Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

Tracking Retail Investor Activity

Ekkehart Boehmer, Charles M. Jones, Xiaoyan Zhang and Xinran Zhang*

April 8, 2020

Abstract

We provide an easy method to identify marketable retail purchases and sales using recent, publicly available U.S. equity transactions data. Individual stocks with net buying by retail investors outperform stocks with negative imbalances, and the magnitude is approximately 10 basis points over the following week. Less than half of the predictive power of order imbalance can be attributed to order flow persistence. The rest is consistent with the hypothesis that retail investors are informed. We provide supportive evidence that retail investors are mainly informed about firm-level news, and that they are likely to have valuable private information.

*Boehmer is with Singapore Management University, LKC School of Business. Jones is with Columbia Business School. Zhang and Zhang are with Tsinghua University, PBC School of Finance. We thank Kevin Crotty, Larry Glosten, Frank Hatheway, Eric Kelley, Jamil Nazarali, Jeff Pontiff, Meijun Qian, Tarun Ramadorai, Paul Tetlock; seminar participants at Brigham Young University, University of California at Riverside, the Frankfurt School, Humboldt University of Berlin, INSEAD, Purdue University, Rice University, Singapore Management University, Southern Methodist University, Tsinghua University, University of Utah, National Singapore University, Nanyang Technology University; and conference audiences at the 2017 SFS Cavalcade, 2017 German Finance Association, 2017 ABFER, the 2017 CICF and the 2018 AFA for their helpful comments.

Electronic copy available at: https://ssrn.com/abstract=2822105

Can retail equity investors predict future stock returns? Or do they make systematic, costly mistakes in their trading decisions? The answers to these questions are important for other market participants looking for useful signals about future price moves, for behavioral finance researchers, and for policymakers who need to decide whether these investors should be protected from themselves.

Many researchers have concluded that retail equity investors are generally uninformed and, if anything, make systematic mistakes when selecting equity investments (e.g., Barber and Odean (2008), Barber, Odean, and Zhu (2009)). However, some more recent evidence, including Kaniel, Saar, and Titman (2008), Kaniel, Liu, Saar, and Titman (2012), Kelley and Tetlock (2013), Fong, Gallagher, and Lee (2014), and Barrot, Kaniel, and Sraer (2016), suggests otherwise. These studies show that retail investors' trading can predict future stock returns. Unfortunately, most existing studies of retail order flow are based on proprietary datasets with relatively small subsets of overall retail order flow. For example, Barber and Odean (2000) and Barber, Odean, and Zhu (2009) analyze data from a single U.S. retail brokerage firm; Kelley and Tetlock (2013) use data from a single U.S. wholesaler; Fong, Gallagher, and Lee (2014) analyze data from the Australian Securities Exchange (ASX); and Barrot, Kaniel, and Sraer (2016) use data from a single French brokerage firm. Kaniel, Saar, and Titman (2008), Kaniel, Liu, Saar, and Titman (2012), and Boehmer, Jones, and Zhang (2008) use proprietary account-type data from the NYSE during the early 2000s. During that sample period, only a small number of brokerages sent their retail order flow to the NYSE. As a result, the NYSE's market share of overall retail order activity was (and has remained) quite small.

In existing work, many researchers use trade size as a proxy for retail order flow. Before the spread of computer algorithms that "slice and dice" large institutional parent orders into a sequence of small child orders, small trades were much more likely to come from retail customers, while institutions were likely behind the larger reported trades. For example, Lee and Radhakrishna (2000) use a $20,000 cutoff point to separate smaller individual trades from larger institutional trades. More recently, Campbell, Ramadorai, and Schwartz (2009) effectively allow these cutoff points to vary through a regression approach that is calibrated to observed quarterly changes in institutional ownership, but they maintain the same basic assumption that small trades are more likely to arise from individual trading. However, once algorithms become an important

Electronic copy available at: https://ssrn.com/abstract=2822105

feature of institutional order executions, in the early 2000s, this trade-size partition becomes far less useful as a proxy for retail order flow. In fact, the tendency for algorithms to slice orders into smaller and smaller pieces has progressed so far that we find that during our recent sample period, the retail order flow that we identify actually has a slightly larger average trade size compared to other flow.

Given the current automated and algorithm-driven market structure, researchers need an easily implementable method to isolate retail order flow. We introduce such a measure in this paper. As one of our main contributions, we show that our measure can identify a broad swath of marketable retail order flow. Our measure builds on the fact that, due to regulatory restrictions in the U.S. and the resulting institutional arrangements, retail order flow, but not institutional order flow, can receive price improvement, measured in small fractions of a cent per share. We use this fact to identify retail price-improved orders from the TAQ data, a publicly available data set that contains all transactions for stocks listed on a national exchange in the U.S. We do this by identifying trades that execute at share prices with fractional pennies. Most such price-improved transactions take place off-exchange and are reported to a Trade Reporting Facility (TRF). Using this TRF data, we identify transactions as retail buys if the transaction price is slightly below the round penny, and retail sells if the transaction price is slightly above the round penny. This approach isolates retail investors' marketable orders from institutional ones, because institutional trades cannot receive this type of fractional penny price improvement.[1] We discuss our approach in greater detail in the data section. Overall, we believe that our method of retail trade identification is conservative, and we cross-validate the accuracy of our approach using a small sample of Nasdaq TRF audit trail data.

We analyze retail marketable order flow from the U.S. equity market for six years between January 2010 and December 2015. We find that retail investors are slightly contrarian at a weekly horizon, and that the cross-section of weekly retail order imbalances predicts the cross-section of returns over the next several weeks, consistent with the findings of Kaniel, Saar, and Titman (2008), Kaniel, Liu, Saar, and Titman (2012), Kelley and Tetlock (2013), Fong, Gallagher, and

[1] In contrast, institutional trades often occur at the midpoint of the prevailing bid and ask prices. If the bid-ask spread is an odd number of cents, the resulting midpoint trade price ends in a half-penny. Many of these midpoint trades take place on crossing networks and are reported to the TRF. Thus, trades at or near a half-penny are likely to be from institutions and are not assigned to the retail category.

Electronic copy available at: https://ssrn.com/abstract=2822105

Lee (2014), and Barrot, Kaniel, and Sraer (2016), but inconsistent with the findings of many others. The predictability of retail order flow for future returns is consistent with three hypotheses: persistence in retail order flow, liquidity provision, and informed trading. We conduct a decomposition exercise and separate the retail order imbalance into proxies for these components. The empirical findings show that the persistence in order flow, and order flow driven by return reversals (our proxy for liquidity provision), accounts for about half of the predictive power of the retail order imbalance for future returns, and we attribute the other half to informed trading. We go one step further and investigate the nature of the information embedded in retail trading. Our results show that the retail order imbalance is positively correlated with some firm-level surprises in public news, and the retail order flow has predictive power beyond public news, consistent with their possession of firm-level private information. Finally, we conduct a battery of robustness checks and provide further discussion. Our results are robust, and we provide additional evidence that, despite the predictive power of retail order flows in the cross section, aggregate retail flows cannot predict future market returns.

Given the nature of our data, our work is also related to recent studies of off-exchange trading in the U.S. For instance, Kwan, Masulis, and McInish (2015) study the competition between traditional stock exchanges and new dark trading venues and find that the minimum pricing increment regulation (typically one penny) drives orders to dark pools and limits the competitiveness of the exchanges. Battalio, Corwin, and Jennings (2016) examine make-take fees and how brokers route order flow, and suggest that current order routing practices may not maximize the quality of limit order execution. Menkveld, Yueshen, and Zhu (2017) directly investigate the pecking order of trading venues in dark pools and document that investors strategically put low-cost-low-immediacy orders in front of high-cost-high-immediacy orders.

Compared to earlier literature on retail orders and studies of off-exchange trades, we make three main contributions. First and most importantly, we propose a novel methodology for identifying and signing marketable retail trades using publicly available data with substantial coverage. Second, our empirical results show that the retail trades we identify can predict future stock returns. Third, we analyze the nature of the predictive power of retail order flow and show that half of its predictability is likely driven by order imbalance persistence and liquidity provision, while the other half is consistent with informed trading. We also track potential informed trading to different types of news and provide evidence that these retail investors seem to be informed

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

about some types of firm-level news and that the predictive power of their trades extends beyond public news.

Two studies, Kaniel, Saar, and Titman (2008) and Kelley and Tetlock (2013), study similar questions, and are closely related to our research, but with different data and different interpretations. For instance, using proprietary data from the NYSE between January 2000 and December 2003, Kaniel, Saar, and Titman (2008) document that retail order flows can predict stock returns. In addition, Kaniel, Saar and Titman (2008) examine the contemporaneous relation between their retail order flows and stock returns. They find that the contemporaneous return is significantly positive for stocks that retail investors sell, and negative for stocks that they buy, which is consistent with a liquidity provision interpretation and inconsistent with the information story. We follow their approach using our new retail order flow variables. We are able to replicate the predictive relation between retail order flow and future stock returns. However, our results for the contemporaneous relation are different: the contemporaneous return is significantly negative for stocks that retail investors sell, and positive for stocks that they buy. Our findings are more in line with an information interpretation than a liquidity provision interpretation.

Kelley and Tetlock (2013) obtain data from a major retail wholesaler between February 2003 and December 2007. Their data allow them to separate retail orders into market orders and limit orders. They find that retail market orders and limit orders can both predict future stock returns, but for different reasons. The aggressive market orders can correctly predict future news, suggesting these trades are informed, while the passive limit orders are contrarian, consistent with the liquidity provision hypothesis. Our retail order flow measure only identifies market orders, and for these marketable orders we follow their tests and replicate their results. In addition, we decompose our retail order imbalance into components related to order flow persistence, contrarian trading, public news and a residual, which potentially contains private information. The decomposition exercise shows that public news barely contributes to the predictive power of the retail trades, and the residual part is more important. With more recent data and wider coverage, our study provides interesting new findings, which complement the original studies by Kaniel, Saar, and Titman (2008) and Kelley and Tetlock (2013).

The remainder of this paper is organized as follows. We describe the data and our identification method in Section I. Section II presents our main empirical results. We provide

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

further discussion of the results and perform robustness and plausibility checks in Section III. Finally, Section IV concludes the paper.

I. Identifying Retail Order Flows

A. Data Sources

We collect information on retail investor activity from TAQ trade data. We keep only trades that occur off-exchange, with exchange code "D." We merge these TAQ data with stock returns and accounting data from CRSP and Compustat, respectively. We include only the common stocks with share code 10 or 11 (which excludes mainly ETFs, ADRs, and REITs) listed on the NYSE, NYSE MKT (formerly the Amex), and Nasdaq. We remove low-priced stocks by requiring the minimum stock price to be $1 at the previous month-end.

Our sample period covers January 3, 2010 to December 31, 2015. Data on subpenny price improvement actually extend back to 2005. In Appendix Figure 1, we show the time series from January 2005 (the start of Regulation NMS, which established the current regulatory framework for subpenny price improvement in the U.S.) to December 2017. We choose to study the period from 2010 to 2015 for two reasons. First, for the first few years of Reg NMS, there is a strong upward trend in the number of subpenny trades, possibly because an increasing number of brokerage firms were adopting the practice of providing fractional cents of price improvement to retail investors via internalization or wholesalers. The upward trend disappears and stabilizes after 2009. Second, from 2016 to September 2018, the SEC adopted a tick size pilot program (TSPP) that affects tick size and brokers' ability to provide price improvement for many stocks, which likely affects the prevalence of subpenny price improvements unevenly in the cross section. Therefore, our main results are focused on the middle part of the data, 2010 to 2015. For each day, we have an average of around 3,000 firms included in the sample.

B. Institutional Background and Methodology

In the U.S., most marketable equity orders initiated by retail investors do not take place on one of the dozen or so registered exchanges. Instead, these retail orders are typically executed either by wholesalers or via internalization, meaning that orders are filled from a broker's own inventory. Orders executed by wholesalers or through internalization must be publicly reported; they are usually reported to a FINRA Trade Reporting Facility (TRF), which provides broker-dealers with a mechanism through which to report transactions that take place off-exchange. These

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

TRF executions are then included in the TAQ "consolidated tape" of all reported transactions with exchange code "D." Many orders that are internalized or executed by wholesalers are given a small price improvement relative to the National Best Bid or Offer (NBBO).[2] For instance, wholesalers are willing to provide a small price improvement to induce the retail trader's broker to route the order to the wholesaler. Internalizers, who are subject to Regulation 606T, need to show that they execute their clients' orders optimally, and thus also have incentives to provide price improvement to their clients. This price improvement is typically only a small fraction of a cent. Common price improvement amounts are 0.01, 0.1, and 0.2 cent.

Brokerage firms in the U.S. are required to provide regular summary statistics in SEC Rule 606 filings about their order routing practices for non-directed orders. A directed order instructs the broker to execute an order on a given exchange or trading venue; a non-directed order gives the broker discretion regarding the execution venue. The vast majority of retail orders are non-directed. For example, Charles Schwab reports that 98.6% of their security orders during the second quarter of 2016 were non-directed orders. The corresponding figure for TD Ameritrade is 99%. According to the Rule 606 filings by these two retail brokerage firms, more than 90% of these orders receive price improvement.

Our communications with a major retail wholesaler and a major exchange suggest that these types of price improvement are not a feature of institutional order executions, as institutional orders are almost never internalized or sold to wholesalers. Instead, their orders are sent to exchanges and dark pools, and Regulation NMS prohibits these orders from having subpenny limit prices. Thus, institutional transaction prices are usually in round pennies. The only exception applies to midpoint trades. Reg NMS has been interpreted to allow executions at the midpoint between the best bid and best offer. As a result, institutions are heavy users of crossing networks and midpoint peg orders that generate transactions at this midpoint price. Since the quoted spread is now typically one cent per share, this means that many institutional transactions are reported at a half-penny. There is also a dark pool that, for a time, allowed some negotiation around the midquote and thus printed trades at 0.4, 0.5, or 0.6 cent.

[2] As a rough estimate of the frequency of subpenny price improvement, we find in a Nasdaq subsample used for robustness tests (introduced in Section I.D) that 60% of trades on "retail" venues receive subpenny price improvements, with 14% reported at the halfpenny and 46% taking place at a different subpenny. For subpenny trades that do not execute at half pennies and constitute the focus of our study, more than 99% are reported to a TRF with exchange code "D."

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

Based on these institutional arrangements, identifying transactions initiated by retail customers is fairly straightforward. Transactions with a retail seller tend to be reported on a TRF at prices that are just above a round penny due to the small price improvement, while transactions with a retail buyer tend to be reported on a TRF at prices just below a round penny. To be precise, for all trades reported to a FINRA TRF (exchange code "D" in TAQ), let P_{it} be the transaction price in stock i at time t, and let $Z_{it} \equiv 100 * \mathrm{mod}(P_{it}, 0.01)$ be the fraction of a penny associated with that transaction price. Z_{it} can take any value in the unit interval [0,1). If Z_{it} is in the interval (0,0.4), we identify it as a retail sell transaction. If Z_{it} is in the interval (0.6,1), then the transaction is coded as a retail buy transaction. To be conservative, transactions at a round penny ($Z_{it} = 0$) or near the half-penny ($0.4 \leq Z_{it} \leq 0.6$) are not assigned to the retail category.

As discussed above, Reg NMS requires that limit orders be priced at round pennies, so our approach will by definition identify only marketable retail orders.[3] The 606 filings by brokerage firms are also partitioned into market and limit orders, which allows us to gauge the relative prevalence of these two types of orders. For example, the Charles Schwab brokerage firm reports that for the second quarter of 2016, market orders account for 50.0% of its customers' non-directed orders in NYSE-listed securities, while limit orders account for 45.1%, and other orders account for the remainder. For securities listed on Nasdaq, limit orders are slightly more prevalent than market orders at Schwab, with market orders accounting for 44.0% and limit orders, 50.7%. Note also that limit orders may be cancelled without execution, and may in fact be marketable, so most overall retail trading activity is likely to arise from marketable orders. Thus, our approach probably picks up a majority of the overall retail trading activity.[4]

C. Summary Statistics

Table I presents summary statistics on the retail investor activity identified by our method. We pool observations across stocks and days, and compute the mean, standard deviation, median, and 25th and 75th percentiles. Our sample comprises over 4.6 million stock-day observations. For the number of shares traded per day (*vol*), the mean share volume is around 1.23 million, and the

---

[3] Marketable orders tend to be more aggressive. According to Kelley and Tetlock (2013), market orders are more informed than are limit orders. Thus, the predictive power of retail market orders is likely to be stronger than that of retail limit orders and overall retail orders, as has been documented in Kelley and Tetlock (2013).

[4] One might wonder whether the market or marketable orders can be offset, in aggregate, by limit or non-marketable orders in the opposite direction. This is possible. Unfortunately, we do not have data to directly check this possibility.

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

standard deviation is about 6.85 million shares. The average stock has 5,917 trades each day (*trd*). These numbers suggest that the average trade size over this sample period is about 200 shares. Our identified retail investor activity represents only a small part of the overall trading volume. The identified average daily buy volume from retail investors (*indbvol*) is 42,481 shares, and the average daily sell volume from retail investors (*indsvol*) is 42,430 shares. Thus, we identify an average of 84,911 shares per stock-day traded by retail investors, about 6.91% of the average total shares traded each day. The average number of buy trades from retail investors (*indbtrd*) each day is 110, and the average number of sell trades from retail investors (*indstrd*) each day is 108. Thus, the total number of identified trades per stock-day from retail investors is 218, around 3.68% of the total number of trades. Interestingly, the buy volumes closely match the sell volumes, and the number of buy trades match the number of sell trades, both indicating that many retail trades offset each other. In terms of average share volumes and number of trades, there is slightly more buying than selling by retail investors over our sample period.

Information on odd lot trades (trades of fewer than 100 shares) is reported on the TRF and on the consolidated tape beginning in December 2013 (see O'Hara, Yao, and Ye, 2014). During the sample period from December 2013 to December 2015, for which odd lot data are available, the daily averages of marketable odd lot retail buy and sell volumes (*oddindbvol* and *oddindsvol*, respectively) are 506 and 443 shares respectively, totaling 949 shares traded by retail investors in odd lots per average stock-day. This is about one-third of the total odd lot share volume, at 3,027 shares. The pattern for the number of trades is similar. Prior studies of odd lots generally find that these retail-dominated orders are virtually uninformed, so we study odd lots separately to determine whether the information content in marketable odd lots executed by retail customers differs from that of retail round lots.

Figure 1 provides further statistics on the overall properties of our identified retail trades. Panel A presents trade sizes in dollars. For each retail trade, we compute the trade size in dollars by multiplying the number of executed shares by the transaction price. For each year of our sample, we compute the 25th percentile, the median, and the 75th percentile of retail trade size. The median retail trade size is around \$8,000, and the interquartile range is mostly between \$2,000 and \$25,000. Panel B reports the distribution of subpenny prices. We separate all trades into 12 groups or bins. We separate out trades that take place at a round penny or half penny; the other bins are each 0.1 cent wide. We pool the sample across days and stocks, and we report the number of shares

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

reported in the different subpenny buckets. Not surprisingly, most of the share volume occurs at round and half-pennies, with average stock-day share volumes of around 27,000 and 7,000, respectively. The next most prevalent occurrence, averaging around 3,000 shares per day per stock, is a subpenny price within 0.1 cent of a round penny. Other subpenny bins are less prevalent, with most averaging around 1,000 shares per stock per day.

We measure retail investors' directional trades by computing four order imbalance measures for each stock *i* on each day *t*:

$$oibvol(i,t) = \frac{indbvol(i,t) - indsvol(i,t)}{indbvol(i,t) + indsvol(i,t)}, \tag{1}$$

$$oibtrd(i,t) = \frac{indbtrd(i,t) - indstrd(i,t)}{indbtrd(i,t) + indstrd(i,t)}, \tag{2}$$

$$oddoibvol(i,t) = \frac{oddindbvol(i,t) - oddindsvol(i,t)}{oddindbvol(i,t) + oddindsvol(i,t)}, \tag{3}$$

$$oddoibtrd(i,t) = \frac{oddindbtrd(i,t) - oddindstrd(i,t)}{oddindbtrd(i,t) + oddindstrd(i,t)}. \tag{4}$$

The first two measures are calculated using retail round lot executions between January 2010 and December 2015 and by aggregating round lot and odd lot executions thereafter, while the last two measures are calculated using only retail odd lots, and thus these latter measures begin in December 2013 instead of December 2010.

Summary statistics on the retail order imbalance measures are reported at the bottom of Table I. Across all stocks and all days, the mean order imbalance for share volume, *oibvol*, is -0.038, with a standard deviation of 0.464, and the mean order imbalance for trade, *oibtrd*, is -0.032, with a standard deviation of 0.437. The correlation between *oibtrd* and *oibvol* is around 85%. Our later discussions mostly focus on *oibvol*, but the results using these two measures are quite similar given the high correlation between the two. Overall, the order imbalance measured in shares is close to zero on average, but with sells slightly more prevalent than buys, which is consistent with findings in Kaniel, Saar, and Titman (2008). More importantly, the sizable standard deviation measures show that there is substantial cross-sectional variation in the activity levels and trading direction of retail investors. The odd lot order imbalance measures exhibit similar patterns.

Electronic copy available at: https://ssrn.com/abstract=2822105

In Figure 2, we plot the time-series of the cross-sectional mean, median, and 25th and 75th percentiles of the retail order imbalance measures over the six-year sample period. Across all four order imbalance measures, the means and medians are all close to zero, while the 25th percentiles are mostly around -0.3, and the 75th percentiles are mostly around 0.2. There are no obvious time trends or structural breaks in the time-series observations.

We extensively examine other properties of the retail order imbalance measures. To save space, we put them in the Appendix. The order imbalance measure's daily autocorrelations are reported in Appendix Figure 2 Panel A. The daily order imbalance measures are mostly significantly positively correlated with their nearby lags, while the cross-firm median correlation is 0.15. This positive autocorrelation is statistically significant over horizons up to a few months. The persistence of marketable retail order flow is slightly higher for larger firms than for smaller firms. Appendix Figure 2 Panel B presents the time-series correlation between the retail order imbalance measure and past returns. The results display a V-shape, indicating that the correlation between the current retail imbalance and the previous one-day return is positive on average (which is consistent with momentum trading), then becomes negative (which is consistent with contrarian trading) for the next 30 trading days. Finally, Appendix Table I reports the results for the measure's seasonality and its relation with variables reflecting firm fundamentals.

