific implementations is required. After calibrating BVC using an iterative algorithm, we find it classifies aggressors nearly as well as the tick-test and LR in the Euronext sample. We believe our calibration procedure can be tailored to any equity data based on the distribution of price changes in those specific bars.

Importantly, we find—using both spread regressions and a returns event study—that BVC has a consistent advantage in capturing information rather than just trade aggressors, which suggests that BVC offers real advantages over methods built on signing individual trades. On further examination, we find that this is not driven by the tick test per se, but by a fundamental shift in markets where the aggressor side of trades does not appear to tell us as much about the underlying intentions of informed traders. This has implications for traditional measures of adverse selection designed to capture informed trading using trade aggressors (e.g., PIN measure of Easley et al., 1996 and Easley et al., 2002), since it suggests that these measures can be problematic in modern markets of fast executions and smart order trading. It also provides support to models of informed trading that either don't use trade aggressors, such as Johnson and So (2017) who propose a measure of informed trading based on abnormal volume imbalances across stock and options markets, or measures that use the BVC algorithm, such as VPIN (Easley et al., 2012).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jbankfin.2019.04.001.

References

Andersen, T.G., Bondarenko, O., 2015. Assessing measures of order flow toxicity and early warning signals for market turbulence. Rev. Finance 19 (1), 1–54.

Bakshi, G., Kapadia, N., Madan, D., 2003. Stock return characteristics, skew laws, and the differential pricing of individual equity options. Rev. Financial Stud. 16, 101–143.

Baruch, S., Glosten, L.R., 2016. Strategic foundation for the tail expectation in limit order book markets. Working paper. University of Utah.

Baruch, S., Panayides, M., Venkataraman, K., 2016. Informed trading and price discovery before unscheduled corporate events. J. Financial Econ. 125 (3), 561–588.

Battalio, R., Corwin, S., Jennings, R., 2016. Can brokers have it all? On the relationship between make-take fees and limit order execution quality. J. Finance 71, 2193–2238.

Berkman, H., Brailsford, T., Frino, A., 2005. A note on execution costs for stock index futures: information versus liquidity effects. J. Bank. Finance 29, 565–577.

Boehmer, E., Kelley, E.K., 2009. Institutional investors and the information efficiency of prices. Rev. Financial Stud. 22, 3563–3594.

Bouchaud, J.P., Farmer, J., Lillo, F., 2009. How markets slowly digest changes in supply and demand. Handbook of Financial Markets: Dynamics and Evolution. Elsevier, North-Holland.

Chakrabarty, B., Pascual, R., Shkilko, A., 2015. Evaluating trade classification algorithms: bulk volume classification versus the tick rule and the Lee-Ready algorithm. J. Financial Mark. 25, 52–79.

Chan, L.K.C., Lakonishok, J., 1993. Institutional trades and stock price behavior. J. Financial Econ. 33, 173–199.

Chordia, T., Roll, R., Subrahmanyam, A., 2000. Commonality in liquidity. J. Financial Econ. 56, 3–28.

Chordia, T., Roll, R., Subrahmanyam, A., 2005. Evidence on the speed of convergence to market efficiency. J. Financial Econ. 76, 271–292.

Conrad, J.S., Wahal, S., Xiang, J., 2015. High frequency quoting, trading, and the efficiency of prices. J. Financial Econ. 116 (2), 271–291.

Corwin, S.A., Schultz, P., 2012. A simple way to estimate bid-ask spreads from daily high and low prices. J. Finance 67 (2), 719–759.

Easley, D., Hvidkjaer, S., O'Hara, M, 2002. Is information risk a determinant of asset returns? J. Finance 57 (5), 2185–2221.

Easley, D., Kiefer, N., O'Hara, M., Paperman, J., 1996. Liquidity, information, and infrequently traded stocks. J. Finance 51, 1405–1436.

Easley, D., López de Prado, M.M., O'Hara, M., 2011. The microstructure of the flash crash: flow toxicity, liquidity crashes, and the probability of informed trading. J. Portfolio Manag. 37, 118–128.

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

Easley, D., López de Prado, M.M., O'Hara, M., 2016. Discerning information from trade data. J. Financial Econ. 120 (2), 269–285.

Ellis, K., Michaely, R., O'Hara, M., 2000. The accuracy of trade classification rules: evidence from Nasdaq. J. Financial Quant. Anal. 35 (4), 529–551.

Glosten, L.R., Milgrom, P.R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. J. Financial Econ. 14, 71–100.

Hasbrouck, J., 2018. High frequency quoting: short-term volatility in bids and offers. J. Financial Quant. Anal. 53 (2), 613–641.

Hendershott, T., Moulton, P., 2011. Automation, speed, and stock market quality: the NYSE's hybrid. J. Financial Mark. 14, 568–604.

Hendershott, T., Riordan, R., 2013. Algorithmic trading and the market for liquidity. J. Financial Quant. Anal. 48, 1001–1024.

Holden, C.W., Jacobsen, S., 2014. Liquidity measurement problems in fast, competitive markets: expensive and cheap solutions. J. Finance 69 (4), 1747–1785.

Holthausen, R.W., Leftwich, R.W., Mayers, D., 1987. The effect of large block transactions on security prices: a cross-sectional analysis. J. Financial Econ. 19 (2), 237–267.

Jain, P.K., 2005. Financial market design and the equity premium: electronic versus floor trading. J. Finance 60 (6), 2955–2985.

Johnson, T.L., So, E.C., 2017. Time will tell: information in the timing of scheduled earnings news. J. Financial Quant. Anal. Forthcoming.

Kirilenko, A., Kyle, A.S., Samadi, M., Tuzun, T., 2017. The flash crash: high-frequency trading in an electronic market. J. Finance 72 (3), 967–998.

Kyle, A.S., 1985. Continuous auctions and insider trading. Economet. J. Economet. Soc. 1315–1335.

Lee, C., Ready, M., 1991. Inferring trade direction from intraday data. J. Finance 46, 733–746.

Lesmond, D.A., Ogden, J.P., Trzinka, C.A., 1999. A new estimate of transaction costs. Rev. Financial Studies 12 (5), 1113–1141.

Mahmoodzadeh, S., Gençay, R., 2017. Human vs. high-frequency traders, penny jumping, and tick size. J. Bank. Finance 85, 69–82.

Menkhoff, L., Osler, C.L., Schmeling, M., 2010. Limit-order submission strategies under asymmetric information. J. Bank. Finance 34, 2665–2677.

Menkveld, Albert J., 2013. High-frequency trading and the new-market makers. J. Financial Mark. 16, 712–740.

Muravyev, D., Picard, J., 2016. Does trade clustering reduce trading costs? Evidence from periodicity in algorithmic trading. Working paper. Boston College.

Novy-Marx, R., Velikov, M., 2016. A taxonomy of anomalies and their trading costs. Rev. Financial Stud. 29, 104–147.

Odders-White, E., 2000. On the occurrence and consequences of inaccurate trade classification. J. Financial Mark. 3, 259–286.

O'Hara, M., 2015. High frequency market microstructure. J. Financial Econ. 116 (2), 257–270.

Pöppe, T., Moos, S., Schiereck, D., 2016. The sensitivity of VPIN to the choice of trade classification algorithm. J. Bank. Finance 73, 165–181.

Smidt, S., 1985. Trading floor practices on futures and securities exchanges: economics, regulation, and policy issues. In: Futures Markets: Regulatory Issues. American Enterprise Institute for Public Policy Research, pp. 49–152.

Zhang, Z., 2013. Informed liquidity provision and adverse selection measures. Working paper. Indiana University.