D. Cross Validation with Nasdaq TRF Data

Our main data source is TAQ, which does not provide direct information on the direction of the trade or the identity of the traders. We mainly validate our retail order imbalance algorithm through a small sample of proprietary Nasdaq data.[5] The same dataset is used in Menkveld, Yueshen, and Zhu (2017), who provide more details about the data. The Nasdaq sample covers all intraday transactions on its TRF for 117 stocks for the month of October 2010. The 117 stocks are chosen from different size groups, but they are generally larger than a typical firm in TAQ.[6]

For each trade, the Nasdaq TRF data provide a trade direction indicator: "buy," "sell," or "cross." Our algorithm identifies all subpenny trades with subpenny prices between 0.61 and 0.99 cents inclusive as "buy" trades. We separate all subpenny "true buy" trades (as indicated in the

[5] We thank Nasdaq for generously providing the data.
[6] The smallest market cap of the 117 Nasdaq firms is 257 million dollars, while our sample's 40 percentile market cap is merely 243 million dollars.

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

Nasdaq TRF data) with price below $100 into two categories: “identified buy” and “identified sell (false identification).” We falsely identify 1.37% of all subpenny “true buy” trades as “sell.” Similarly, our algorithm identifies all subpenny trades with subpenny prices between 0.01 and 0.39 basis points as “sell” trades. In this case, we falsely identify 2.12% of all subpenny “true sell” trades with price below $100 as “buy.” If we put identified retail “buys” and “sells” together, for stocks with a share price below $100, our subpenny approach matches the Nasdaq TRF’s correct buy/sell sign 98.2% of the time, while the standard Lee and Ready (1991) trade-signing algorithm gets the trade sign right 96.7% of the time. Overall, we find our algorithm is quite accurate for trade direction identification.

Menkveld et al. (2017) describe that when order flows come in, they will be routed to different types of off-exchange venues, depending on the cost and immediacy of the trade execution. The Nasdaq TRF data identify five types of off-exchange venues: “DarkNMid,” “DarkMid,” “DarkOther,” “DarkPrintB,” and “DarkRetail.” Our communication with a major retail-wholesaler and the Nasdaq indicates that other than “DarkRetail”, the venue types are a mix of all kinds of traders. Thus, the venue is not a precise indicator for a trader’s identity, and even if one had access to the Nasdaq TRF sample for a larger cross-section over a longer period of time, there would still be an important role for our algorithm in identifying retail buys and sells.

Our main measure in this article is order imbalance. The correlation between our order imbalance measure and the one calculated using the “DarkRetail” order imbalance for the 117 stocks is 0.70. This correlation is less than one for two main reasons: First, our order imbalance measure includes trades printed on the competing NYSE TRF, while the Nasdaq TRF dataset does not; second, our order imbalance measure includes some subpenny trades from the “DarkNMid” and “DarkMid” venues, in addition to those in “DarkRetail.” Finally, some retail market orders do not receive price improvement or receive a full half-cent of price improvement. We do not sign these trades or include them in our retail sample, because we cannot be sure that we have the correct trade direction. Nevertheless, the high correlation between our retail order imbalance measure and the actual Nasdaq “DarkRetail” venue data strongly suggests that our order imbalance measures closely reflect the true marketable buy and sell activities of retail investors.[7]

[7] Kelley and Tetlock (2013) compute retail order imbalance measures using data from one large wholesaler. We are grateful to Eric Kelley for calculating and sharing these correlations with us. Kelley computed the retail order

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

II. Empirical Results

In the data section, we measure order imbalances at the daily level to minimize the amount of aggregation. For our main empirical results, we focus on weekly horizons to reduce the impact of microstructure noise on our results. That is, our main variables of interest are firm-level average retail order imbalances over five-day horizons and five-day firm-level stock returns. Blume and Stambaugh (1986) show that using the end-of-day closing price to compute daily returns can generate an upward bias due to bid-ask bounce. Therefore, we compute two versions of weekly returns, one by compounding CRSP daily returns, based on daily closing prices, and one by compounding daily returns using the end-of-day bid-ask average price. We always report the results for both types of returns but focus our attention on returns based on closing bid-ask averages.

We start by investigating the properties of the order imbalance measures in Section II.A. In Section II.B, we examine whether past retail order imbalance measures can predict future stock returns using Fama-MacBeth regressions and long-short portfolios. In Section II.C, we compare alternative hypotheses for the predictive power of retail order imbalances for future stock returns. In Section II.D, we discuss the nature of the information contained in retail flow by linking it to Thomson Reuters News Analytics data.

A. What Explains Retail Investor Order Imbalances?

We start our empirical investigation by examining what drives the trading of retail investors. Specifically, we examine how retail investors' order flow is related to past order flow and past returns. To allow maximal time-series flexibility and focus on cross-sectional patterns, we adopt the Fama and MacBeth (1973) two-stage estimation. At the first stage, for each day, we estimate the following predictive regression:

$$oib(i,w) = b0(w) + b1(w)'ret(i,w-1) + b2(w)'controls(i,w-1) + u1(i,w), \tag{5}$$

imbalance measure for 2007 using our algorithm and found that the correlations between our measure and their measure ranged between 0.345 and 0.507 when using different subpenny ranges. For instance, 0.345 is the correlation of our measure and their measure using the number of shares for subpenny price within (0, 0.4) and (0.6, 1), while 0.507 is the correlation of our measure and their measure using the number of trades for subpenny price at 0.99 cent and 0.01 cent. These correlations should be less than one because their flow comes from only one wholesaler, while our measure comes from TRF, which covers nearly all retail order executions.

Electronic copy available at: https://ssrn.com/abstract=2822105

where we use various horizons of past weekly returns, $ret(i, w-1)$ and various control variables from the past to explain the order imbalance measure, $oib(i, w)$, for firm i during week w. The first-stage estimation generates a daily time-series of coefficients, $\{b0(w), b1(w)', b2(w)'\}$. At the second stage, we conduct statistical inference using the time-series of the coefficients. Because we use overlapping daily frequency data for weekly order imbalance and return measures, the standard errors are calculated using the Newey-West (1987) with six lags.[8]

To explain the order imbalance over week *w*, from day 1 to day 5, we first include its own lag, the past week order imbalance measure from day -4 to day 0, or *oib*(*w*-1). We also include past returns over three different horizons: the previous week (*ret*(*w*-1)), the previous month (*ret*(*m*-1)), and the previous six months (*ret*(*m*-7,*m*-2)). For control variables, we use log market cap, log book-to-market ratio, turnover (share volume over shares outstanding), and daily return volatility, all computed from the previous month's data.

The results are presented in Table II, with regressions I and II explaining the order imbalance measured in shares, and regressions III and IV explaining the order imbalance measured using the number of trades. In the first regression, the order imbalance using share volume, *oibvol*, has a positive correlation with its own lag, with a highly significant coefficient of 0.22, indicating that directional retail trading activity is somewhat persistent over successive weeks, as suggested in Chordia and Subrahmanyam (2004). The coefficients for the past week, past one month, and past six-month returns are -0.9481, -0.2778, and -0.0586, respectively. All three coefficients are negative and highly significant, which shows that retail investors are contrarian, for horizons ranging between one week and six months. The control variables indicate that investors tend to buy more aggressively in larger firms, growth firms, and firms with higher turnover and higher volatility. All coefficients are highly significant. The average adjusted R^2 from the first stage cross sectional estimation is about 6%.

We use different return and order imbalance measures for regressions II, III, and IV. At the weekly horizon, the results are similar across methods of computing returns and order imbalances. Henceforth, we focus our discussion on bid-ask midpoint returns, which do not have bid-ask bounce and thus exhibit a smaller degree of time-series predictability than returns based on

[8] The optimal lag number is chosen using BIC.

Electronic copy available at: https://ssrn.com/abstract=2822105

transaction prices. We also include CRSP returns in the results for the sake of completeness and robustness.

The negative coefficients on past returns match some of the findings in the literature. For example, retail traders are found to be contrarian in Kaniel, Saar, and Titman (2008) over monthly horizons, and by Barrot, Kaniel, and Sraer (2016) over daily and weekly horizons. In contrast, Kelley and Tetlock (2013) paint a more complex picture. They find that at weekly horizons, retail order imbalance measures are contrarian and have negative coefficients on past returns. Over shorter (daily) horizons, however, they find that market order imbalances actually have a positive coefficient on the lagged one-day return, which implies momentum rather than contrarian behavior.

Appendix Figure 2 Panel B plots the correlation of daily order imbalance with past returns for the previous one to 80 trading days. Similar to Kelley and Tetlock (2013), the correlation between the current retail order imbalance and the previous-day return is positive, indicating a momentum trading pattern on average. However, at lags between two days and 30 days, the average correlation is slightly negative. Our results are thus consistent with the findings of Kelley and Tetlock (2013) at short horizons and with those of other researchers at longer horizons. [9,10]

Our results in Table II reveal two important drivers affecting weekly order imbalance. The first is its own lag, which indicates that the retail order imbalance measures are persistent. The second are past returns, and we show a mixed result of both contrarian and momentum patterns, with the contrarian pattern prevailing at weekly horizons.

B. Predicting Future Stock Returns with Retail Order Imbalance Measures

B.1. Methodology and Overall Predictive Power

[9] Lee, Liu, Roll, and Subrahmanyam (2004) also find a mixed pattern of contrarian and momentum, using the overall market order imbalance. They find evidence that, after up-market moves, overall trades tend to follow a momentum pattern, while overall trades tend to be contrarian after down-market moves. We provide similar results using daily retail order flows in Internet Appendix Table I Panel A. When we use weekly retail order flows, the patterns both become contrarian, as shown in Internet Appendix Table I Panel B.

[10] In addition, we examine how firm-level order imbalance measures are related to firm fundamentals, as in Chordia, Huh, and Subrahmanyam (2007). The results shown in Internet Appendix Table I Panel E indicate that retail order imbalances are positively related to firm size, number of analysts, analyst dispersion, and leverage and negatively related to past return, firm age, and book-to-market ratio.

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

Can retail investors' activity provide useful information for future stock returns? In this section, we examine the predictive power of our order imbalance measures using Fama-MacBeth regressions as follows:

$$ret(i,w) = c0(w) + c1(w)oib(i,w-1) + c2(w)'controls(i,w-1) + u2(i,w), \tag{6}$$

where we use the retail order imbalance measure from the previous week, $oib(i,w-1)$, and various control variables to predict the next week's stock return, $ret(i,w)$, for firm i during week w. As in the previous section, because we use overlapping daily frequency data for weekly order imbalances and return measures, the standard errors of the time-series are adjusted using Newey-West (1987) with five lags. If past retail order imbalance can predict future returns in the right direction, we expect coefficient $c1$ to be significantly positive. If coefficient $c1$ turns out to be significantly negative, then retail investors might be making systematic trading mistakes, and if c1 is close to zero, we cannot reject the null that retail trading is uninformative on average about the cross-section of future stock returns.

We again include past returns as control variables, using three different horizons: the previous week, the previous month, and the previous six months (from month *m*-7 to month *m*-2). In addition, we include log market cap, log book-to-market ratio, turnover, and daily return volatility, all from the previous month. We report the estimation results in Table III. In regression I, we use the order imbalance based on share volume, *oibvol*, to predict the next week's return based on bid-ask midpoints. The coefficient on *oibvol* is 0.0009, with a *t*-statistic of 15.60. The positive and significant coefficient shows that, if retail investors buy more than they sell in a given week, the return on that stock in the next week is significantly higher. In terms of magnitude, we report at the bottom of the table that the inter-quartile range for the *oibvol* measure is 1.1888 per week. Multiplying the interquartile difference by the regression coefficient of 0.0009 generates a weekly return difference of 10.89 basis points (or 5.66% per year) when moving from the 25th to the 75th percentile of the *oibvol* variable. The same pattern is present when we use different order imbalance and return measures, and the weekly interquartile difference in the conditional mean return ranges from 9.31 basis points to 11.44 basis points (4.84% to 5.94% per year). That is, past week retail order imbalance can predict future returns in the correct direction.

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

For the control variables, we observe negative coefficients on the previous week's return, which indicates weekly return reversals, and positive coefficients on the other longer-horizon returns, which indicates momentum. Size, book-to-market, turnover, and volatility all carry the expected signs, and most are not statistically significant. This result also confirms that the predictability we find is not simply a manifestation of some other size, book-to-market, turnover, or volatility anomaly. The average adjusted R^2's from the first stage cross-sectional estimation are mostly around 3.85%.

B.2. Subgroups in the Cross Section

Our sample includes on average more than 3,000 firms each day. Is the predictive power of retail investor order imbalances restricted to a particular type of firm? Do informed retail investors have preferences for particular types of firms? We investigate these questions by analyzing various firm subgroups in this section. We first sort all firms into three groups based on a firm or stock characteristic observed at the end of the previous month. Then, we estimate equation (6) within each characteristic group. That is, we allow all coefficients in equation (6) to be different within each group, which allows substantial flexibility in the possible predictive relationship across these different groups.

To save space, we include only the results on weekly returns that are computed using the end-of-day bid-ask average price. We first sort all stocks into three different size groups based on market capitalization: small, medium, and large. The results are reported in Panel A of Table IV. In the left panel, we report coefficients on *oibvol*, the order imbalance computed from share volume. When we move from the smallest one-third of firms by market cap to the largest tercile, the coefficient on *oibvol* decreases from 0.0013 to 0.0003, and the *t*-statistic decreases from 13.90 to 3.68. Clearly, the predictive power of retail order imbalance is much stronger for smaller firms than for larger-cap firms, but the predictability remains reliably present in all three groups. Economically, the interquartile difference in weekly returns is 21.9 basis points for the smallest firms (11.39% per year), and 2.6 basis points for the largest firms (1.35% per year). The results in the right panel using order imbalance based on the number of trades (*oibtrd*) are quite similar.

In Panel B of Table IV, we sort all firms into three groups based on the previous month-end share price. In the left panel, moving from the lowest share-price firms to the highest, the coefficient on *oibvol* decreases from 0.0014 to 0.0002, and the *t*-statistics go from 13.34 to 3.23.

Electronic copy available at: https://ssrn.com/abstract=2822105

In terms of magnitude, the interquartile weekly return difference is 20.5 basis points (10.66% per year) for the lowest-price firms and only 2.0 basis points for the firms with the highest share price (1.04% per year). The results are similar for specifications using *oibtrd*, reported in the right panel, with slightly lower coefficients and *t*-statistics. The pattern is clear: the predictive power of retail investor order imbalances for future returns is stronger for low-price firms.

Next, we sort all firms based on previous-month turnover, which is a proxy for liquidity. In the left panel, moving from the tercile of low trading activity to the firms with more turnover, the coefficient on *oibvol* decreases from 0.0011 to 0.0007, and the *t*-statistic decreases from 15.60 to 4.98. In terms of magnitude, the interquartile weekly return difference is 20.5 basis points (10.66% per year) for the firms with the lowest turnover and 6.5 basis points for the firms with the highest turnover (3.38% per year). For specifications based on *oibtrd* in the right panel, the results are similar, with slightly lower coefficients and *t*-statistics. Overall, retail investor order imbalances better predict returns for firms with lower trading activity.

In this section, we find that the predictive power of the retail investor order imbalance is significant and positive for all but one subgroup, which shows that the predictive power is not driven by special subgroups. However, a clear cross-sectional pattern for the predictive power is observed. The predictive power of the retail order imbalance is much stronger for small firms and firms with low share-price and low liquidity.

B.3. Longer Horizons

The results in the previous section show that retail investor order imbalances can predict next week's returns positively and significantly. It is natural to now ask whether the predictive power is transient or persistent. If the predictive power quickly reverses, retail investors may be capturing price reversals; if the predictive power continues over time and then vanishes beyond some horizon, retail investors may be informed about information relating to firm fundamentals. To answer this question, we extend equation (6) to longer horizons as follows:

$$ret(i,w+k) = c0(w) + c1(w)oib(i,w) + c2(w)'controls(i,w) + u3(i,w+k). \tag{7}$$

That is, we use one week of order imbalance measures to predict k-week ahead returns, *ret*(*i*,*w*+*k*), with k=1 to 12. To observe the decay of the predictive power of retail order imbalance, the return

Electronic copy available at: https://ssrn.com/abstract=2822105

to be predicted is a weekly return over a one-week period, rather than a cumulative return over *n* weeks, which is an average over all weeks involved. If retail order imbalances have only short-lived predictive power for future returns, we might observe the coefficient $c1$ decrease to zero within a couple of weeks. Alternatively, if the retail order imbalance has longer predictive power, the coefficient $c1$ should remain statistically significant for a longer period. In our empirical estimation, we choose *k* ranging from two to 12 weeks.

We report the results in Table V, with results based on bid-ask average returns in Panel A, and those based on closing transaction prices in Panel B. In Panel A, when we extend the window from two to 12 weeks, the coefficient on *oibvol* monotonically decreases from 0.00055 to 0.00007, and the coefficient on *oibtrd* gradually decreases from 0.00048 to 0.00006. The coefficients are statistically significant up to six or eight weeks ahead. The results in Panel B are similar. There is no evidence of price reversals at any horizon. Thus, the retail order imbalances potentially capture longer-lived information.

B.4. Long-Short Portfolios

One might wonder whether we can use retail order imbalances as a signal to form a profitable trading strategy. As discussed earlier, both *oibvol* and *oibtrd* are publicly available information. In this section, we form quintile portfolios based on the previous week's average order imbalance and then hold the quintile portfolios for up to 12 weeks. If retail investors on average can select the right stocks to buy and sell, then firms with higher or positive retail order imbalance would outperform firms with lower or negative order imbalance. Notice that this exercise uses retail order imbalance measures merely as a signal to predict future stock returns, and it thus provides no information on whether retail investors make profits from their own trades. We ignore trade frictions and transaction costs here, and the results are therefore not definitive on whether outsiders can profit from these signals.

Table VI reports long-short portfolio returns, where we buy the stocks in the highest order imbalance quintile and short the stocks in the lowest order imbalance quintile each day using the previous 5-day retail order flow measures, and hold them for the next few weeks. Portfolio returns are value-weighted using the previous month-end market cap. Because the holding period can be as long as 12 weeks, we report both the raw and risk-adjusted returns using the Fama-French three-factor model. Given the usage of overlapping data, we adjust the standard errors of the portfolio

Electronic copy available at: https://ssrn.com/abstract=2822105

return time-series using Hansen and Hodrick's (1980) standard errors with the corresponding number of lags.[11]

In Panel A, the long-short strategy is based on the previous week's *oibvol*, and we report bid-ask average returns. Over a one-week horizon, the long-short portfolio return is 0.092%, or 4.78% per year annualized. The *t*-statistic is 2.66. Risk adjustment using the Fama-French three-factor model does not make much difference: the weekly Fama-French alpha for the long-short portfolio is 0.084%, with a *t*-statistic of 2.43. When we increase the holding horizon to 12 weeks, the mean return becomes 0.588%, with a *t*-statistic of 2.09. The general pattern is that holding-period returns (and alphas) continue to grow at a decreasing rate over time. We observe no evidence of a reversal in returns. In terms of statistical significance, the *t*-statistics are significant or marginally significant up to the 12-week horizon. These results are slightly weaker than those of the Fama-MacBeth regressions, mainly because, in this section, we value-weight the portfolio returns across firms, while the Fama-MacBeth approach implicitly weights each stock equally.

When we restrict portfolio formation to one of the three market cap groups, the one-week return is 0.403% (or 20.96% per year) with a *t*-statistic of 9.16 for the smallest firms, while the one-week return is 0.067% (or 3.48% per year) with a *t*-statistic of 1.78 for the largest firms. When the holding horizon becomes longer, the return on the long-short strategy is still significant and positive for up to 12 weeks for the smallest third of firms, but the results are statistically insignificant for the largest tercile. The results in Panel B, obtained using *oibtrd*, are qualitatively similar but with smaller magnitude and lower statistical significance. This result is expected since, as mentioned, the information provided by *oibvol* is similar but finer than that provided by *oibtrd*.[12]

To make sure that the statistical significance in return differences is not driven by particular sample periods, we provide a time-series plot of the return differences between quintiles 1 and 5 in Figure 3, where the portfolios are sorted on *oibvol* and the holding period is one week. Over our

[11] For example, for one-week holding period portfolio, we use Hansen and Hodrick (1980) with 5 lags; if two-weeks holding period portfolio, we use Hansen and Hodrick (1980) with 10 lags, etc.

[12] We also conduct a rough calculation that includes transaction costs. Frazzini, Israel and Moskowitz (2018) state that a reasonable estimate of one-way transaction cost on value-weighted US stocks is about 12 basis points for the period January 2006 to June 2016. To be conservative, we assume that for each rebalance, we change 100% of the positions. That is, each rebalance we incur a 2*12 bps = 24 bps rebalance cost. For instance, for a weekly rebalance or 1 week holding period, each year's transaction cost would be 52 rebalances * 2 * 12bps = 1248 bps. After this drastic transaction cost adjustment, the mean returns and alphas remain positive and significant for the small firms over all holding horizons. For the medium and big firms, the mean returns and alphas stay positive for longer holding periods, but they are mostly insignificant.

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

six-year sample period, we observe both time-variation in the return differences and positive and negative spikes. However, most data points are positive, and the positive returns are not driven by particular sample subperiods. Unreported plots of alphas show the same pattern.

C. Alternative Hypotheses for Retail Order Imbalance Predictive Power for Future Returns

The predictive power of retail order imbalances for future stock returns is consistent with three hypotheses. First, as in Chordia and Subrahmanyam (2004), order flows are persistent, and, as the buying/selling pressure is persistent, this could lead directly to the predictability of future returns. Second, as in Kaniel, Saar, and Titman (2008), the retail traders are contrarian at weekly horizons, and their contrarian trading provides liquidity to the market, and thus their trades might positively predict future returns. Third, as in Kelley and Tetlock (2013), retail investors, especially the aggressive investors using market orders, may have valuable information about the firm, and thus their trading could correctly predict the direction of future returns. The above three hypotheses are not exclusive. In Section II.C.1, we conduct a simple decomposition to separate alternative hypotheses. In Section II.C.2, we provide more evidence regarding the liquidity provision hypothesis.

C.1. Two-Stage Decomposition

To distinguish among these alternative hypotheses, we adopt a two-step decomposition procedure. In the first step, we decompose the weekly retail order imbalance into three components for each week *w*, with the following cross-sectional regressions:

$$oib(i,w) = d0(w) + d1(w)oib(i,w-1) + d2(w)'ret(i,w-1) + u4(i,w). \tag{8}$$

After we obtain the time-series of coefficients, $\{\widehat{d0}(w), \widehat{d1}(w), \widehat{d2}(w)'\}$, we define the following terms:

$$\begin{aligned} \widehat{oib}_{i,w}^{persistence} &= \widehat{d1}(w)oib(i,w-1), \\ \widehat{oib}_{i,w}^{contrarian} &= \widehat{d2}(w)'ret(i,w-1), \\ \widehat{oib}_{i,w}^{other} &= \widehat{u4}(i,w). \end{aligned} \tag{9}$$

That is, we denote the part related to the past order imbalance as the “persistence,” which is related to the price pressure hypothesis. The part related to past returns over different horizons is denoted as “contrarian,” which relates to the liquidity provision hypothesis. After we take out predictability

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

due to "persistence" and "contrarian" trading, we denote the residual part as "other," which we attribute to retail investors' information about future returns.

At the second stage, we estimate the following regression using the Fama-MaBeth methodology, which is parallel to equation (6):

$$ret(i,w) = e0(w) + e1(w)\widehat{oib}_{i,w-1}^{persistence} + e2(w)\widehat{oib}_{i,w-1}^{contrarian} + e3(w)\widehat{oib}_{i,w-1}^{other} + e4(w)'controls(i,w-1) + u5(i,w). \tag{10}$$

Since we decompose the original order imbalance measure into three parts, related to order flow persistence, a contrarian trading pattern, and the residual, the coefficient estimates in equation (10) reveal how each component helps to predict future stock returns.

We report the decomposition results in Table VII. Panel A presents the first-stage estimation as in equation (8), which is quite similar to those reported in Table II. Take the first regression as an example. The order imbalance measure, *oibvol*, has a highly significant and positive coefficient on its own lag at 0.22, which indicates order persistence. In terms of past returns, the coefficients for the past week, past month, and past six-month returns are -0.9286, -0.2029 and -0.0267, respectively, all implying contrarian trading patterns.

After we decompose the previous week's order imbalance into "persistence," "contrarian," and "other," we use them separately and together to predict future stock returns, as in equation (10). In the first regression, we use the past week's *oibvol* to predict future bid-ask return. The coefficient estimate on *oib (persistence)* is 0.0027, with a t-statistic of 8.75, which implies that price pressure significantly and positively contributes to the predictive power of the retail flow. The coefficient estimate on *oib (contrarian)* is -0.0044, and it is insignificantly different from zero, implying that we cannot reject the null hypothesis that the contrarian component does not contribute to the predictive power of retail order imbalances. Finally, for the *oib (other)* component, the coefficient is 0.0008, with a significant t-statistic of 14.47.[13]

In terms of economic magnitude, we compute the interquartile range of all three components of the order imbalance measure. For the oib *(persistence)*, if we move from the 25th

[13] We also try to include the past order imbalance as a control variable for the second stage estimation. We cannot directly include oib(w-1) or oib(w-2), because it will create collinearity issues. Therefore, here we include either oib(w-3) or oib(m-1) to control for past oib. These results are presented in Appendix Table II. No matter which specification we use, the main results stay quite similar to those in Table VII.

Electronic copy available at: https://ssrn.com/abstract=2822105

percentile firm to the 75th percentile firm, the difference in future one-week stock return is 0.0688% (3.58% per year). For the *oib (other)* variable, if we move from the 25th percentile firm to the 75th percentile firm, the difference in future one-week stock return is 0.0915% (4.76% per year). For the *oib (contrarian)* measure, the sign is the opposite and has no statistical significance. The results in other specifications are quite similar.

Our decomposition exercise shows that close to half of the predictive power of the retail order imbalance comes from the persistence of the order imbalance measures,[14] and that most of the rest comes from the residual component, after we take out order persistence and the contrarian trading pattern. Since this residual component significantly predicts future stock returns, it is consistent with the hypothesis that retail investor trading contains valuable information about future stock price movements.

C.2. A Close Look at the Liquidity Provision Hypothesis

The liquidity provision hypothesis receives substantial attention in the existing literature, so here we take a closer look at this hypothesis. Kaniel, Saar, and Titman (2008) argue that retail investors' contrarian trading provides liquidity to the market, and this leads to the positive predictive power of past retail order imbalance for future stock returns. Therefore, in equation (8), we use the part of the retail order imbalance related to past returns, *oib (contrarian)*, as a proxy for the "liquidity provision" hypothesis. Now our results in Table VII show that the contrarian component of retail order flow cannot significantly predict future stock returns. Does this finding completely rule out the "liquidity provision" hypothesis for the predictive power of retail order flow? We are afraid not. We can only rule out the liquidity provision hypothesis, under the assumption that the contrarian trading pattern captured by *oib (contrarian)* is a perfect proxy for the liquidity provision hypothesis. This seems to us to be a reasonable assumption, but as far as we can tell it cannot be directly confirmed by any data that we observe.[15] In this subsection, we

[14] To be more specific, the retail order imbalance has a low autocorrelation coefficient between 10-20%, but the positive autocorrelation lasts for a long period. Here the persistence refers to the long horizon rather than the magnitude.

[15] For example, recent studies, such as Arif, Ben-Rephael and Lee (2016), and Chakrabarty, Moulton, and Trzcinka (2017), show that directional trading by active funds is highly persistent and price destabilizing. If the retail trades provide liquidity to these active funds, then liquidity provision can also go through the persistence channel, rather than the contrarian channel.

Electronic copy available at: https://ssrn.com/abstract=2822105

provide more results regarding the liquidity provision hypothesis through different approaches beyond the predictive regression.

An important piece of evidence in support of the liquidity provision hypothesis in Kaniel, Saar, and Titman (2008) is the contemporaneous relation between retail order imbalance and stock returns. [16] To be more specific, Kaniel, Saar, and Titman (2008) examine the past, contemporaneous, and future returns of intense buy and sell portfolios of retail investors. In their paper, the buy and sell order flows by retail investors are measured using the "net individual trading" (NIT) measure. For each week, all firms are first sorted into decile groups using the previous-week NIT, and then Kaniel, Saar, and Titman (2008) track the excess returns to these different groups for the four weeks before and after the portfolio construction. The excess returns of each portfolio is computed by subtracting the return on a market proxy (the equal-weighted portfolio of all stocks in the sample). Here we follow their approach, while using our retail order flow measures, *oibvol* and *oibtrd*. Results using oibvol are reported in Table VIII, and results using oibtrd are reported in Appendix Table III.

The main results of Kaniel, Saar and Titman (2008) are reported in their Table III, which contains three main findings. First, the stocks the retail investors sell during the portfolio construction week (week 0), the intense selling group, experience significantly positive excess return, before week 0; while the stocks the retail investors buy during week 0, the intense buying group, experience negative excess returns. This is a typical contrarian trading pattern of selling winners and buying losers. In Panel A of Table VIII, the first row contains the firms intensely sold by the retail investors, and the mean excess return in the 20 days prior to the selling week is 0.67%. The bottom row contains the firms intensely bought by retail investors, and the mean excess return on these stocks in the 20 days prior to the selling week is -1.29%. Both numbers are highly significant, and confirm Kaniel, Saar and Titman's first finding.

The second finding of Kaniel, Saar and Titman (2008) is that after retail investors buy or sell, the stocks the retail investors sell during week 0, the intense selling group, experience negative excess returns, while the stocks the retail investors buy during week 0, the intense buying group, experience positive excess returns. This shows that retail trading can predict returns in the correct direction. In Panel A of Table VIII, we find that firms intensely sold by retail investors (in the first

[16] We thank an anonymous referee for this suggestion.

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

row) experience a mean excess return in the 20 days after the selling week of -0.30%, while the firms intensely bought by retail investors in the bottom row experience a mean excess return of 0.57%. Again, both numbers are highly significant, and confirm Kaniel, Saar, and Titman's second finding.

Finally, for the contemporaneous relation over week 0, Kaniel, Saar, and Titman (2008) find that the contemporaneous excess return is significantly *positive* for stocks retail investors *sell*, and *negative* for stocks they *buy*. Since the return signs are opposite of the retail trading direction, they interpret this finding in favor of the liquidity provision hypothesis. From the column of k=0 in Table VIII Panel A, however, we find that for firms intensely sold by retail investors, the contemporaneous return is significantly negative at -0.24% with a t-statistic of -5.30. For the intensely bought firms in the bottom row, the contemporaneous return is significantly positive at 0.11% with a t-statistic of 2.69. Our findings show consistent, rather than opposite, signs between contemporaneous retail trading and return direction, which does not line up with the liquidity provision hypothesis proposed in Kaniel, Saar, and Titman (2008).

What might cause the differences in our results? It might come from differences in the sample period or in terms of coverage. In terms of the sample period, the Kaniel, Saar and Titman (2008) sample is January 2000 through December 2003, and our sample is January 2010 through December 2015, which are about ten years apart. Coverage wise, Kaniel, Saar and Titman (2008)'s sample is from NYSE's Consolidated Equity Audit Data (CAUD) which contains only retail trades that are executed on that exchange. During the Kaniel, Saar and Titman (2008) sample period, only a small number of brokerages sent their retail order flow to the NYSE. As a result, the NYSE's market share of overall retail order activity was (and has remained) quite small. In comparison, our sample is from TAQ which contains all trades. But our retail order flow data only captures the part related to market orders, not limit orders, which on average accounts for 6.91% of the total trading volume in each stock each day. To summarize, the liquidity provision hypothesis receives at most mixed support in our data sample.

D. Public News and Retail Order Imbalance

Our earlier results indicate that retail investor order flows may contain valuable information about future stock price movements, which might be surprising to many. As Kaniel, Saar, and Titman (2008) note, "… *it is unclear how individuals, who have far fewer resources than*

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

institutions, could gain the upper hand in discovering private information and trading on it profitably in such a widespread fashion."

Therefore, to better understand whether the retail investors can be informed traders and the nature of information they might possess, we examine the relation between retail order flow and public news in this section. We introduce the news data in Section D.1, and investigate whether the information in retail flow is related to public news in Section D.2.

D.1 Retail Order Imbalance and Future Returns across News Topics

We obtain news data from Thomson Reuters News Analytics (TRNA), which contains prominent public news articles for a broad set of firms starting from 2003. TRNA provides key information about each news item, such as the ticker, the time stamp of the news story, the news topics the story belongs to, and sentiment scores for each article. News topics are grouped into five categories: cross market, general news, economy, equities, and money/debt. Each category contains several news subtopics, and we collect 58 such subtopics in our sample. The sentiment score measures the probabilities of the article being positive, negative, or neutral, computed using Thomson Reuters' proprietary algorithm. We compute a net sentiment score as the difference between the positive and negative sentiment score for each stock each day. The news data are available from January 2010 to December 2014, which covers most of our main sample. We use tickers to match the news data with our retail order imbalance data, generating a merged sample of 3,854,813 stock-day observations.

We first provide some simple statistics for the relation among news, returns, and retail order flow. To examine whether the public news can predict future stock returns, we estimate the following Fama-MacBeth regression:

$$ret(i,w) = f0(w) + f1(w) \times sent(i,w-1) + f2(w)'controls(i,w-1) + u6(i,w). \quad (11)$$

Here, variable $sent(i,w-1)$ is the average TRNA net sentiment score for firm *i* during week *w*, calculated by averaging non-missing news sentiment for firm *i* within week *w*-1. The results are reported in regression I and II of Table IX Panel A. In regression I, the coefficient of past week public news sentiment is 0.0008 with a t-statistic of 3.31. The positive and significant coefficient indicates that the public news can predict next week's stock returns. When we include the past retail order imbalance in regression III and IV, the predictive power of the public news sentiment stays about the same.

Electronic copy available at: https://ssrn.com/abstract=2822105

To understand how retail order imbalances are related to public news, we estimate the contemporaneous[17] relation between the two using the following Fama-MacBeth specification:

$$sent(i,w) = g0(w) + g1(w) \times oib(i,w) + g2(w)'controls(i,w-1) + u7(i,w). \qquad (12)$$

We find that the current week's retail order imbalance is significantly related to the same-week public news sentiment when there is news in 10 out of the 58 subtopics. To save space, we present the coefficient estimates for the 10 cases in Panel B of Table IX. These 10 subtopics represent about 38% of total news days, and they mostly contain firm-level news. For instance, for the subtopic "RESF" (results forecast) in the news type "equities", the coefficient g1 is 0.0054, with a significant t-statistic of 3.90, indicating that the retail order imbalance has a positive and significant contemporaneous relation with news related to forecasts of company results. Out of the ten subtopics, four are from the category of "money and debt", and three are from the category of "equity", with the highest two significant t-statistics for the subtopics "results forecast" and "debt rating news". Interestingly, the retail order imbalances are never statistically significantly correlated with the "economy" type of news. This finding implies that retail investors may have valuable information at the firm level rather than at the market level.

D.2 Public Information and Other Information

The above results show that retail order imbalances are associated with some types of public news, particularly firm-level news. In this subsection, we probe deeper into the fraction of retail investors' predictive power is associated with these public news releases, because it is also possible that retail traders possess and trade on private information that eventually makes its way into prices, but not via an identifiable news release.

We investigate this issue empirically using a two-step decomposition procedure similar to that in Section II.C. In the first step, we estimate a Fama-MacBeth regression and decompose the weekly order imbalance into four components, as follows:

$$oib(i,w) = g0(w) + g1(w)oib(i,w-1) + g2(w)'ret(i,w-1) + g3(w)sent(i,w) + u7(i,w). \qquad (13)$$

[17] In Appendix Table IV, we also examine whether retail order imbalance can directly predict public news. We fail to find evidence that retail order flow can predict future public news.

Electronic copy available at: https://ssrn.com/abstract=2822105

Here, variable $sent(i,w)$ is the average TRNA net sentiment score for firm *i* during week w, which we use to capture information in contemporaneous public news releases. After we obtain the time-series of coefficients, $\{\widehat{g0}(w), \widehat{g1}(w), \widehat{g2}(w)', \widehat{g3}(w)\}$, we define the following terms:

$$\begin{aligned}
\widehat{oib}_{i,w}^{persistence} &= \widehat{g1}(w)oib(i,w-1), \\
\widehat{oib}_{i,w}^{contrarian} &= \widehat{g2}(w)'ret(i,w-1), \\
\widehat{oib}_{i,w}^{publicnews} &= \widehat{g3}(w)sent(i,w), \\
\widehat{oib}_{i,w}^{other} &= \widehat{u7}(i,w).
\end{aligned} \tag{14}$$

As before, we denote the part related to past order imbalance as the "persistence" component, which is related to the price pressure hypothesis; the part related to past returns is denoted as the "contrarian" component, which is connected to the liquidity provision hypothesis. We define the part related to contemporaneous public news sentiment as the "public news" component. Finally, we denote the residual part as the "other" component, which we attribute to retail investors' private information that is not incorporated into prices via an identifiable news release.

At the second stage, we estimate the following regression using the Fama-Macbeth methodology, which is parallel to equation (6):

$$ret(i,w) = h0(w) + h1(w)\widehat{oib}_{i,w-1}^{persistence} + h2(w)\widehat{oib}_{i,w-1}^{contrarian} +$$
$$h3(w)\widehat{oib}_{i,w-1}^{publicnews} + h4(w)\widehat{oib}_{i,w}^{other} + h5(w)'controls(i,w-1) + u8(i,w). \tag{15}$$

Since we decompose the original order imbalance measure into four parts, related to persistence, contrarian trading pattern, public information, and the residual, the coefficients in equation (15) reveal how each component helps to predict future stock returns.

Notice that for the first stage estimation, the public news component is derived from a contemporaneous relation between the current news and current retail order flow, rather than past news. From the perspective of the empirical design, we can link the retail order imbalance with the past, contemporaneous, or future public news, but the interpretations would be different. If we use future public news, the interpretation would be whether and how retail order flow "anticipates" future public news. If we use past public news, the interpretation would be that previously "incorporated" public news can be a component of the retail order imbalance. When we use

Electronic copy available at: https://ssrn.com/abstract=2822105

contemporaneous public news sentiment, we interpret the related part of retail order flow as contemporaneously "processed" public news. Here we choose not to use future public news, because if we project *oib(w-1)* on *sent(w)* at the first stage, then $\widehat{oib}_{i,w}^{publicnews}$ would capture news from week *w*, and would have a mechanical correlation with *ret(w)*, which is the dependent variable in the second stage estimation. We also choose not to use past public news, because we would like to maximize the explanatory power of public news for retail order flow, while contemporaneous public news sentiment very likely nests the information in the past public news.

Table X Panel A provides results for the first-stage decomposition. The patterns of how past retail order imbalance and past returns affect the current order imbalance are very similar to those in Table II. The coefficient on the contemporaneous sentiment ranges between 0.0249 and 0.0305, all with t-statistics higher than 10. This clearly indicates that more positive news is associated with more contemporaneous purchases by retail investors. The average adjusted R^2 for the first stage estimation are mostly between 5.50% and 8.6%.

Panel B of Table X reports the results on the second-stage decomposition. From the top half, the coefficients on oib(persistence) are positive and highly significant, and the coefficients on oib(contrarian) are mostly insignificant, similar to the findings in Table VII. Most importantly, the coefficients on the public news components of order imbalance, $\widehat{oib}_{i,w-1}^{publicnews}$, are also all insignificant, indicating that the contemporaneous public news component of retail order imbalances do not help to predict future returns significantly. In contrast, the "other" component of the retail order imbalance measure is always positive and significant in the regressions. For instance, in the first regression, it has a coefficient of 0.0008 with a highly significant t-statistic of 13.98. This result is consistent with the hypothesis that retail investors trade on private information that is not incorporated into prices via public news releases. The bottom half panel shows that when we move from the 25th percentile firm to the 75th firm, the "other" component of the retail order flow accounts for 0.07% to 0.10% of weekly return differences, which is more than half of the return difference that the retail order imbalance can generate overall. These results suggest that the predictive power of retail investors' order imbalance is likely not related to an identifiable public news release.

Electronic copy available at: https://ssrn.com/abstract=2822105

Returning to the interesting question raised at the beginning of this subsection, how can retail investors, with far fewer resources than institutions, get the upper hand in discovering private information? Here we offer two possible explanations.

First, as an investor group, retail investors can be heterogeneous. For instance, it is possible that some individual investors might simply be endowed with private and valuable firm-specific information. These individuals might work in the same industry or for a customer or supplier and might naturally obtain value-relevant information in this way. They may also have some resources for information discovery. For example, it may be possible for individuals to study the parking lots of retailers to assess demand growth. If other retail investors are uninformed and simply add noise in their order flow, the net trading of individual investors could still predict future price movements. Unfortunately, our current data does not contain identities of the retail investors, and we cannot directly check the heterogeneity of the retail investors.

Second, our data only contains retail market orders, but not retail limit orders. Retail limit orders could have negative information, which would at least partially offset our findings for retail market orders. In addition, it is not clear that the counterparties to retail orders are necessarily "better informed" institutional investors. For example, Chakrabarty, Moulton, and Trzcinka (2017) document the presence of uninformed "short-term" institutional investors as a non-trivial part of the market.

III. Further Discussion

Retail order imbalances can predict future stock returns. This predictive ability lasts up to eight weeks and is stronger for smaller and lower-priced firms. In this section, we discuss several related issues to put the retail order imbalance's predictive power in perspective. In Section III.A, we discuss whether retail investors' trading can predict the market's overall movement, and we examine whether the predictive power is related to overall market conditions in Section III.B. We investigate the predictive power of odd lot retail orders in section III.C. Retail trades occur with different sizes, and we examine the predictive power of large vs. small trade sizes in Section III.D. It is important to understand the role of wholesalers in this setup. Thus, Section III.E examines the magnitude of price improvement and the profitability of interacting with retail order flow. We identify the nature of the information captured by retail order flows by linking retail order imbalances to earnings news in Section III.F. Finally, we examine whether retail order imbalances

Electronic copy available at: https://ssrn.com/abstract=2822105

can still predict future returns if we control for overall market order imbalances in Section III.G. To save space, all returns in this section are bid-ask returns.

A. Aggregate Retail Order Imbalance

If retail order imbalances can predict future stock returns in the cross section, retail investors may also be able to predict aggregate market moves. To investigate this possibility, we aggregate retail order imbalances across all firms to predict aggregate stock market returns. We estimate the following equation:

$$mkt(w+1, w+k) = m0 + m1 \times aggoib(w) + u9(w+1, w+k), \qquad (16)$$

where $mkt(w+1,w+k)$ is the future k-week cumulative market return from week $w+1$ to week $w+k$, and $aggoib(w)$ is the current aggregated retail order imbalance measure for week w. We compute *aggoib* using either value-weighted or equal-weighted *oibvol* or *oibtrd* measures. The results are shown in Table XI Panel A. They are the same regardless of the weighting scheme or order imbalance measure: There is no evidence that retail investors can reliably predict future market returns.

Our approach can also be used to identify the retail trading of exchanged-traded funds (ETFs). In Table XI Panel B, we examine retail order flow in a large cross section of ETFs over the same time period. In cross-sectional predictive regressions of the form in equation (6), the coefficient is mostly around or below one basis point, which is much smaller than the comparable coefficients shown in Table III, and the t-statistics are mostly insignificant. This result suggests that retail traders cannot predict sector returns or overall equity market returns. To separate sector-oriented information from broader market-wide information, we select the six largest ETFs that focus on the overall U.S. equity market by tracking comprehensive U.S. equity indexes: SPY, IVV, VTI, VOO, IWM, and IWB. The results are reported in the last row of Panel B. Consistent with the market timing results in Panel A, we find little evidence that retail order flow can predict future returns on broad equity market ETFs.

B. Market Conditions

Barrot, Kaniel, and Srear (2016) find that retail trades contain more information when markets are volatile, specifically when the VIX option-implied volatility index is high. Their sample spans from 2002 to 2010, during which the VIX experiences dramatic changes. In contrast,

Electronic copy available at: https://ssrn.com/abstract=2822105

our sample period is 2010 to 2015, where the VIX is far less volatile. Nevertheless, we divide our sample in half: one portion when VIX is higher than the historical median of 18% and the other when the VIX is below this historical median.

We re-estimate equation (6) for the high- and low-VIX subsamples. The results are presented in Panel C of Table XI. Comparing the low- and high-VIX regimes, we find that the coefficient on *oibvol* is quite similar, yet the *t*-statistic is higher when VIX is low than when it is high. This result might not be surprising, given that the volatility of all variables increases when VIX is high. Overall, the predictive power in both high- and low-VIX regimes is positive and significant.

C. Odd Lots

In this section, we investigate the behavior of odd lot retail trades over the post-December 2013 period when odd lot transactions are reported to the consolidated tape. Can odd lot retail order flow predict future firm-level returns? We estimate regression (6) using odd lot retail order imbalances and present the results in Table XI Panel D. Both coefficients are positive but not statistically significant. In unreported results, we find that daily odd lot order imbalance measures can significantly predict returns for the next trading day but not at longer horizons than that. We conclude that the odd lot retail order imbalance measure's predictive power is much weaker than that of the overall retail order imbalance.

D. Retail Order Sizes

As Figure 1 Panel A shows, a median market order submitted by a retail investor is around \$7,000. The median retail trade is about 400 shares. The "stealth trading" literature argues that medium-size orders are more likely to be informed and that large orders are usually broken into smaller orders.

To determine if information content differs according to order size, we partition the orders into large vs. small groups using 400 shares as the cutoff, and we estimate the predictive regression for each group separately. The results are reported in Table XI Panel E. We find that both large and small orders predict future stock returns, but the larger orders' predictive power is stronger. Our results suggest that more informed retail investors may demand immediacy by using larger market orders and that stealth trading does not seem to characterize the trading of retail investors.

Electronic copy available at: https://ssrn.com/abstract=2822105

E. Wholesaler/Internalizer's Perspective: Profitability of Marketable Retail Order Flow

If marketable retail order flow is sufficiently informed, trading with these orders would be unprofitable. This might raise the question of whether our results are consistent with the apparently profitable business model of internalizers and wholesalers. Ultimately, as long as the information content of retail order flow is less than the bid-ask spread being charged, internalizers and wholesalers on average can still earn positive revenues by trading with these orders. For example, if a retail buy and a retail sell order arrive at the same time, they offset each other, and a wholesaler earns the full bid-ask spread charged (the quoted spread less the price improvement given). Ultimately, internalizers and wholesalers are only exposed to adverse selection on retail order imbalances. The summary statistics in Table I show that there is a substantial amount of offsetting retail order flow. The interquartile range for the volume-based daily order imbalance measure is from -0.301 to 0.217, indicating that, even at the ends of these ranges, more than two-thirds of the retail order flow in such a stock on a given day is offsetting buys and sells.

To get a better sense of the profitability of interacting with retail order flow, we compute standard microstructure information-content measures for the retail trades in our sample. Specifically, we calculate proportional effective spreads, one-minute price impacts, and one-minute realized spreads for all retail buys and sells during 2015. Realized spreads are a standard proxy for trading revenue earned by a liquidity provider such as a wholesaler. We apply standard data filters, eliminating all trades where effective spreads exceed $1, and we calculate dollar-volume weighted averages across all stocks. We find that the mean effective half-spread is 16 basis points. The one-minute price impact is four basis points, leaving a realized half-spread of 12 basis points. In other words, interacting with our identified retail order flow appears to be profitable (at least before other costs) for the wholesalers/internalizers because the bid-ask spreads are sufficiently large. The liquidity provider (in this case, the wholesaler or internalizer) loses about four basis points (the price impact) of the bid-ask spread to short-term price moves, but this leaves about 12 basis points (the realized spread) of the bid-ask spread as average trading revenue to the liquidity provider. Note that the realized spread is a very crude measure of trading revenue. Furthermore, we cannot measure payments made by wholesalers to introducing brokers, nor can we measure the other costs associated with a wholesaling or internalization operation. However, these realized spreads are considerably higher than the realized spreads associated with on-exchange transactions, so we feel comfortable in concluding that the price-improvement business

Electronic copy available at: https://ssrn.com/abstract=2822105

model is quite profitable for wholesalers and internalizers who can successfully segment order flow.

We can also examine some of the segmentation that is performed by these liquidity providers. For instance, the magnitude of price improvement is chosen by the internalizers/wholesalers. They can rationally incorporate the potential information embedded in retail orders and offer price improvement only up to the point at which they can still profit from the trade. That is, on the one hand, if they infer there might be relevant information in the retail order, they might offer less price improvement, and on the other hand, if they conclude that the retail order is unlikely to contain relevant information, they might be willing to offer more price improvement. If this is true, the predictive power of retail order imbalances should be higher for retail trades with less price improvement.

In the earlier sections, we group all orders with subpenny prices between 0.6 and 1 as retail-initiated buy orders and group those between 0 and 0.4 as retail-initiated sell orders. In this section, we further divide orders into "less price improvement" and "more price improvement" types. For transactions with less improvement, we define buyer-initiated trades as transactions with prices between 0.8 and the round penny, and seller-initiated trades as trades with transaction prices between the round penny and 0.2 cents. For the "more price improvement" category, we define buyer-initiated trades as trades with transaction prices between 0.6 and 0.8, and seller-initiated trades as trades with transaction price between 0.2 and 0.4. We compute retail order imbalances following equation (1) and (2). We compare the predictive power of retail order imbalances for "more" vs. "less" price improvement by estimating equation (6) on each order imbalance measure separately.

Recall that the distribution of subpenny price improvements is displayed in Figure 1 Panel B. Most transaction occur at a round penny or half penny. Based on the other bins, each covering 0.1 cent, there is slightly more trading volume for the "less price improvement" category than for the "more price improvement." The regression results for the cross-section of future returns are shown in Table XI Panel F. For the "less price improvement" type, the coefficients range from 0.0004 to 0.0007, both with *t*-statistics above 5. For the "more price improvement" type, the coefficients range from 0.0001 to 0.0002, both with *t*-statistics below 4. Clearly, both sets of retail order imbalances have predictive power for future stock returns, but the retail trades with less price

Electronic copy available at: https://ssrn.com/abstract=2822105

improvement have stronger predictive power, indicating that internalizers/wholesalers successfully price-discriminate against retail orders with potentially more information content. Similar to the presence of large realized spreads, this observation also supports the viability of the business model, particularly for internalizers and wholesalers who can successfully distinguish between more- and less-informed order flows.

F. Earnings Announcements and Marketable Retail Order Flow

Kelley and Tetlock (2013) use the Dow Jones news archive to identify whether retail investors are informed about cash flow news, and find that retail market orders can predict earnings surprises.

Here, we examine whether retail order flow becomes more predictive around earnings news. Specifically, we estimate a variant of equation (6) that allows the predictive relationship to differ based on the variable eventday, an indicator that takes a value of 1 if day t is an earnings announcement day and zero otherwise. The results are shown in Table XI Panel G. They show that the predictive power of retail order flow is greater on announcement days, but the difference is not statistically significant. In Appendix Table VI, we directly replicate the results in Kelley and Tetlock (2013), and find that our retail order flows can predict earnings news positively. The predictive power is statistically significant at the 1-day horizon, but insignificant over longer horizons. That is, we are able to partially confirm KT's results. The difference can be attributed to the different samples periods and coverage we use. While we cover all subpenny trades for all stocks over 2010 to 2015, Kelley and Tetlock cover about 1/3 of all retail trades between 2003 and 2007.

G. Controlling for Overall Order Imbalances

Previous studies such as Chordia and Subrahmanyam (2004) find that overall order imbalances (calculated using all reported transactions, including individual and institutional types) can predict future stock returns. We use the Lee-Ready algorithm to compute the overall order imbalance from TAQ data. In our dataset, overall order imbalances and retail order imbalances are significantly correlated at around 30%. An interesting question is whether overall and retail order imbalances are relatively orthogonal to each other: Specifically, if we control for the overall order imbalance, can the retail order imbalance still predict future stock returns?

Electronic copy available at: https://ssrn.com/abstract=2822105

We proceed in two steps to address this question and report the results in Table XI Panel H. In the first step, we re-estimate equation (6) using the overall order imbalance from the previous week rather than retail order imbalance as a key predicting variable. Consistent with the literature, we find that overall order imbalances significantly predict future stock returns, with a coefficient of 0.0004 and a significant *t*-statistic of 3.32.

In the second step, we estimate equation (6) using the retail order imbalance variables as key predicting variables, and include the overall order imbalance as a control. With both retail and market order imbalances in the model, retail imbalances are significantly positive, and they completely drive out the effect of market order imbalances. Thus, the predictive power of the retail order imbalance seems to be stronger than that of the overall order imbalance measure.

Here, we want to be cautious about the interpretation in the sense that this finding does not necessarily indicate that the retail investors are more informed than the institutional investors for the following two reasons. First, due to the different calculation methods for the two oib measures, the difference between the overall oib and the retail oib is not the oib from the institutional investors. Second, we only calculate the order flow from marketable retail orders, which account for about half of the trades from retail investors, and the overall oib's weaker predictive power might be partially a result of uninformed trading by other participants in the market.

H. When the Effective Spread is Less Than 1 Cent

Our identification for buy and sell orders relies on the implicit assumption that price improvements are always a small fraction (less than half) of a cent. If price improvements are larger, our method may not correctly sign trades. For example, if a stock has a bid price of $50.01 and an ask price of $50.04, and a retail market buy order arrives and is improved by 0.75 cents, the reported transaction price would be $50.0325, and our trade-signing approach would erroneously conclude that this is a sell order. We investigate whether our identification method is reliable in three ways. First, recall that when we cross-validate using the 2010 Nasdaq TRF sample, we find a trade sign error rate of only about 2%. Second, we examine intraday quote data from TAQ. For all 2015 trades that we can sign using our approach, we compare our buy-sell assignment to the trade sign from the Lee and Ready algorithm, and we find that the trade signs match for 89.9% of the observations. Last, we impose a strict filter that requires the average effective spread from the previous month to be at most one cent, and re-examine our results. For stocks with a one-

Electronic copy available at: https://ssrn.com/abstract=2822105

Produced with a Trial Version of PDF Annotator - www.PDFAnnotator.com

cent spread, our trade-sign approach for subpenny-priced trades should match the Lee-Ready algorithm exactly and should be virtually error-free overall. This strict filter excludes more than 80% of the data, and we retain only the most liquid stocks in the sample. The results are shown in Table XI Panel I. We find the retail order imbalance still significantly predicts the next week's stock returns, with a coefficient of 0.0008 and a significant *t*-statistic of 4.48, consistent with the findings shown in Table III.

IV. Conclusions

In this paper, we exploit the fact that most retail order flows in U.S. equity markets are internalized or sold to wholesalers. As a part of this routing process, retail orders are typically given a small fraction of a penny per share of price improvement relative to the national best bid or offer price, and this price improvement can be observed when the trade is reported to the consolidated tape. Institutional orders almost never receive this kind of price improvement, so it becomes possible to use subpenny trade prices to identify a broad swath of marketable retail order flow. It is also straightforward to identify whether the retail trader is buying or selling stock: transactions at prices that are just above a round penny are classified as retail sales, while transactions that are just below a round penny are retail purchases.

We use this methodology to characterize the trading behavior and information content of retail orders. We find that retail investors are on average contrarian over weekly horizons, buying stocks that have experienced recent price declines and selling stocks that have risen in the past week. More significantly, we find that the retail order flow can predict the cross section of future stock returns. Over the next week, stocks with more positive retail order imbalances outperform stocks with relatively negative retail order imbalances by about 10 basis points, which is on the order of 5% annualized. This predictability extends to about 12 weeks before dying off. Through a decomposition exercise, we attribute less than half of the predictive power of retail order imbalances to the order imbalance's persistence and potential liquidity provision by retail investors' contrarian trading. The remainder of the predictive power (over half of it) is consistent with the hypothesis that the retail order flow contains valuable information about future returns. Concerning the information content of the retail trades, we provide supportive evidence that retail investors are better informed about firm-level news and are likely to have valuable private information.

Electronic copy available at: https://ssrn.com/abstract=2822105

An important advantage of our method is that it is based on widely available intraday transaction data: Anyone with access to TAQ can easily identify retail buys and sells using our approach. Our approach has many possible research applications. For example, future researchers can investigate certain behavioral biases to determine whether individual traders as a group exhibit them. Another possibility is studying the seasonality and time-series variation of retail trading, including tax-related and calendar-driven trading, as well as activity around corporate events, such as dividends, stock splits, and equity issuance.[18]

[18] Our measure is already used in a few studies. For instance, Farrell, Green, Jame, and Markov (2018) find our retail order imbalances are strongly correlated with the sentiment of "Seeking Alpha" articles and that the ability of retail order imbalances to predict returns is roughly twice as high on research article days. Israeli, Kasznik, and Sridharan (2019) use our methodology to identify retail investor trading and then use abnormal retail trading volume and Bloomberg searches as specific measures of retail and institutional investor attention.

Electronic copy available at: https://ssrn.com/abstract=2822105

References

Barber, Brad M., and Terrance Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *Journal of Finance* 55, 773−806.

Barber, Brad M., and Terrance Odean, 2008, All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies* 21, 785-818.

Barber, Brad M., Terrance Odean, and Ning Zhu, 2009, Do retail trades move markets? *Review of Financial Studies* 22, 151−186.

Barrot, Jean-Noel, Ron Kaniel, and David Alexandre Sraer, 2016, Are Retail Traders Compensated for Providing Liquidity? *Journal of Financial Economics* 120, 146-168.

Battalio, Robert, Shane A. Corwin, and Robert Jennings, 2016, Can Brokers Have It All? On the Relation between Make-Take Fees and Limit Order Execution Quality, *Journal of Finance* 71, 2193-2238.

Blume, Marshall E., and Robert F. Stambaugh, 1983, Biases in Computed Returns: An Application to the Size Effect, *Journal of Financial Economics* 12, 387-404.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang, 2008. Which Shorts Are Informed? *Journal of Finance* 63, 491-527.

Boehmer, Ekkehart, Charles M. Jones, Yuan Wu, and Xiaoyan Zhang, 2019. What Do Short Sellers Know? *Review of Finance*, forthcoming.

Campbell, John Y., Tarun Ramadorai, and Allie Schwartz, 2009, Caught on tape: Institutional Trading, Stock Returns, and Earnings Announcements, *Journal of Financial Economics*, 92(1), 66-91.

Chakrabarty, Bidisha, Pamela C. Moulton, and Charles Trzcinka, 2017, The Performance of Short-Term Institutional Trades. *Journal of Financial and Quantitative Analysis* 52.4 (2017), 1403-1428.

Chordia, Tarun, and Avanidhar Subrahmanyam, 2004, Order Imbalance and Stock Returns: Theory and Evidence, *Journal of Financial Economics* 72, 485−518.

Chordia, Tarun, Sahn-Wook Huh, and Avanidhar Subrahmanyam, 2006, The Cross-Section of Expected Trading Activity, *Review of Financial Studies* 20, 709-740.

Electronic copy available at: https://ssrn.com/abstract=2822105

Fama, Eugene F., and James D. MacBeth, 1973, Risk, Return, and Equilibrium: Empirical Tests, *Journal of Political Economy* 81, 607−636.

Fong, Kingsley YL, David R. Gallagher, and Adrian D. Lee, 2014, Individual Investors and Broker Types, Journal of Financial and Quantitative Analysis 49(2), 431-451.

Hansen, Lars Peter, and Robert J. Hodrick, 1980, Forward exchange rates as optimal predictors of future spot rates: An econometric analysis, *Journal of political economy*, *88*(5), 829-853.

Hendershott, Terrence, Dmitry Livdan, and Norman Schürhoff, 2015, Are Institutions Informed About News? *Journal of Financial Economics*, 117(2), 249-287.

Kaniel, Ron, Saar Gideon, and Titman, Sheridan, 2008, Individual Investor Sentiment and Stock Returns, *Journal of Finance* 63, 273−310.

Kaniel, Ron, Liu, Shuming, Saar, Gideon, and Titman, Sheridan, 2012, Individual Investor Trading and Return Patterns Around Earnings Announcements, *Journal of Finance* 67, 639-680.

Kelley, Eric K. and Paul C. Tetlock, 2013, How Wise Are Crowds? Insights from Retail Orders and Stock Returns, *Journal of Finance* 68, 1229-1265.

Kwan, Amy, Ronald Masulis, and Thomas H. McInish, 2015, Trading rules, Competition for Order Flow and Market Fragmentation, *Journal of Financial Economics*, 115, 330-348.

Lee, Charles M. and Balkrishna Radhakrishna, 2000, Inferring Investor Behavior: Evidence from TORQ Data, *Journal of Financial Markets*, 3(2), 83-111.

Lee, Charles M.C. and Mark J. Ready, 1991, Inferring Investor Behavior from Intraday Data, *Journal of Finance*, 46(2), 733-746.

Lee, Yi-Tsung, Liu, Yu-Jane, Roll, Richard, and Subrahmanyam, Avanidhar, 2004, Order Imbalances and Market Efficiency: Evidence from the Taiwan Stock Exchange. *Journal of Financial and Quantitative Analysis*, 39, 327-341.

Menkveld, Albert J., Bart Z. Yueshen, and Haoxiang Zhu, 2017, Shades of Darkness: A Pecking Order of Trading Venues, *Journal of Financial Economics*, 124, 503-534.

Newey, Whitney K., and Kenneth D. West, 1987, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55, 703−708.

O'Hara, Maureen, Chen Yao, and Mao Ye, 2014, What's Not There: Odd Lots and Market Data, *Journal of Finance*, 69(5), 2199-2236.

Electronic copy available at: https://ssrn.com/abstract=2822105

Table I. Summary Statistics

This table reports summary statistics of our measure of marketable retail investor trading activity. Our sample period covers January 2010 to December 2015, and our sample firms are common stocks listed on all U.S. stock exchanges with a share price of at least $1. Across all stocks and all days, we report the pooled sample mean for the daily number of shares traded (*vol*), retail buy volume (*indbvol*), retail sell volume (*indsvol*), number of trades (*trd*), retail buy trades (*indbtrd*), retail sell trades (*indstrd*), as well as their odd lot counterparts (prefix *odd*). Odd lot measures are available starting at the end of 2013. We include odd lot-related data starting January 2014. We compute order imbalance measures (variables containing *oib*) as in equations (1) to (4).

	N	Mean	Std	Median	Q1	Q3
Round lots and odd lots						
Vol	4,628,957	1,229,004	6,849,849	221,234	51,768	819,615
Trd	4,628,957	5,917	13,909	1,505	312	5,502
Indbvol	4,628,957	42,481	280,474	5,165	1,200	20,681
Indsvol	4,628,957	42,430	264,704	5,635	1,369	21,828
Indbtrd	4,628,957	110	410	22	5	79
Indstrd	4,628,957	108	355	24	6	81
Oibvol	4,628,957	-0.038	0.464	-0.027	-0.301	0.217
Oibtrd	4,628,957	-0.032	0.437	-0.010	-0.276	0.205
Odd lots only						
Oddvol	1,446,749	6,561	20,141	1,811	629	5,250
Oddtrd	1,446,749	222	669	64	21	186
Oddindbvol	1,446,749	1,108	5,054	211	58	690
Oddindsvol	1,446,749	968	3,488	210	62	663
Oddindbtrd	1,446,749	37	171	7	2	23
Oddindstrd	1,446,749	33	114	7	2	23
Oddoibvol	1,446,749	-0.004	0.559	0.014	-0.338	0.331
Oddoibtrd	1,446,749	-0.017	0.506	0.000	-0.290	0.250

Electronic copy available at: https://ssrn.com/abstract=2822105

Table II. Determinants of Retail Investor Order Imbalances

This table reports determinants of retail investor trading activity. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate Fama-MacBeth regressions as specified in equation (5). The dependent variables are two scaled retail order imbalance measures: *oibvol* (based on the number of shares traded) and *oibtrd* (based on the number of trades). As independent variables, we include the previous-week return, *ret*(*w*-1), previous-month return, *ret*(*m*-1), and previous 6-month return, *ret*(*m*-7, *m*-2). We compute the weekly returns in two ways: using the end-of-day bid-ask average price or the CRSP closing price. The control variables are monthly turnover (*lmto*), monthly volatility of daily returns (*lvol*), log market cap (*size*), and log book-to-market ratio (*lbm*), all measured at the end of the previous month. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with six lags.

Reg	I		II		III		IV	
Dep.var	Oibvol		Oibvol		Oibtrd		Oibtrd	
Return	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	-0.4013	-20.03	-0.4065	-20.19	-0.4326	-22.00	-0.4357	-22.01
Oib(w-1)	0.2200	92.53	0.2201	92.57	0.2865	150.01	0.2866	150.06
Ret(w-1)	-0.9481	-40.60	-0.9620	-41.43	-0.9003	-35.92	-0.9156	-36.74
Ret(m-1)	-0.2778	-19.24	-0.2784	-19.30	-0.2258	-14.84	-0.2262	-14.87
Ret(m-7,m-2)	-0.0586	-11.49	-0.0584	-11.46	-0.0380	-6.50	-0.0378	-6.48
Lmto	0.0003	5.31	0.0003	5.19	0.0002	3.93	0.0002	3.83
Lvol	0.8100	8.37	0.8478	8.79	0.4366	4.24	0.4633	4.51
Size	0.0154	12.06	0.0157	12.31	0.0209	16.37	0.0211	16.48
Lbm	-0.0275	-17.66	-0.0274	-17.61	-0.0274	-18.09	-0.0273	-18.05
Adj.R2	6.00%		6.01%		9.49%		9.50%	

Electronic copy available at: https://ssrn.com/abstract=2822105

Table III. Predicting Next-week Returns Using Retail Order Imbalances

This table reports estimation results on whether retail investors' trading activity can predict the cross-section of one-week-ahead returns. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate Fama-MacBeth regressions as specified in equation (6). The dependent variable is weekly individual stock returns, computed in two ways: using end-of-day bid-ask average price or CRSP closing price. The independent variables are two scaled retail order imbalance measures: *oibvol* (based on the number of shares traded) and *oibtrd* (based on the number of trades). As independent variables, we include the previous-week return, *ret*(*w*-1), previous-month return, *ret*(*m*-1), and previous 6-month return, *ret*(*m*-7, *m*-2). The control variables are log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*), and monthly volatility of daily returns (*lvol*), all measured at the end of the previous month. To account for serial correlation in the coefficients, the standard errors of the estimated coefficients are adjusted using Newey-West (1987) with five lags.

Reg	I		II		III		IV	
Order imbalance	Oibvol		Oibvol		Oibtrd		Oibtrd	
Dep.var	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	0.0050	2.58	0.0056	2.85	0.0050	2.58	0.0056	2.85
Oib(w-1)	0.0009	15.60	0.0010	16.29	0.0008	12.30	0.0008	13.20
Ret (w-1)	-0.0185	-5.83	-0.0220	-6.85	-0.0186	-5.88	-0.0222	-6.91
Ret (m-1)	0.0006	0.35	0.0006	0.34	0.0005	0.29	0.0005	0.29
Ret (m-7, m-2)	0.0008	1.16	0.0008	1.16	0.0008	1.12	0.0008	1.12
Lmto	0.0000	-3.37	0.0000	-3.76	0.0000	-3.36	0.0000	-3.75
Lvol	-0.0223	-1.41	-0.0205	-1.31	-0.0217	-1.37	-0.0198	-1.27
Size	-0.0001	-0.86	-0.0001	-0.92	-0.0001	-0.90	-0.0001	-0.96
Lbm	-0.0001	-0.39	0.0000	-0.07	-0.0001	-0.42	0.0000	-0.10
Adj.R2	3.85%		3.85%		3.84%		3.84%	
Interquartile	1.1888		1.1888		1.2292		1.2292	
Interquartile weekly return diff	0.1089%		0.1144%		0.0931%		0.0997%	

Electronic copy available at: https://ssrn.com/abstract=2822105

Table IV. Retail Return Predictability within Subgroups

This table reports whether retail investor order imbalances can predict the cross section of returns for subsets of stocks. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We first sort all firms into 3 groups based on previous month-end characteristics. Then we estimate Fama-MacBeth regressions, specified in equation (6), for each subgroup. The dependent variable is weekly returns on individual stocks, computed using end-of-day bid-ask average price. The independent variables are two scaled retail order imbalance measures: *oibvol* (based on number of retail shares traded) and *oibtrd* (based on number of retail trades). To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with five lags. For each regression, we also provide the interquartile range for the relevant explanatory order imbalance along with the difference in predicted week-ahead returns for observations at the two ends of the interquartile range. Control variables are the same as those in Table 3; those coefficients are not reported.

Panel A. Market cap groups

Oib measure	Oibvol				Oibtrd			
Mkt cap	Coef.	t-stat	Interquartile	Weekly return diff	Coef.	t-stat	Interquartile	Weekly return diff
Small	0.0013	13.90	1.662	0.219%	0.0012	11.58	1.736	0.207%
Medium	0.0007	9.18	1.323	0.087%	0.0004	5.63	1.346	0.059%
Big	0.0003	3.68	0.892	0.026%	0.0002	2.52	0.929	0.019%

Panel B. Share price groups

Oib measure	Oibvol				Oibtrd			
Price groups	Coef.	t-stat	Interquartile	Weekly return diff	Coef.	t-stat	Interquartile	Weekly return diff
Low	0.0014	13.34	1.432	0.205%	0.0012	10.34	1.586	0.185%
Medium	0.0007	10.00	1.289	0.089%	0.0005	7.56	1.309	0.070%
High	0.0002	3.23	0.961	0.020%	0.0002	2.19	0.961	0.015%

Panel C. Turnover groups

Oib measure	Oibvol				Oibtrd			
Turnover groups	Coef.	t-stat	Interquartile	Weekly return diff	Coef.	t-stat	Interquartile	Weekly return diff
Low	0.0011	15.60	1.837	0.205%	0.0011	14.71	1.777	0.195%
Medium	0.0008	10.21	1.219	0.094%	0.0006	7.05	1.228	0.071%
High	0.0007	4.98	0.910	0.065%	0.0004	2.55	1.005	0.037%

Electronic copy available at: https://ssrn.com/abstract=2822105

Table V. Predicting Returns k-weeks Ahead

This table reports estimation results on whether retail investor trading activity can predict the cross-section of stock returns at more distant horizons. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate Fama-MacBeth regressions, as specified in equation (7). The dependent variable is the weekly individual stock return n-weeks ahead, computed in two ways: using the end-of-day bid-ask average price (Panel A) or CRSP closing price (Panel B). The independent variables are two scaled retail order imbalance measures, *oibvol* (based on the number of retail shares traded) or *oibtrd* (number of retail trades), respectively. To account for serial correlation in the coefficients, the standard deviations of the time-series are adjusted using Newey-West (1987) with five lags. Control variables are the same as those in Table 3; those coefficients are not reported.

Panel A. Predict bid-ask average return k weeks ahead

# of weeks ahead	Oibvol Coef.	t-stat	Oibtrd Coef.	t-stat
1 week	0.00092	15.60	0.00076	12.30
2 weeks	0.00055	9.35	0.00048	7.89
4 weeks	0.00031	5.56	0.00026	4.66
6 weeks	0.00022	3.90	0.00015	2.60
8 weeks	0.00021	3.47	0.00011	1.75
10 weeks	0.00010	1.82	0.00002	0.35
12 weeks	0.00007	1.29	0.00009	1.52

Panel B. Predict CRSP return k weeks ahead

# of weeks ahead	Oibvol Coef.	t-stat	Oibtrd Coef.	t-stat
1 week	0.00096	16.29	0.00081	13.20
2 weeks	0.00058	9.99	0.00052	8.57
4 weeks	0.00032	5.92	0.00028	5.05
6 weeks	0.00024	4.18	0.00017	2.93
8 weeks	0.00021	3.50	0.00011	1.80
10 weeks	0.00011	2.04	0.00005	0.81
12 weeks	0.00008	1.39	0.00010	1.76

Electronic copy available at: https://ssrn.com/abstract=2822105

Table VI. Long-short Strategy Returns Based on Retail Order Imbalances

This table reports portfolio returns using a long-short strategy wherein we buy the stocks in the highest quintile of scaled retail order imbalance, and we short the stocks in the lowest retail order imbalance quintile. The order imbalance is computed during the previous week. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. Portfolio returns are value-weighted, and market cap terciles are based on the previous month-end market cap. Because the holding period can be as long as 12 weeks, we report both the raw returns and risk-adjusted returns using the Fama-French three-factor model. As our data are overlapping, we adjust the standard errors of the portfolio return time-series using Hansen-Hodrick (1980) standard errors with the corresponding number of overlapping lags.

Panel A. Form portfolios on the previous week retail order imbalance based on number of shares traded

Holding	Whole sample				Small		Medium		Big	
Period	Mean	t-stat	alpha	t-stat	alpha	t-stat	alpha	t-stat	alpha	t-stat
1 week	0.092%	2.66	0.084%	2.43	0.403%	9.16	0.170%	6.24	0.067%	1.78
2 weeks	0.147%	2.45	0.135%	2.46	0.669%	9.01	0.292%	6.81	0.105%	1.70
4 weeks	0.223%	1.89	0.208%	2.00	1.124%	10.43	0.423%	6.36	0.143%	1.22
6 weeks	0.310%	1.72	0.277%	1.73	1.399%	13.02	0.558%	6.07	0.171%	1.05
8 weeks	0.448%	1.92	0.460%	2.26	1.709%	17.13	0.623%	4.18	0.342%	1.69
10 weeks	0.515%	1.99	0.484%	1.81	1.704%	11.17	0.578%	3.87	0.381%	1.53
12 weeks	0.588%	2.09	0.629%	1.89	1.857%	7.65	0.556%	3.20	0.477%	1.48

Panel B. Form portfolios on the previous week retail order imbalance based on number of trades

Holding	Whole sample				Small		Medium		Big	
Period	Mean	t-stat	alpha	t-stat	alpha	t-stat	alpha	t-stat	alpha	t-stat
1 week	0.056%	1.34	0.061%	1.44	0.343%	7.04	0.104%	3.52	0.055%	1.42
2 weeks	0.137%	1.72	0.143%	1.89	0.557%	6.72	0.194%	4.02	0.119%	1.61
4 weeks	0.238%	1.61	0.251%	1.88	0.880%	6.98	0.277%	3.75	0.214%	1.61
6 weeks	0.311%	1.50	0.350%	1.93	1.145%	6.25	0.313%	2.62	0.304%	1.84
8 weeks	0.427%	1.58	0.523%	2.26	1.468%	6.40	0.353%	1.91	0.449%	2.19
10 weeks	0.454%	1.41	0.539%	1.74	1.442%	5.37	0.292%	1.56	0.483%	1.64
12 weeks	0.529%	1.47	0.667%	1.70	1.672%	5.30	0.228%	1.05	0.567%	1.51

Electronic copy available at: https://ssrn.com/abstract=2822105

Electronic copy available at: https://ssrn.com/abstract=2822105

Table VII. Predictability Decomposition

This table reports estimation results on a decomposition of retail order flow's predictive power for the cross-section of future stock returns. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate two-stage Fama-MacBeth regressions. Panel A reports the first-stage estimation results, where the order imbalance measures are decomposed into three components, as specified in equation (8). Panel B reports the second-stage decomposition of order imbalance's predictive power, as specified in equation (9) and (10). The weekly returns are computed in two ways: using end-of-day bid-ask average price or CRSP closing price. The scaled order imbalance measures are *oibvol* (based on the number of retail shares traded) and *oibtrd* (based on the number of retail trades). The variable oib(w-1, persistence) is estimated in the first stage using past order imbalance and reflects price pressure. The variable oib(w-1, contrarian) is estimated in the first stage using past returns over different horizons and is connected to the liquidity provision hypothesis. The residual part of the previous-week order imbalance from the first-stage estimation is denoted as "other," which can be attributed to private information about future returns on the part of these retail investors. As additional control variables, we include previous-week return, *ret*(*w*-1), previous-month return, *ret*(*m*-1), and previous 6-month return, *ret*(*m*-7, *m*-2). The control variables are log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*), and monthly volatility of daily returns (*lvol*), all measured at the end of the previous month. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with five lags.

Panel A. First stage of projecting order imbalance on persistence and past return

Reg	I		II		III		IV	
Dep.var	Oibvol		Oibvol		Oibtrd		Oibtrd	
Return	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	-0.1413	-24.66	-0.1408	-24.61	-0.1054	-17.23	-0.1049	-17.19
Oib(w-1)	0.2227	96.20	0.2228	96.20	0.2906	149.82	0.2907	149.85
Ret(w-1)	-0.9286	-38.93	-0.9422	-39.80	-0.8926	-34.92	-0.9076	-35.81
Ret(m-1)	-0.2029	-13.93	-0.2025	-13.90	-0.1591	-10.72	-0.1588	-10.70
Ret(m-7,m-2)	-0.0267	-4.98	-0.0268	-4.99	-0.0054	-0.86	-0.0055	-0.88
Adj.R2	5.62%		5.63%		8.99%		9.00%	

Panel B. Second-stage decomposition of order imbalance's predictive power

Reg	I		II		III		IV	
Order Imbalance	Oibvol		Oibvol		Oibtrd		Oibtrd	
Dep.var	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	0.0045	2.20	0.0051	2.48	0.0045	2.20	0.0051	2.49
Oib(w-1,persistence)	0.0027	8.75	0.0029	9.41	0.0018	7.80	0.0019	8.56
Oib(w-1,contrarian)	-0.0044	-0.42	-0.1310	-1.46	-0.0073	-0.73	0.0328	1.62
Oib(w-1,other)	0.0008	14.47	0.0009	15.48	0.0006	10.51	0.0007	11.64
Ret(w-1)	-0.0176	-5.41	-0.0206	-6.27	-0.0177	-5.45	-0.0207	-6.30
Ret(m-1)	-0.0060	-0.67	0.0002	0.03	0.0017	0.56	0.0093	1.13
Ret(m-7,m-2)	-0.0009	-0.65	-0.0127	-1.12	0.0017	0.95	-0.0008	-0.34
Lmto	0.0000	-3.49	0.0000	-3.80	0.0000	-3.48	0.0000	-3.78
Lvol	-0.0230	-1.48	-0.0231	-1.50	-0.0224	-1.44	-0.0225	-1.46
Size	-0.0001	-0.61	-0.0001	-0.67	-0.0001	-0.65	-0.0001	-0.72
Lbm	-0.0001	-0.46	0.0000	-0.14	-0.0001	-0.56	-0.0001	-0.23
Adj.R2	4.26%		4.27%		4.25%		4.26%	
	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff
Oib(w-1,persistence)	0.2591	0.0688%	0.2593	0.0739%	0.3498	0.0620%	0.3500	0.0679%
Oib(w-1,contrarian)	0.0627	-0.0277%	0.0631	-0.8265%	0.0614	-0.0445%	0.0619	0.2031%
Oib(w-1,other)	1.1141	0.0915%	1.1141	0.0977%	1.1326	0.0718%	1.1327	0.0792%

Electronic copy available at: https://ssrn.com/abstract=2822105

Table VIII. Retail order imbalance and contemporaneous returns, replicating KST Table III using oibvol

This table presents analysis of market-adjusted returns around net buying and selling activity as given by our scaled retail order imbalance measure *oibvol* (based on the number of shares traded). The sample extends from Jan 2010 to Dec 2015. For each (non-overlapping) week in the sample period, we aggregate the daily order imbalance measures to weekly to form *Oib* deciles. Each stock is put into 1 of 10 deciles according to the *Oib* value in the current week. Decile 1 contains the stocks with the most net selling (negative *Oib*) while decile 10 contains the stocks with the most net buying (positive *Oib*). We present the results for four portfolios: (i) decile 1, (ii) deciles 1 and 2, (iii) deciles 9 and 10, and (iv) decile 10. Let k be the number of days prior to or following the portfolio formation each week. In Panel A, we calculate eight cumulative return numbers for each of the stocks in a portfolio: $CR(t-k,t-1)$, where $k \in \{20,15,10,5\}$ days and t is the first day of the formation week, and $CR(t+1,t+k)$, where $k \in \{5,10,15,20\}$ days and t is the last day of the formation week. The return on each portfolio is then adjusted by subtracting the return on a market proxy (the equal-weighted portfolio of all stocks in the sample). We present the time-series mean and t-statistic for each market-adjusted cumulative return measure and for the market-adjusted return during the intense trading week (k=0). In Panel B, we present the time-series mean and t-statistic for weekly market-adjusted returns in the 4 weeks around the formation week (i.e., $\mathrm{CR}(\mathrm{t}-\mathrm{k},\mathrm{t}-\mathrm{k}+4)$, where $k \in \{20,15,10,5\}$ days and t is the first day of the formation week, and $\mathrm{CR}(\mathrm{t}+\mathrm{k}-4,\mathrm{t}+\mathrm{k})$, where $k \in \{5,10,15,20\}$ and t is the last day of the formation week). ** indicates significance at 1% level and * indicates significance at 5% level (both against a two-sided alternative). The t-statistic is computed using Newey-West standard errors.

Panel A. Cumulative market-adjusted return

Oibvol	Bid-ask return	k=-20	k=-15	k=-10	k=-5	k=0	k=5	k=10	k=15	k=20
Intense Selling	Mean	0.0067**	0.0056**	0.0041**	0.0027**	-0.0024**	-0.0016**	-0.0023**	-0.0025*	-0.0030*
(decile 1)	t-stat	5.48	5.62	5.40	6.06	-5.30	-3.89	-3.16	-2.45	-2.36
Selling	Mean	0.0063**	0.0055**	0.0042**	0.0028**	-0.0019**	-0.0012**	-0.0018**	-0.0022**	-0.0025**
(decile 1&2)	t-stat	8.10	8.71	9.04	10.07	-5.65	-5.07	-4.35	-3.97	-3.63
Buying	Mean	-0.0111**	-0.0096**	-0.0074**	-0.0047**	0.0011**	0.0018**	0.0028**	0.0036**	0.0043**
(decile 9&10)	t-stat	-19.00	-20.87	-20.83	-24.09	4.02	9.58	8.38	8.39	8.89
Intense Buying	Mean	-0.0129**	-0.0109**	-0.0084**	-0.0053**	0.0011**	0.0024**	0.0036**	0.0046**	0.0057**
(decile10)	t-stat	-12.49	-12.91	-12.93	-15.62	2.69	6.99	5.89	5.51	5.74

Panel B. Weekly market-adjusted return

Oibvol	Bid-ask return	k=-20	k=-15	k=-10	k=-5	k=0	k=5	k=10	k=15	k=20
Intense Selling	Mean	0.0010**	0.0016**	0.0014**	0.0027**	-0.0024**	-0.0016**	-0.0006	-0.0001	-0.0005
(decile 1)	t-stat	2.75	4.12	3.59	6.06	-5.30	-3.89	-1.51	-0.21	-1.34
Selling	Mean	0.0008**	0.0014**	0.0015**	0.0028**	-0.0019**	-0.0012**	-0.0006*	-0.0004	-0.0003
(decile 1&2)	t-stat	3.36	5.66	5.88	10.07	-5.65	-5.07	-2.42	-1.74	-1.29
Buying	Mean	-0.0016**	-0.0022**	-0.0028**	-0.0047**	0.0011**	0.0018**	0.0010**	0.0009**	0.0006**
(decile 9&10)	t-stat	-8.15	-12.70	-12.66	-24.09	4.02	9.58	5.38	4.61	3.84
Intense Buying	Mean	-0.0019**	-0.0026**	-0.0030**	-0.0053**	0.0011**	0.0024**	0.0013**	0.0010**	0.0010**
(decile10)	t-stat	-5.90	-7.97	-7.62	-15.62	2.69	6.99	3.68	2.77	2.80

Electronic copy available at: https://ssrn.com/abstract=2822105

Electronic copy available at: https://ssrn.com/abstract=2822105

Table IX. Relation between Public News and Retail Order Flow

This table reports analysis of the relation between public news and our retail investor order imbalance. Our sample period covers January 2010 to December 2014, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. In Panel A, we examine whether public news can predict the cross-section of future stock returns. We estimate the Fama-MacBeth regression specified in equation (11). The dependent variable is weekly returns, computed in two ways: using end-of-day bid-ask average price or CRSP closing price. The independent variables are *sent(w-1)*, which is the average TRNA net sentiment score for firm *i* during week *w* by averaging non-missing news sentiment for firm *i* within week *w*-1. We also include past retail order imbalance *oibvol* (based on the number of shares traded by retail) in regressions III and IV. As control variables, we include the previous-week return, *ret*(*w*-1), previous-month return, *ret*(*m*-1), and previous 6-month return, *ret*(*m*-7, *m*-2), log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*), and monthly volatility of daily returns (*lvol*), measured at the end of the previous month. In Panel B, we examine the relation between contemporaneous public news sentiment and retail order imbalance across subtopics. We estimate the Fama-MacBeth regression specified in equation (12). The dependent variable is weekly net sentiment score, *sent(i,w)*. The independent variable is the retail order imbalance measure *oibvol* . As control variables, we include the previous-week return, *ret(w-1)*, previous-month return, *ret(m-1)* and previous 6-month return, *ret(m-7,m-2)*, log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*), monthly volatility of daily returns (*lvol*). Coefficients on controls are not reported for brevity. To account for serial correlation in the coefficients, the standard errors of the estimated coefficients are adjusted using Newey-West (1987) with 5 lags.

Panel A. Predicting returns using public news and retail order flow

Reg	I		II		III		IV	
Dep.var	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
Order Imbalance					Oibvol		Oibvol	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	0.0057	2.48	0.0066	2.83	0.0061	2.66	0.0070	3.02
Sent(w-1)	0.0008	3.31	0.0009	3.64	0.0008	3.33	0.0009	3.66
Oib(w-1)					0.0009	10.61	0.0010	11.53
Ret(w-1)	-0.0088	-2.65	-0.0105	-3.10	-0.0090	-2.70	-0.0107	-3.16
Ret(m-1)	0.0008	0.38	0.0009	0.44	0.0013	0.64	0.0015	0.72
Ret(m-7,m-2)	0.0001	0.15	0.0001	0.11	0.0002	0.21	0.0001	0.18
Lmto	0.0000	-1.03	0.0000	-1.29	0.0000	-1.14	0.0000	-1.41
Lvol	-0.0435	-2.15	-0.0465	-2.34	-0.0444	-2.20	-0.0477	-2.40
Size	-0.0001	-0.90	-0.0001	-1.04	-0.0001	-0.99	-0.0002	-1.14
Lbm	0.0001	0.29	0.0001	0.61	0.0001	0.44	0.0002	0.79

Adj.R2	5.01%	5.01%	5.06%	5.08%

Panel B. contemporaneous relation between sentiment and order imbalance

Topic	Type	Description	N	g1	t-stat
RESF	equities	results forecast	102,515	0.0054	3.90
AAA	money/debt	debt rating news	23,405	0.0131	3.66
DIP	general news	diplomacy	6,057	0.0362	3.34
DIV	equities	Dividend	24,282	0.0093	2.77
IGD	money/debt	investment grade debt	6,760	0.0261	2.76
DRV	cross market	derivatives	18,061	0.0238	2.58
DBT	money/debt	debt markets	73,600	0.0060	2.55
MTG	money/debt	mortgage-backed debt	7,764	0.0264	2.52
RES	equities	corporate results	176,699	0.0031	2.36
JUDIC	general news	Judicial	28,280	0.0096	2.31

Electronic copy available at: https://ssrn.com/abstract=2822105

Table X. Predictability Decomposition using Public News Releases

This table reports estimation results on a decomposition of our retail order flow measure's predictive power for future returns. Our sample period covers January 2010 to December 2014, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate two-stage Fama-MacBeth regressions specified in equations (13) and (15). The dependent variable is weekly returns, computed in two ways: using end-of-day bid-ask average price or CRSP closing price. The independent variables are scaled retail order imbalance measures: *oibvol* (based on the number of shares traded) and *oibtrd* (based on the number of trades). In the first-stage estimation, the order imbalance measures are decomposed into four components. The variable oib(w-1, persistence) is estimated in the first stage using past order imbalance and reflects price pressure. The variable oib(w-1, contrarian) is estimated in the first stage using past returns over different horizons, which is connected to the liquidity provision hypothesis. The variable oib(w-1, public) is estimated in the first stage using week w news sentiment, which proxies for retail order imbalances that predict returns associated with future news releases. The residual part of previous-week order imbalance from first-stage estimation is denoted as "other," which we attribute to retail investors' valuable private information about future returns that is incorporated into prices but is not associated with an identifiable public news release. As additional control variables, we include previous-week return, *ret*(*w*-1), previous-month return, *ret*(*m*-1), and previous 6-month return, *ret*(*m*-7, *m*-2). Other control variables are log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*), and monthly volatility of daily returns (*lvol*), all measured at the end of the previous month. To account for serial correlation in the coefficients, the standard deviations of the time-series are adjusted using Newey-West (1987) with five lags.

Panel A. First stage of projecting order imbalance on persistence, past return, and public news.

Reg	I		II		III		IV	
Dep.var	Oibvol		Oibvol		Oibtrd		Oibtrd	
Return	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef	t-stat	Coef	t-stat	Coef	t-stat	Coef	t-stat
Intercept	-0.1510	-24.08	-0.1505	-24.03	-0.1132	-16.89	-0.1127	-16.85
Oib(w-1)	0.2208	112.64	0.2209	112.67	0.2827	125.44	0.2829	125.48
Ret(w-1)	-0.8918	-36.22	-0.9051	-37.00	-0.8940	-34.04	-0.9098	-35.01
Ret(m-1)	-0.2169	-13.68	-0.2156	-13.60	-0.1702	-10.56	-0.1687	-10.47
Ret(m-7,m-2)	-0.0264	-4.54	-0.0264	-4.55	-0.0080	-1.17	-0.0081	-1.19
Sent(w)	0.0249	11.60	0.0249	11.60	0.0305	13.81	0.0305	13.81
Adj.R2	5.49%		5.50%		8.58%		8.59%	

Electronic copy available at: https://ssrn.com/abstract=2822105

Panel B. Second-stage decomposition of order imbalance's predictive power

Reg	I		II		III		IV	
Order Imbalance	Oibvol		Oibvol		Oibtrd		Oibtrd	
Dep.var	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef	t-stat	Coef	t-stat	Coef	t-stat	Coef	t-stat
Intercept	0.0054	2.35	0.0059	2.57	0.0054	2.35	0.0059	2.58
Oib(w-1,persistence)	0.0027	8.50	0.0029	9.16	0.0018	7.28	0.0020	8.02
Oib(w-1,contrarian)	0.0088	0.55	0.6947	1.03	-0.0182	-1.00	0.0419	1.18
Oib(w-1,public news)	0.1150	1.17	-0.0134	-0.35	-0.0386	-0.84	0.0020	0.04
Oib(w-1,other)	0.0008	13.98	0.0009	15.16	0.0006	10.15	0.0007	11.35
Ret(w-1)	-0.0217	-6.25	-0.0250	-7.12	-0.0218	-6.28	-0.0251	-7.16
Ret(m-1)	0.0059	0.54	0.6585	1.03	0.0106	1.05	0.0119	1.41
Ret(m-7,m-2)	0.0014	1.01	0.0511	1.02	0.0059	0.99	-0.0004	-0.22
Lmto	0.0000	-2.40	0.0000	-2.75	0.0000	-2.35	0.0000	-2.71
Lvol	-0.0273	-1.63	-0.0252	-1.52	-0.0266	-1.59	-0.0244	-1.47
Size	-0.0001	-0.68	-0.0001	-0.73	-0.0001	-0.73	-0.0001	-0.79
Lbm	0.0001	0.31	0.0001	0.66	0.0001	0.25	0.0001	0.59
Adj.R2	4.22%		4.23%		4.21%		4.22%	
	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff
Oib(w-1,persistence +contrarian+public news)	0.2760	0.0707%	0.2763	0.0761%	0.3609	0.0614%	0.3611	0.0678%
Oib(w-1,other)	1.12019	0.0932%	1.12032	0.1006%	1.16542	0.0745%	1.16541	0.0829%

Electronic copy available at: https://ssrn.com/abstract=2822105

Electronic copy available at: https://ssrn.com/abstract=2822105

Table XI. Additional Analysis

Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. Standard errors are calculated using Newey-West (1987). In Panel A, we estimate equation (16). The dependent variable is the n-week ahead weekly value-weighted market return. The independent variables are two scaled retail order imbalance measures, *oibvol* (based on the number of retail shares traded), and *oibtrd* (based on the number of retail trades), respectively. For all other panels, the regression is specified in equation (6) and estimated using Fama-MacBeth regressions. In Panel B, the dependent variable is weekly returns on approximately 1000 ETFs. In Panel C, we estimate the coefficients for different VIX regimes. In Panel D, the independent variables are two scaled odd lot retail order imbalance measures, *oddoibvol* (based on the number of odd lot shares traded), and *oddoibtrd* (based on the number of odd lot trades), respectively. In Panel E, we estimate the coefficients for different trade size. In Panel F, we estimate the coefficients for different amounts of price improvement. In Panel G, we estimate a variant of equation (6) that allows the predictive relationship to differ based on the variable event day, an indicator that takes a value of 1 if day t is an announcement day and zero otherwise. In Panel H, we estimate the coefficients controlling for overall order imbalance computing by Lee-Ready algorithm. In Panel I, we estimate the coefficient when the effective spread is less than 1 cent. The dependent variable is weekly returns, computed in two ways: using end-of-day bid-ask average price or CRSP closing price. The independent variable is one of the two scaled retail order imbalance measures *oibvol* or *oibtrd*. Control variables for the cross-sectional regressions are the same as those shown in Table III, except that we do not include a book-to-market variable in the ETF regression; those coefficients are not reported.

Panel A. Predicting future n-week market return

Weights	Oibvol Value weight		Oibvol Equal weight		Oibtrd Value weight		Oibtrd Equal weight	
Horizon	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
1 week	0.0037	0.50	-0.0053	-0.57	0.0054	0.92	-0.0038	-0.46
2 weeks	0.0101	0.79	-0.0030	-0.20	0.0120	1.21	0.0007	0.06
4 weeks	0.0044	0.20	-0.0236	-1.04	0.0073	0.43	-0.0136	-0.63
6 weeks	-0.0061	-0.22	-0.0356	-1.25	0.0022	0.10	-0.0216	-0.80
8 weeks	0.0075	0.20	-0.0046	-0.10	0.0118	0.41	0.0044	0.11
10 weeks	0.0051	0.11	-0.0114	-0.23	0.0101	0.28	-0.0038	-0.08
12 weeks	-0.0059	-0.10	-0.0315	-0.58	0.0000	0.00	-0.0227	-0.46

Panel B. Using retail oib to predict ETF returns

Order imbalance	Oibvol		Oibtrd	
Dep.var	Bid-ask return		Bid-ask return	
	Coef.	t-stat	Coef.	t-stat
All ETFs	0.0001	2.04	0.0001	1.68
Interquartile	1.4726		1.4737	
Return diff	0.0153%		0.0118%	
Broad market ETFs	-0.0004	-0.81	0.0005	1.52

Panel C. Different market conditions

	Vix <=18%		Vix>18%	
Dep.var	Bid-ask return		Bid-ask return	
Indep.var	coef.	t-stat	coef.	t-stat
Oibvol	0.0009	13.49	0.0010	9.36
Oibtrd	0.0007	10.32	0.0008	7.60

Panel D. Predicting stock returns using odd-lot order imbalances

Order imbalance	Oibvol		Oibtrd	
Dep.var	Bid-ask return		Bid-ask return	
	Coef.	t-stat	Coef.	t-stat
Odd lot	0.0001	1.41	0.0001	0.77
Interquartile	1.2734		1.1314	
Return diff	0.0154%		0.0086%	

Panel E. Different retail trade sizes

Order imbalance	Oibvol		Oibtrd	
Dep.var	Bid-ask return		Bid-ask return	
	Coef.	t-stat	Coef.	t-stat
Small trades (< 400 shares)	0.0004	5.77	0.0004	4.48
Large trades (≥ 400 shares)	0.0009	7.25	0.0008	5.85

Panel F. Different price improvement amounts

Order imbalance	Oibvol		Oibtrd	
Dep.var	Bid-ask return		Bid-ask return	
	Coef.	t-stat	Coef.	t-stat
Less price improvement	0.00071	9.30	0.00042	5.57
More price improvement	0.00021	3.04	0.00018	2.43

Electronic copy available at: https://ssrn.com/abstract=2822105

Panel G. Earnings surprises

Order imbalance	Oibvol		Oibtrd	
Dep.var	Bid-ask return		Bid-ask return	
	Coef.	t-stat	Coef.	t-stat
Oib	0.0003	8.16	0.0004	11.98
Oib* eventday	0.0003	1.47	0.0002	1.31

Panel H. Retail vs. overall order imbalance

Order imbalance	Overall Oib		Oibvol		Oibtrd	
Dep.var	Bid-ask return		Bid-ask return		Bid-ask return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Retail Oib			0.0011	6.14	0.0006	3.33
Overall Oib	0.0004	3.32	0.0000	0.10	0.0001	0.51

Panel I. When effective spread is less than one cent

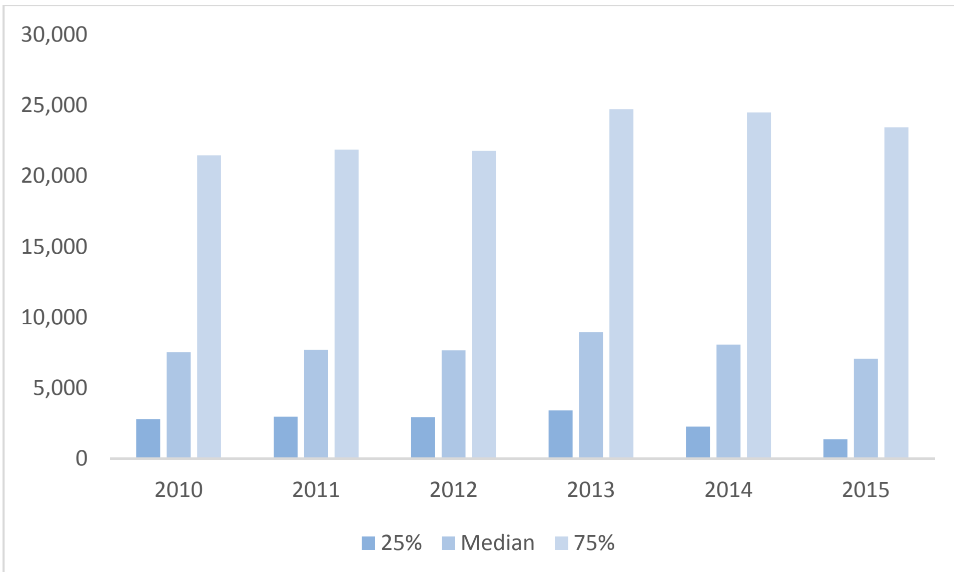
Order imbalance	Oibvol		Oibtrd	
Dep.var	Bid-ask return		Bid-ask return	
	Coef.	t-stat	Coef.	t-stat
Oib	0.0008	4.48	0.0004	2.45

Electronic copy available at: https://ssrn.com/abstract=2822105

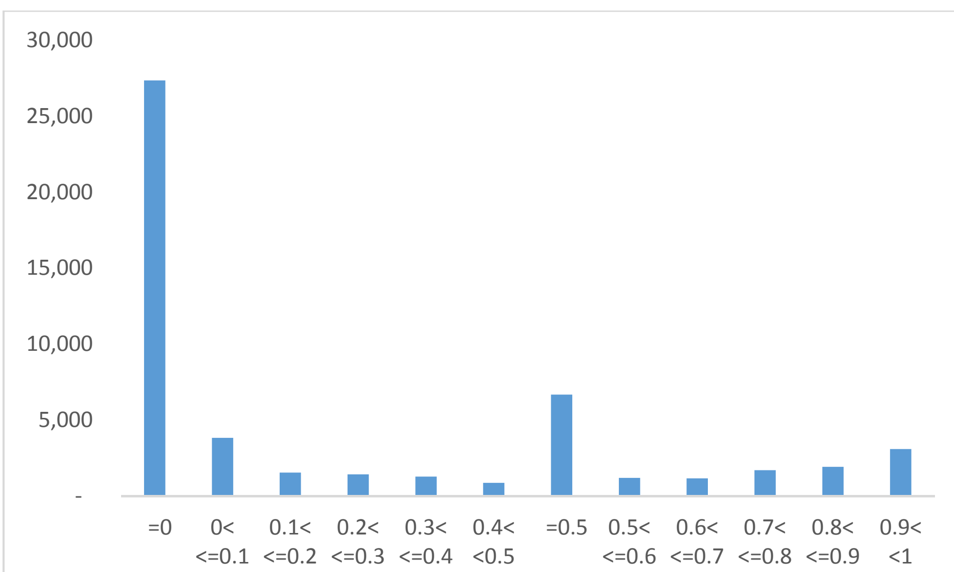
Figure 1. Distribution of Trade Size and Subpenny Prices for Retail Orders

These figures report summary statistics for the retail investor trading we identify. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. In Panel A, we compute the trade size in dollars as the number of shares multiplied by transaction price. For each year, we report the cross-sectional median, q1 (25th percentile) and q3 (75th percentile). In Panel B, we separate trades into 12 groups based on subpenny increments: trades at the whole penny, at the half penny, and in buckets that are 0.1 cent wide. We report the cross-sectional median of the daily number of shares traded in each group.

Panel A. Retail order trade size in dollars



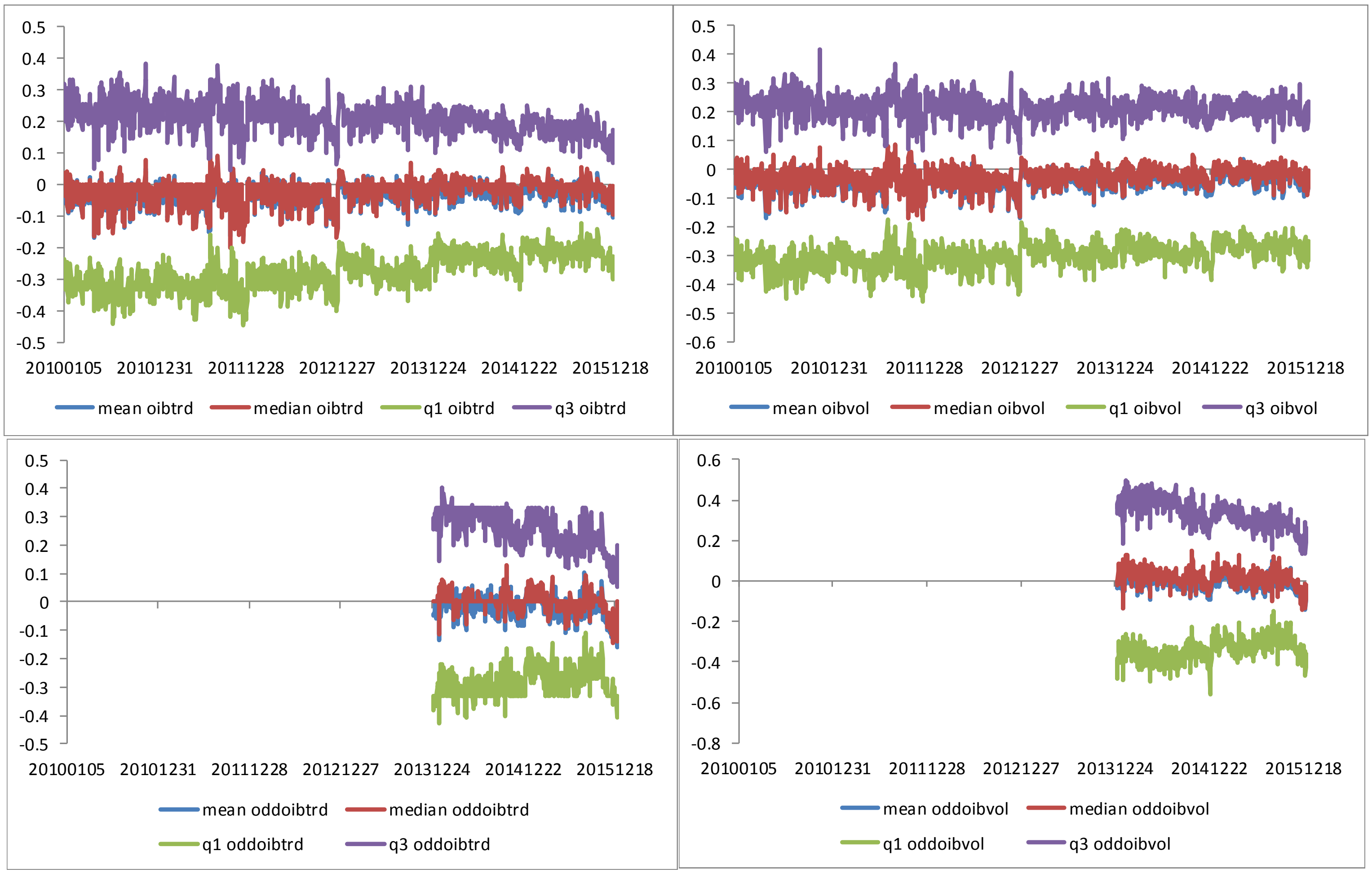
Panel B. Median share volumes for different subpenny groups



Electronic copy available at: https://ssrn.com/abstract=2822105

Figure 2. Time Series of Retail Investor Order Imbalances

These figures report time series statistics of our identified retail investor trading activity. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We present cross-sectional means, medians, q1 (25th percentile), and q3 (75th percentile) for each day.

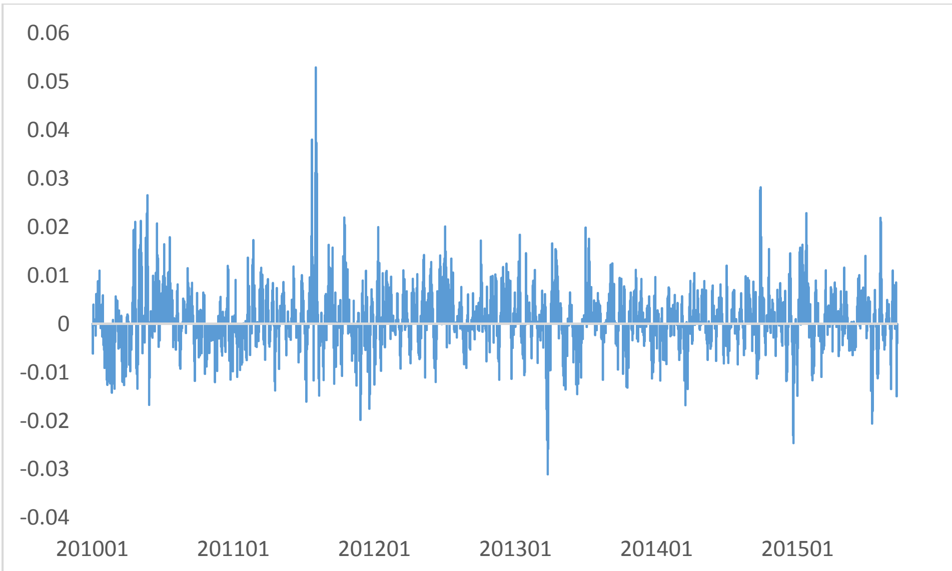


Electronic copy available at: https://ssrn.com/abstract=2822105

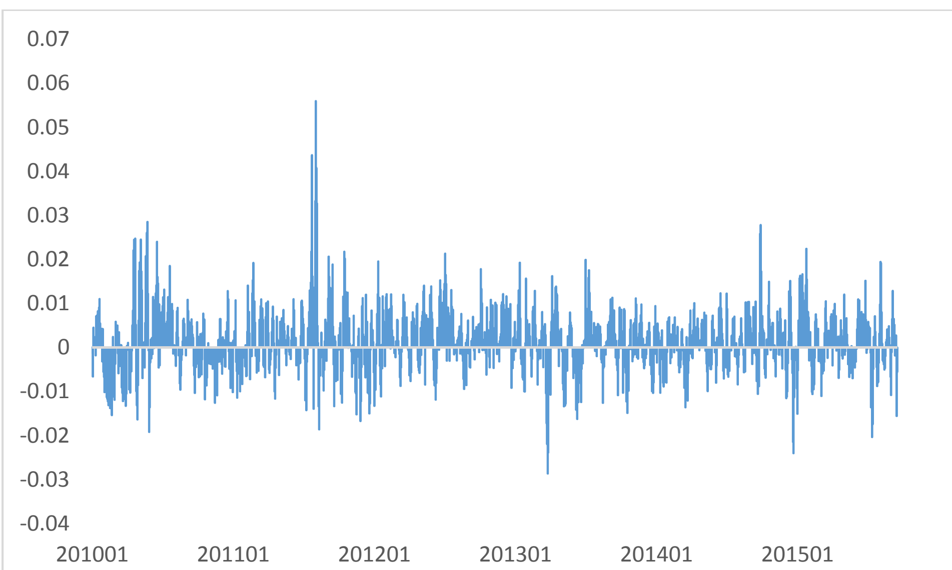
Figure 3. Portfolio Return Difference Using Previous Week Retail Order Imbalance

These figures plot weekly value-weighted portfolio return differences between quintile 5 and quintile 1, where stocks are sorted on the previous-week retail order imbalance calculated using the number of shares traded (*oibvol*). Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. The portfolio returns are computed using the end-of-day bid-ask average price (*bidaskret*) in Panel A and the CRSP closing price (*crspret*) in Panel B.

Panel A. Weekly portfolio return difference using end-of-day bid-ask average prices



Panel B. Weekly portfolio return difference using CRSP closing prices



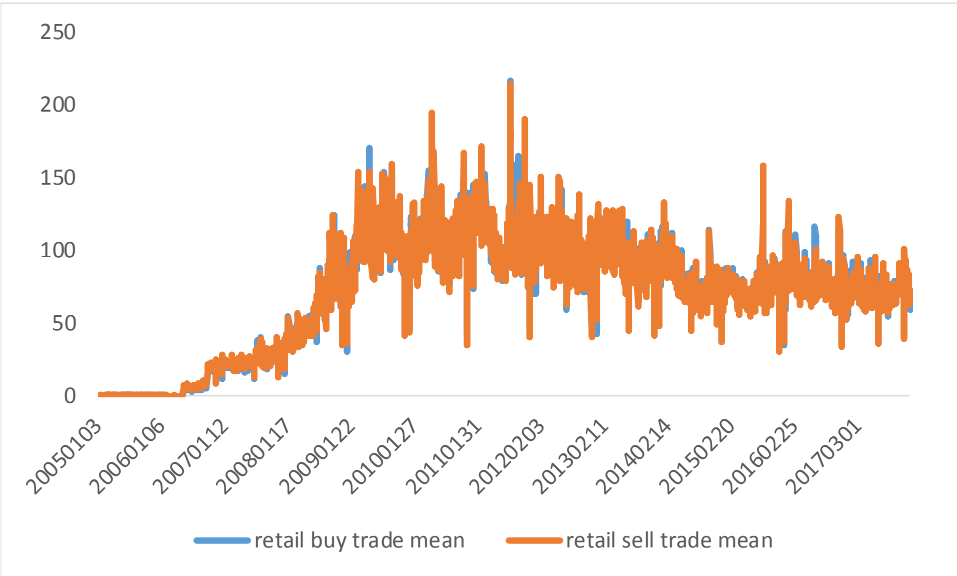
Electronic copy available at: https://ssrn.com/abstract=2822105

Internet Appendix.

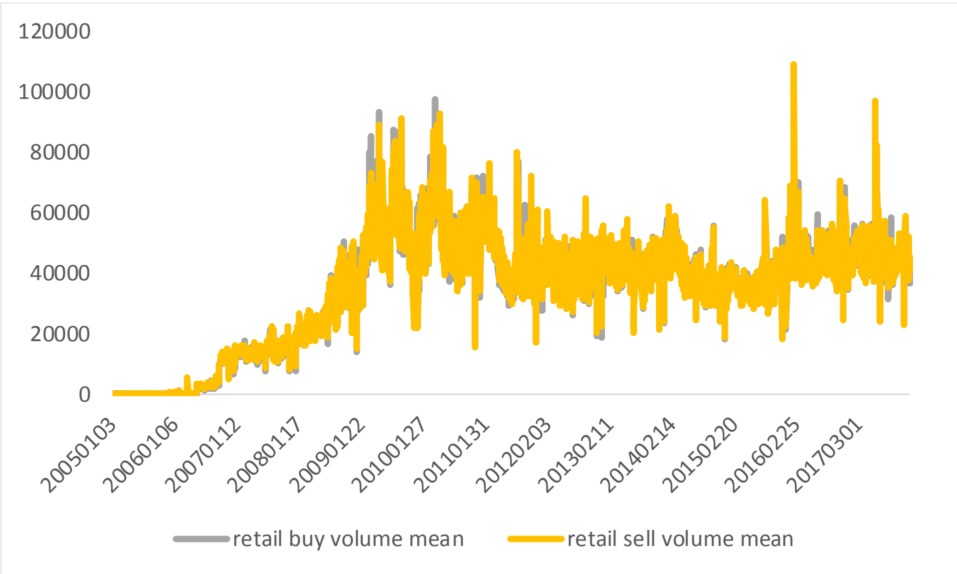
Figure 1. Retail Order Flows between 2005 and 2017

These figures report the time series mean of retail investor trading activity from January 2005 (the start of Reg NMS, which allows subpenny price improvement in its current form) to December 2017. Our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. In Panel A, we separately present the cross-sectional mean for each stock-day for retail buy trades and retail sell trades. In Panel B, we present the cross-sectional mean for each stock-day for retail buy volume and retail sell volume.

Panel A. Number of Trades



Panel B. Share Volume

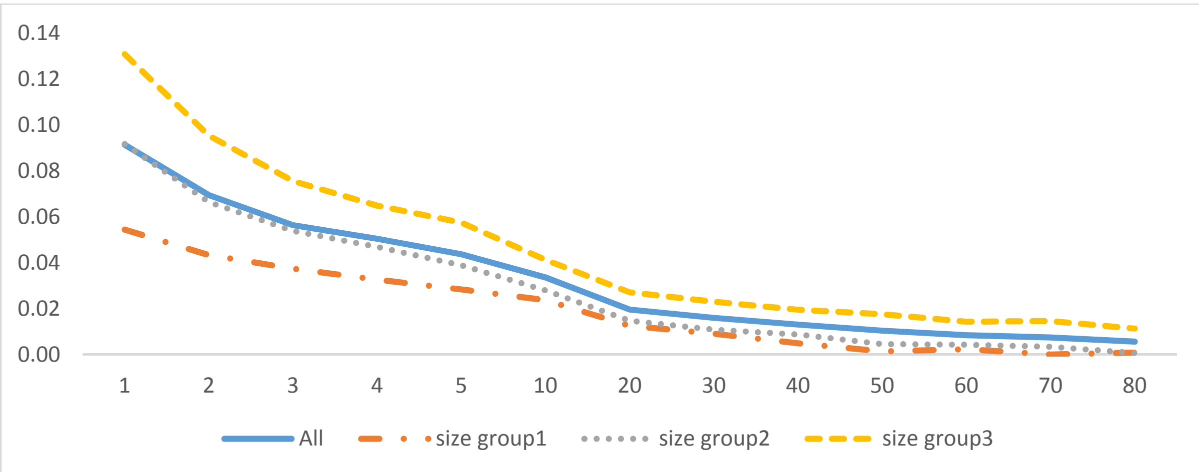


Electronic copy available at: https://ssrn.com/abstract=2822105

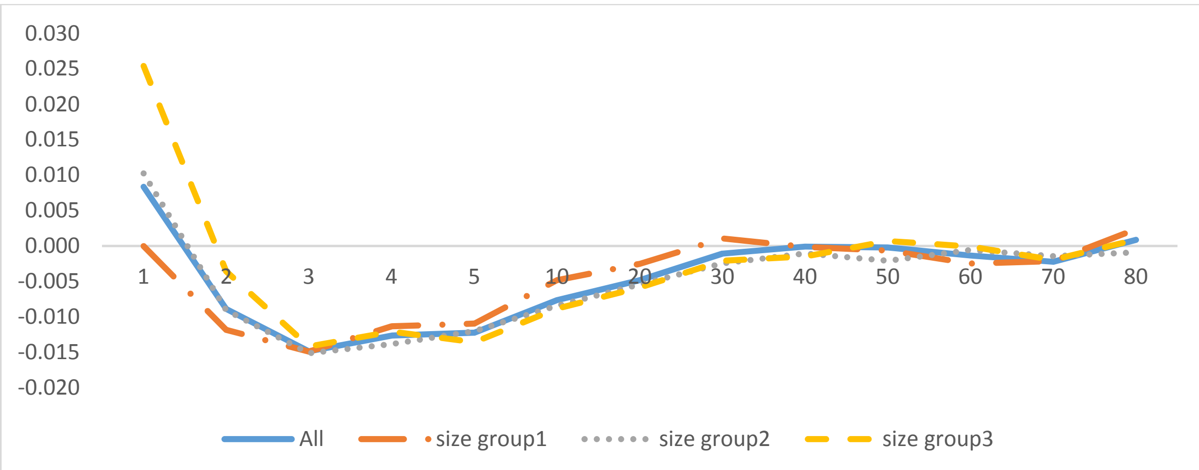
Appendix Figure 2. Retail Order Imbalance Persistence and Cross-correlations with Returns

These figures report the autocorrelation of retail order imbalances and cross-autocorrelations with past returns. Our sample period covers January 2010 to December 2015, and our sample firms are common stocks listed on all U.S. stock exchanges with a share price of at least $1. We also report subgroup correlations based on market cap (size) measured at the end of the previous month. Panel A shows the auto-correlations of retail order imbalances using the number of shares for each firm at a lag of N days (N = 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, and 80), reporting the cross-firm median of autocorrelations. Panel B shows the cross-autocorrelations of retail order imbalances using number of trades with past returns for each firm, with the correlation calculated at a lag of N days (N = 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, and 80) and reporting the cross-firm median of these cross-autocorrelations.

Panel A. Auto-correlations of Retail Order Imbalance Measure Using Number of Shares



Panel B. Correlation of Retail Order Imbalance Measure Using Number of Trades with Past Returns



Electronic copy available at: https://ssrn.com/abstract=2822105

Appendix Table I. What Affects Retail Order Imbalances?

This table reports covariates with our retail order imbalance measure. Our sample period covers January 2010 to December 2015, and our sample firms are common stocks listed on all U.S. stock exchanges with a share price of at least $1. We estimate Fama–MacBeth regressions in all panels. The dependent variables are two scaled retail order imbalance measures: *oibvol* (based on the number of retail shares traded) and *oibtrd* (based on the number of retail trades). As independent variables, we include the previous-week return *ret(w-1)*, previous-month return, and previous 6-month return, *ret(m-7,m-2)*. We compute weekly returns in two ways: using end-of-day bid–ask average price or CRSP closing price. The control variables are monthly turnover (*lmto*), monthly volatility of daily returns (*lvol*), log market cap (*size*), and log book-to-market ratio (*lbm*), all measured at the end of the previous month. In Panels A and B, variable *Max(ret[t],0)* is equal to ret[t] if the return is positive and 0 otherwise; *Min(ret[t],0)* is equal to ret[t] if the return is negative and 0 otherwise. In Panel C, firms are sorted into 3 groups based on previous month-end characteristics. The *Rank2* interactive dummy is 1 if the firm belongs to the medium characteristic group and 0 otherwise. The *Rank3* interactive dummy is 1 if the firm belongs to the large characteristic group and 0 otherwise. In Panel D, the Monday dummy is equal to 1 if the trading day is Monday and 0 otherwise. The dummy variables Friday, December, and January are defined similarly. In Panel E, the dependent variables are two normalized order imbalance measures: *oibvol[m]* (the number of monthly retail buy shares minus the number of monthly retail sell shares divided by the sum of the number of monthly retail buy shares and the number of monthly retail sell shares), *oibtrd[m]* (the number of monthly retail buy trades minus the number of monthly retail sell trades divided by the sum of the number of monthly retail buy trades and the number of monthly retail sell trades). As independent variables, we include the monthly return of an individual stock if positive and 0 otherwise (*Ret+*); the monthly return of an individual stock if negative and 0 otherwise (*Ret-*); the logarithm of month-end market value (*ASIZE*=ln(MV)); the logarithm of one plus the number of months since its listing on an exchange M (*FAGE*=log(1+M)); the logarithm of price (*ALNP*=ln(Prc)); the logarithm of one plus the number of analysts who follow a company and report forecasts to the I/B/E/S database (*ALANA*=log(1+ANA)); the standard deviation of the most recent eight quarterly earnings (*EVOLA*); the absolute value of the most recent quarterly earnings minus the earnings four quarters ago (*ESURP*); the analyst forecast dispersion (*FDISP*), which is defined as the standard deviation of earnings per share forecasts from multiple (two or more) analysts; firm leverage (*LEVRG*), which is the book debt divided by total assets, where book debt is current liabilities, long term debt, and preferred stocks. *Beta* is the ex-ante rolling beta regressed on the market factor using three years of daily returns.

Electronic copy available at: https://ssrn.com/abstract=2822105

Panel A. Momentum or contrarian over daily horizon: past returns

Reg	I		II		III		IV	
Dep.var	Oibvol		Oibvol		Oibtrd		Oibtrd	
Return	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	-0.1173	-26.85	-0.1195	-27.27	-0.1322	-31.78	-0.1335	-31.88
Oib[-1]	0.0854	75.44	0.0854	75.40	0.1357	104.59	0.1357	104.70
Max(ret[-1],0)	0.1975	12.27	0.1919	11.77	0.2701	16.79	0.2563	15.82
Min(ret[-1],0)	-0.5474	-21.69	-0.5510	-21.30	-0.4896	-20.27	-0.5061	-20.41
Max(ret[-6,-2],0)	-0.0987	-12.81	-0.1015	-13.23	-0.0482	-6.56	-0.0514	-7.00
Min(ret[-6,-2],0)	-0.3551	-36.09	-0.3604	-36.64	-0.3229	-32.99	-0.3267	-33.61
Max(ret[-27,-7],0)	-0.0546	-10.32	-0.0548	-10.44	-0.0316	-6.07	-0.0307	-5.90
Min(ret[-27,-7],0)	-0.1281	-21.84	-0.1309	-22.40	-0.1278	-21.25	-0.1301	-21.75
Ret(m-7,m-2)	-0.0147	-13.03	-0.0146	-12.92	-0.0103	-8.27	-0.0102	-8.19
Lmto	0.0000	3.29	0.0000	3.09	0.0000	2.41	0.0000	2.37
Lvol	-0.0254	-0.95	-0.0182	-0.69	-0.1266	-4.87	-0.1283	-4.92
Size	0.0048	18.02	0.0050	18.68	0.0065	24.66	0.0066	25.01
Lbm	-0.0057	-15.90	-0.0056	-15.79	-0.0058	-17.29	-0.0057	-17.20

Panel B. Momentum or contrarian over weekly horizon: past returns

Reg	I		II		III		IV	
Dep.var	Oibvol		Oibvol		Oibtrd		Oibtrd	
Return	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	-0.4355	-21.86	-0.4416	-22.08	-0.4692	-24.00	-0.4731	-24.07
Oib(w-1)	0.2191	94.13	0.2192	94.19	0.2855	151.68	0.2856	151.75
Max(ret(w-1),0)	-0.5250	-17.36	-0.5262	-17.56	-0.4566	-15.05	-0.4582	-15.37
Min(ret(w-1),0)	-1.4741	-35.98	-1.5144	-36.98	-1.4696	-34.09	-1.5140	-34.95
Ret(m-1)	-0.2711	-19.11	-0.2712	-19.13	-0.2187	-14.63	-0.2184	-14.62
Ret(m-7,m-2)	-0.0582	-11.56	-0.0580	-11.53	-0.0376	-6.50	-0.0374	-6.48
Lmto	0.0003	4.92	0.0002	4.81	0.0002	3.49	0.0002	3.40
Lvol	0.5660	5.90	0.5858	6.13	0.1688	1.66	0.1761	1.73
Size	0.0172	13.68	0.0176	14.00	0.0228	18.17	0.0231	18.35
Lbm	-0.0267	-17.36	-0.0266	-17.30	-0.0266	-17.78	-0.0264	-17.73

Panel C. Momentum or contrarian for firms with different characteristics

Reg	I		II		IIII		IV	
Dep.var	Oibvol		Oibvol		Oibvol		Oibvol	
Return	Bid-ask return		Bid-ask return		Bid-ask return		Bid-ask return	
Rank	Size		Price		Mto		Vol	
	Coef	t-stat	Coef	t-stat	Coef	t-stat	Coef	t-stat
Intercept	-0.4018	-21.36	-0.3984	-21.20	-0.3929	-20.76	-0.4002	-21.15
Oib(w-1)	0.2198	99.54	0.2200	99.35	0.2201	99.42	0.2199	99.23
Ret(w-1)	-0.9037	-30.01	-0.9104	-35.86	-1.2887	-30.04	-1.5742	-28.87
Ret(w-1)*Rank2	-0.1524	-3.85	-0.1846	-5.19	0.0676	1.46	0.3994	7.63

Electronic copy available at: https://ssrn.com/abstract=2822105

Ret(w-1)*Rank3	-0.0192	-0.47	-0.0583	-1.48	0.6252	13.25	0.7857	14.12
Ret(m-1)	-0.2775	-20.39	-0.2777	-20.33	-0.2764	-20.35	-0.2754	-20.34
Ret(m-7,m-2)	-0.0586	-12.07	-0.0589	-12.19	-0.0581	-12.00	-0.0587	-12.13
Lmto	0.0003	5.63	0.0003	5.64	0.0002	4.29	0.0003	5.68
Lvol	0.7981	8.58	0.7937	8.57	0.7847	8.52	0.6572	7.02
Size	0.0154	12.85	0.0152	12.67	0.0149	12.39	0.0157	13.13
Lbm	-0.0275	-18.51	-0.0273	-18.45	-0.0273	-18.48	-0.0272	-18.45

Panel D. Seasonality

Reg	I		II		III		IV	
Dep.var	Oibvol		Oibvol		Oibtrd		Oibtrd	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	-0.0383	-138.89	-0.0367	-155.82	-0.0323	-124.24	-0.0303	-136.36
Monday	-0.0034	-6.04			-0.0049	-9.19		
Friday	0.0027	4.94			0.0039	7.50		
December			-0.0343	-43.95			-0.0413	-56.10
January			0.0153	19.01			0.0168	22.08

Panel E. Firm fundamentals vs. monthly order imbalance

Reg	I		II	
Dep.var	Oibvol		Oibtrd	
	Coef.	t-stat	Coef.	t-stat
Intercept	-0.0732	-6.60	-0.0875	-7.37
Ret(-)	-0.1951	-16.39	-0.1980	-15.41
Ret(+)	-0.0102	-1.16	0.0004	0.04
ALNP	-0.0011	-0.59	0.0015	0.68
FAGE	-0.0054	-5.03	-0.0052	-4.78
BTM	-0.0034	-5.42	-0.0044	-7.82
SIZE	0.0022	3.10	0.0047	6.57
ALNAN	0.0082	7.36	0.0033	2.54
Beta	0.0010	0.48	-0.0043	-2.41
ESURP	0.0000	-0.06	0.0000	-0.45
EVOLA	0.0001	0.61	0.0001	0.64
FDISP	0.0009	2.70	0.0007	2.40
LEVRG	0.0357	10.79	0.0203	5.40
Avg adj R2	0.0243		0.0245	

Electronic copy available at: https://ssrn.com/abstract=2822105

Electronic copy available at: https://ssrn.com/abstract=2822105

Appendix Table II. Alternative specifications of decomposition

This table reports the second stage decomposition results of retail order flow's predictive power for the cross-section of future stock returns using different specifications deviated from Table VII. Our sample period covers January 2010 to December 2015, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate two-stage Fama-MacBeth regressions. In the first-stage estimation, the order imbalance measures are decomposed into three components, as specified in equation (8). In the second-stage decomposition of order imbalance's predictive power, as specified in equation (9) and (10). Panel A reports the second stage estimation, including oib(w-3) to control for past order imbalance. Panel B reports the second stage estimation, including oib(m-1) to control for past order imbalance. The weekly returns are computed in two ways: using end-of-day bid-ask average price or CRSP closing price. The scaled order imbalance measures are *oibvol* (based on the number of retail shares traded) and *oibtrd* (based on the number of retail trades). The variable oib(w-1, persistence) is estimated in the first stage using past order imbalance and reflects price pressure. The variable oib(w-1, contrarian) is estimated in the first stage using past returns over different horizons and is connected to the liquidity provision hypothesis. The residual part of the previous-week order imbalance from the first-stage estimation is denoted as "other," which can be attributed to private information about future returns on the part of these retail investors. As additional control variables, we include previous-week return, *ret*(*w*-1), previous-month return, *ret*(*m*-1), and previous 6-month return, *ret*(*m*-7, *m*-2). The control variables are log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*), and monthly volatility of daily returns (*lvol*), all measured at the end of the previous month. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with five lags.

Panel A. Include oib(w-3) as a control for past order imbalance in the second stage

Reg	I		II		III		IV	
Order Imbalance	Oibvol		Oibvol		Oibtrd		Oibtrd	
Dep.var	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	0.0043	2.03	0.0049	2.29	0.0043	2.02	0.0049	2.28
Oib(w-1,persistence)	0.0024	7.89	0.0026	8.44	0.0016	6.73	0.0017	7.34
Oib(w-1, contrarian)	-0.0066	-0.60	-0.1559	-1.50	-0.0067	-0.69	0.0356	1.60
Oib(w-1,other)	0.0008	13.70	0.0009	14.65	0.0006	9.45	0.0007	10.49
Oib(w-3)	0.0002	4.24	0.0003	4.65	0.0002	2.87	0.0002	3.27
Ret(w-1)	-0.0173	-5.31	-0.0201	-6.10	-0.0174	-5.34	-0.0202	-6.13
Ret(m-1)	-0.0062	-0.68	-0.0030	-0.33	0.0020	0.65	0.0117	1.17
Ret(m-7,m-2)	-0.0010	-0.69	-0.0152	-1.13	0.0019	1.03	-0.0013	-0.52

Lmto	0.0000	-3.49	0.0000	-3.78	0.0000	-3.47	0.0000	-3.75
Lvol	-0.0218	-1.39	-0.0224	-1.44	-0.0210	-1.34	-0.0217	-1.39
Size	-0.0001	-0.49	-0.0001	-0.53	-0.0001	-0.52	-0.0001	-0.58
Lbm	-0.0001	-0.47	0.0000	-0.14	-0.0001	-0.59	-0.0001	-0.25
Adj.R2	4.34%		4.36%		4.33%		4.35%	
	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff
Oib(w-1,persistence)	0.2591	0.0628%	0.2593	0.0672%	0.3498	0.0546%	0.3500	0.0597%
Oib(w-1,contrarian)	0.0627	-0.0415%	0.0631	-0.9834%	0.0614	-0.0409%	0.0619	0.2205%
Oib(w-1,other)	1.1141	0.0888%	1.1141	0.0950%	1.1326	0.0676%	1.1327	0.0748%

Panel B. Include oib(m-1) as a control for past order imbalance in the second stage

Reg	I		II		III		IV	
Order Imbalance	Oibvol		Oibvol		Oibtrd		Oibtrd	
Dep.var	Bid-ask return		CRSP return		Bid-ask return		CRSP return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	0.0045	2.16	0.0051	2.44	0.0045	2.16	0.0051	2.45
Oib(w-1,persistence)	0.0026	8.60	0.0028	9.23	0.0017	7.62	0.0019	8.36
Oib(w-1, contrarian)	-0.0058	-0.52	-0.1499	-1.42	-0.0081	-0.82	0.0318	1.58
Oib(w-1,other)	0.0008	14.08	0.0009	15.16	0.0006	10.03	0.0007	11.22
Oib(m-1)	0.0000	1.92	0.0000	2.20	0.0000	1.50	0.0000	1.71
Ret(w-1)	-0.0173	-5.32	-0.0203	-6.17	-0.0174	-5.36	-0.0204	-6.22
Ret(m-1)	-0.0062	-0.68	0.0005	0.07	0.0015	0.52	0.0092	1.12
Ret(m-7,m-2)	-0.0010	-0.69	-0.0152	-1.11	0.0017	0.98	-0.0008	-0.31
Lmto	0.0000	-3.47	0.0000	-3.75	0.0000	-3.46	0.0000	-3.74
Lvol	-0.0232	-1.48	-0.0239	-1.54	-0.0227	-1.44	-0.0233	-1.50
Size	-0.0001	-0.55	-0.0001	-0.61	-0.0001	-0.59	-0.0001	-0.66
Lbm	-0.0001	-0.50	0.0000	-0.17	-0.0001	-0.60	-0.0001	-0.27
Adj.R2	4.28%		4.30%		4.28%		4.29%	
	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff	Interquartile	return diff
Oib(w-1,persistence)	0.2591	0.0683%	0.2593	0.0731%	0.3498	0.0607%	0.3500	0.0665%
Oib(w-1,contrarian)	0.0627	-0.0361%	0.0631	-0.9454%	0.0614	-0.0498%	0.0619	0.1965%
Oib(w-1,other)	1.1141	0.0899%	1.1141	0.0964%	1.1326	0.0697%	1.1327	0.0774%

Electronic copy available at: https://ssrn.com/abstract=2822105

Appendix Table III. Retail order imbalance and contemporaneous returns, replicating KST Table III using oibtrd

This table presents analysis of market-adjusted returns around net buying and selling activity as given by the scaled order imbalance measure *oibtrd* (based on the number of trades). For each non-overlapping week in the sample period (which extends from Jan 2010 to Dec 2015), we aggregate the daily order imbalance measures to form weekly *Oib* deciles. Each stock is put into 1 of 10 deciles according to the *Oib* value in the current week. Decile 1 contains the stocks with the most net selling (negative *Oib*) while decile 10 contains the stocks with the most net buying (positive *Oib*). We present the results for four portfolios: (i) decile 1, (ii) deciles 1 and 2, (iii) deciles 9 and 10, and (iv) decile 10. Let k be the number of days prior to or following the portfolio formation each week. In Panel A, we calculate eight cumulative return numbers for each of the stocks in a portfolio: $CR(t-k, t-1)$, where $k \in \{20,15,10,5\}$ days and t is the first day of the formation week, and $CR(t+1, t+k)$, where $k \in \{5,10,15,20\}$ days and t is the last day of the formation week. The return on each portfolio is then adjusted by subtracting the return on a market proxy (the equal-weighted portfolio of all stocks in the sample). We present the time-series mean and t-statistic for each market-adjusted cumulative return measure and for the market-adjusted return during the intense trading week (k=0). In Panel B, we present the time-series mean and t-statistic for weekly market-adjusted returns in the 4 weeks around the formation week (i.e., $\mathrm{CR}(\mathrm{t}-\mathrm{k}, \mathrm{t}-\mathrm{k}+4)$, where $k \in \{20,15,10,5\}$ days and t is the first day of the formation week, and $\mathrm{CR}(\mathrm{t}+\mathrm{k}-4, \mathrm{t}+\mathrm{k})$, where $k \in \{5,10,15,20\}$ and t is the last day of the formation week). ** indicates significance at 1% level and * indicates significance at 5% level (both against a two-sided alternative). The t-statistic is computed using Newey-West standard errors.

Panel A. Cumulative market adjusted return

Oibtrd	Bid-ask return	k=-20	k=-15	k=-10	k=-5	k=0	k=5	k=10	k=15	k=20
Intense Selling	Mean	0.0051**	0.0047**	0.0035**	0.0022**	-0.0035**	-0.0016**	-0.0025**	-0.0026**	-0.0031**
(decile 1)	t-stat	4.29	4.76	4.72	5.17	-8.02	-3.92	-3.46	-2.68	-2.67
Selling	Mean	0.0047**	0.0044**	0.0033**	0.0022**	-0.0030**	-0.0010**	-0.0016**	-0.0018**	-0.0020**
(decile 1&2)	t-stat	6.58	7.54	7.55	8.77	-9.66	-4.84	-4.16	-3.48	-3.16
Buying	Mean	-0.0101**	-0.0088**	-0.0067**	-0.0041**	0.0025**	0.0015**	0.0023**	0.0029**	0.0033**
(decile 9&10)	t-stat	-18.94	-19.68	-18.91	-19.93	8.19	8.65	7.96	7.64	7.37
Intense Buying	Mean	-0.0131**	-0.0111**	-0.0086**	-0.0051**	0.0024**	0.0019**	0.0029**	0.0038**	0.0046**
(decile10)	t-stat	-14.20	-14.65	-14.30	-15.34	5.69	5.95	5.31	4.86	4.92

Panel B. Weekly market adjusted return

Oibtrd	Bid-ask return	k=-20	k=-15	k=-10	k=-5	k=0	k=5	k=10	k=15	k=20
Intense Selling	Mean	0.0005	0.0013**	0.0013**	0.0022**	-0.0035**	-0.0016**	-0.0008*	-0.0002	-0.0005
(decile 1)	t-stat	1.36	3.39	3.36	5.17	-8.02	-3.92	-2.08	-0.48	-1.37
Selling	Mean	0.0004	0.0012**	0.0011**	0.0022**	-0.0030**	-0.0010**	-0.0005*	-0.0004	-0.0002
(decile 1&2)	t-stat	1.81	5.17	4.66	8.77	-9.66	-4.84	-2.32	-1.77	-0.95
Buying	Mean	-0.0014**	-0.0022**	-0.0027**	-0.0041**	0.0025**	0.0015**	0.0007**	0.0006**	0.0005**
(decile 9&10)	t-stat	-8.03	-12.83	-12.34	-19.93	8.19	8.65	4.02	3.22	2.93
Intense Buying	Mean	-0.0020**	-0.0026**	-0.0035**	-0.0051**	0.0024**	0.0019**	0.0010**	0.0009*	0.0007*
(decile10)	t-stat	-6.52	-8.88	-9.72	-15.34	5.69	5.95	3.17	2.39	2.17

Electronic copy available at: https://ssrn.com/abstract=2822105

Appendix Table IV. Retail Order Imbalance and Future News Sentiment

This table reports estimation results on whether our retail investor order imbalances can predict one-week-ahead sentiment for each news release category. Our sample period covers January 2010 to December 2014, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate Fama-MacBeth regressions for each month as in the following equation using only firm-day observations that have public news for a given subtopic:

$$sent(i,w) = a0 + a1 \times oib(i,w-1) + a2' controls(i,w-1) + u(i,w).$$

The dependent variable is the weekly net sentiment score for a given subtopic, *sent(i,w)*. The independent variables are the scaled retail order imbalance measure *oibvol* (based on the number of shares traded) and *Oib(w-1,other)* from the first stage decomposition in table VII. As control variables, we include the previous-week return, *ret(w-1)*, previous-month return, *ret(m-1)* and previous 6-month return, *ret(m-7,m-2)*, log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*), and monthly volatility of daily returns (*lvol*). Coefficients on controls are not reported for brevity. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey–West (1987) with five lags. We report the coefficient on the retail order imbalance and t-statistics below.

Topic	Type	Description	N	a1(oibvol)	t-stat	a1(oibother)	t-stat
FUND	cross market	investment funds	93,142	-0.0039	-1.19	-0.0003	-0.42
REGS	cross market	regulatory issues	61,513	-0.0044	-1.63	-0.0008	-1.20
MNGISS	cross market	management issues	54,170	-0.0022	-0.96	-0.0002	-0.31
NEWS	cross market	top stories	36,265	-0.0019	-0.30	-0.0016	-1.08
EXCA	cross market	exchange activities	20,533	-0.0114	-1.39	-0.0037	-1.56
DRV	cross market	derivatives	18,061	0.0032	0.45	0.0002	0.12
ISU	cross market	new issues	17,061	0.0073	0.96	0.0011	0.60
INV	cross market	Investing	13,916	0.0090	1.20	0.0015	0.82
PRESS	cross market	press digests	9,783	0.0029	0.22	0.0005	0.19
TRD	cross market	international trade	4,782	0.0192	1.65	0.0055	1.66
BKRT	cross market	bankruptcy	4,700	-0.0090	-0.75	-0.0002	-0.08
RTM	cross market	Retirement	3,075	-0.0187	-1.43	-0.0016	-0.46
HEA	general news	health/medicines	52,110	-0.0099	-3.13	-0.0020	-2.51
POL	general news	domestic politics	49,699	-0.0129	-4.45	-0.0029	-4.26
JUDIC	general news	judicial	28,280	0.0056	1.46	0.0018	1.62
LAW	general news	legislation	17,135	-0.0016	-0.28	0.0003	0.22
ENV	general news	environment/nature	16,189	-0.0178	-3.87	-0.0044	-3.64
SCI	general news	science/technology	11,035	-0.0175	-1.87	-0.0051	-2.32
CRIM	general news	crime	9,866	0.0230	2.27	0.0036	1.76
SECUR	general news	security	7,528	0.0121	1.44	0.0021	0.89
DIP	general news	diplomacy	6,057	0.0167	1.81	0.0024	0.91
WEA	general news	weather	3,860	0.0195	0.88	0.0042	0.58
DIS	general news	disasters/accidents	3,836	-0.0123	-1.31	-0.0055	-1.85

Electronic copy available at: https://ssrn.com/abstract=2822105

Topic	Type	Description	N	a1(oib)	t-stat	a1(oibother)	t-stat
VOTE	general news	elections	1,899	-0.0304	-1.16	-0.0081	-1.05
VIO	general news	civil unrest	1,339	0.0187	0.59	0.0070	0.91
WAR	general news	war/insurgencies	1,053	0.0134	0.32	0.0088	0.58
EMRG	economy	emerging markets	88,990	-0.0069	-3.51	-0.0014	-2.61
WASH	economy	US government news	32,835	-0.0027	-0.48	-0.0006	-0.44
JOB	economy	labor/employment	28,670	-0.0041	-1.24	-0.0005	-0.47
MCE	economy	macroeconomics	18,148	0.0042	0.62	0.0006	0.34
ECI	economy	economic indicators	11,982	0.0077	0.77	0.0017	0.88
CEN	economy	central banks	9,462	-0.0093	-0.80	-0.0043	-1.23
TAX	economy	tax	4,871	-0.0405	-3.63	-0.0096	-3.77
FED	economy	Federal Reserve Board	3,632	-0.0378	-2.85	-0.0131	-4.29
PLCY	economy	policymakers speak	3,263	-0.0187	-1.04	-0.0042	-0.89
INT	economy	interest rates	2,360	0.0004	0.02	0.0038	0.59
RES	equities	corporate results	176,699	-0.0017	-1.83	0.0000	0.14
RCH	equities	broker research	141,765	-0.0137	-6.35	-0.0025	-4.92
RESF	equities	results forecast	102,515	0.0013	1.19	0.0001	0.48
STX	equities	stock markets	101,829	-0.0071	-2.31	-0.0019	-2.91
MRG	equities	ownership changes	69,775	-0.0073	-3.53	-0.0016	-3.28
HOT	equities	hot stocks	66,185	-0.0073	-2.17	-0.0017	-2.14
PVE	equities	private equity	25,740	-0.0167	-3.29	-0.0033	-2.64
DIV	equities	dividend	24,282	0.0021	0.76	0.0002	0.20
IPO	equities	initial public offer	13,913	0.0113	1.94	0.0027	1.90
DBT	money/debt	debt markets	73,600	0.0012	0.45	0.0000	0.04
USC	money/debt	US corporate bonds	26,796	-0.0068	-2.19	-0.0020	-2.50
AAA	money/debt	debt rating news	23,405	0.0072	1.45	0.0015	1.25
LOA	money/debt	loans	17,718	-0.0053	-1.40	-0.0022	-1.86
HYD	money/debt	high-yield debt/junk	10,222	-0.0021	-0.38	-0.0011	-0.77
GVD	money/debt	government debt	8,759	0.0182	1.44	0.0027	0.86
MUNI	money/debt	muni news	7,933	0.0074	0.83	0.0025	0.88
MTG	money/debt	mortgage-backed debt	7,764	0.0063	0.67	0.0004	0.13
FRX	money/debt	forex	7,006	0.0211	1.86	0.0037	1.21
IGD	money/debt	investment grade debt	6,760	0.0110	0.95	0.0023	0.74
ABS	money/debt	asset-backed debt	2,982	0.0120	0.53	0.0028	0.54
TNC	money/debt	bond terms & conditions	1,826	0.0619	1.37	0.0117	1.04
MMT	money/debt	money markets	1,574	-0.0390	-1.25	-0.0066	-0.81

Electronic copy available at: https://ssrn.com/abstract=2822105

Appendix Table V. Retail Return Predictability across News Topics

This table reports estimation results on whether our retail investor order imbalances can predict one-week-ahead returns for each news release category. Our sample period covers January 2010 to December 2014, and our sample firms are all common stocks listed on U.S. stock exchanges with a share price of at least $1. We estimate Fama-MacBeth regression each month as in the following equation, using only firm-day observations that have public news for a given subtopic:

$$ret(i,w) = b0 + b1 \times oib(i,w-1) + b2'controls(i,w-1) + u(i,w)$$

The dependent variable is weekly returns, *ret(i,w)*, using end-of-day bid-ask average prices. The independent variable is the scaled retail order imbalance measure *oibvol* (based on the number of retail shares traded). As control variables, we include the previous-week net sentiment score, *net(w-1)*, the previous-week return, *ret(w-1)*, previous-month return, *ret(m-1)* and previous 6-month return, *ret(m-7,m-2)*, log book-to-market ratio (*lbm*), log market cap (*size*), monthly turnover (*lmto*) and monthly volatility of daily returns (*lvol*). Coefficients on controls are not reported for brevity. To account for serial correlation in the coefficients, standard errors of the time-series are adjusted using Newey–West (1987) with 5 lags. We report the coefficient on the retail order imbalance and t-statistics below. A significant b1 coefficient indicates that retail order imbalances predict returns that occur at the time of news releases in the specified category one week ahead at the 1% level (t>2.58).

Topic	Type	Description	N	b1	t-stat
FUND	cross market	investment funds	93,142	0.0003	0.54
REGS	cross market	regulatory issues	61,513	0.0007	0.85
MNGISS	cross market	management issues	54,170	0.0007	1.54
NEWS	cross market	top stories	36,265	0.0236	1.28
EXCA	cross market	exchange activities	20,533	0.0025	2.10
DRV	cross market	derivatives	18,061	0.0010	1.22
ISU	cross market	new issues	17,061	0.0010	1.08
INV	cross market	Investing	13,916	0.0001	0.09
PRESS	cross market	press digests	9,783	0.0009	0.59
TRD	cross market	international trade	4,782	0.0018	1.91
BKRT	cross market	bankruptcy	4,700	0.0026	1.27
RTM	cross market	Retirement	3,075	-0.0025	-1.29
HEA	general news	health/medicines	52,110	0.0002	0.25
POL	general news	domestic politics	49,699	0.0010	1.63
JUDIC	**general news**	**judicial**	**28,280**	**0.0029**	**4.84**
LAW	general news	legislation	17,135	0.0017	1.66
ENV	general news	environment/nature	16,189	0.0014	2.06
SCI	general news	science/technology	11,035	-0.0003	-0.21
CRIM	general news	crime	9,866	0.0002	0.16
SECUR	**general news**	**security**	**7,528**	**0.0039**	**2.66**
DIP	general news	diplomacy	6,057	0.0029	2.28

Electronic copy available at: https://ssrn.com/abstract=2822105

Topic	Type	Description	N	b1	t-stat
WEA	general news	weather	3,860	0.0010	0.32
DIS	general news	disasters/accidents	3,836	0.0014	0.71
VOTE	general news	elections	1,899	-0.0381	-1.62
VIO	general news	civil unrest	1,339	-0.0069	-0.48
WAR	general news	war/insurgencies	1,053	0.0205	1.69
EMRG	economy	emerging markets	88,990	0.0005	0.79
WASH	economy	US government news	32,835	0.0000	-0.03
JOB	economy	labor/employment	28,670	0.0014	1.98
MCE	economy	macroeconomics	18,148	0.0015	1.88
ECI	economy	economic indicators	11,982	0.0026	1.94
CEN	economy	central banks	9,462	0.0003	0.19
TAX	economy	tax	4,871	0.0022	1.41
FED	economy	Federal Reserve Board	3,632	-0.0024	-1.11
PLCY	economy	policymakers speak	3,263	0.0035	1.72
INT	**economy**	**interest rates**	**2,360**	**0.0084**	**2.61**
RES	**equities**	**corporate results**	**176,699**	**0.0010**	**4.12**
RCH	equities	broker research	141,765	0.0007	1.71
RESF	**equities**	**results forecast**	**102,515**	**0.0013**	**3.16**
STX	equities	stock markets	101,829	0.0004	0.64
MRG	**equities**	**ownership changes**	**69,775**	**0.0019**	**3.14**
HOT	equities	hot stocks	66,185	0.0000	0.04
PVE	equities	private equity	25,740	0.0006	0.60
DIV	equities	dividend	24,282	0.0009	1.53
IPO	equities	initial public offer	13,913	0.0010	0.85
DBT	money/debt	debt markets	73,600	0.0011	1.77
USC	money/debt	US corporate bonds	26,796	0.0014	1.77
AAA	**money/debt**	**debt rating news**	**23,405**	**0.0021**	**3.42**
LOA	money/debt	loans	17,718	0.0009	1.36
HYD	money/debt	high-yield debt/junk	10,222	0.0014	1.61
GVD	money/debt	government debt	8,759	0.0007	0.55
MUNI	money/debt	muni news	7,933	0.0001	0.18
MTG	money/debt	mortgage-backed debt	7,764	0.0025	1.22
FRX	money/debt	forex	7,006	0.0027	1.43
IGD	money/debt	investment grade debt	6,760	0.0011	1.08
ABS	money/debt	asset-backed debt	2,982	-0.0265	-1.15
TNC	money/debt	bond terms & conditions	1,826	-0.0024	-0.19
MMT	money/debt	money markets	1,574	-0.0011	-0.10

Electronic copy available at: https://ssrn.com/abstract=2822105

Appendix Table VI. Replicate Kelley and Tetlock (2013) Table IV

This table presents results from daily logistic regressions of earnings forecast errors on scaled retail imbalances (*Oibvol[0]*) and control variables. The dependent variable *PosFE[x,y]* is one if the analyst forecast error for quarterly earnings announcements occurring from day *t+x* through day *t+y* is positive and zero if the forecast error is negative. The forecast error is the difference between actual earnings-per-share and the median analyst forecast from I/B/E/S. At least 50 earnings announcements with corresponding forecast data during the window of the dependent variable are required for each daily logistic regression. Average coefficients and Newey and West (1987) t-statistics are reported with lags equal to twice the horizons of the dependent variable. We use the negative probability in TRNA as Neg, while Kelley and Tetlock (2013) use negative news from Dow Jones archives.

	Kelly and Tetlock (2013) Table IV				Our paper							
Reg	I		II		I		II		III		IV	
Dep.var	PosFE[1,5]		PosFE[6,20]		PosFE[1]		PosFE[1,3]		PosFE[1,5]		PosFE[6,20]	
Order Imbalance	Imb mkt		Imb mkt		Oibvol		Oibvol		Oibvol		Oibvol	
Ret	CRSP Return		CRSP Return		CRSP Return		CRSP Return		CRSP Return		CRSP Return	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	-1.614	-7.58	-1.664	-7.43	-1.334	-5.34	-3.197	-2.92	-2.783	-9.37	-2.758	-9.78
Oib[0]	0.126	5.73	0.047	2.61	**0.121**	**2.81**	0.043	1.39	0.035	1.30	0.007	0.38
Neg[0]	0.050	0.68	-0.015	-0.27	3.357	1.60	1.611	1.03	2.812	2.35	0.288	1.62
Neg[-5,-1]	-0.014	-1.00	-0.029	-2.23	0.249	0.49	0.054	0.18	0.043	0.12	-0.300	-1.85
Neg[-26,-6]	-0.093	-6.20	-0.110	-5.50	-0.599	-0.97	-1.465	-2.46	-0.651	-0.94	-1.398	-2.94
Ret[0]	0.033	6.60	0.028	9.33	5.608	3.92	5.785	6.20	3.017	5.19	1.951	4.88
Ret[-5,-1]	0.031	7.75	0.025	8.33	3.330	5.34	2.827	2.24	3.243	7.18	1.661	5.33
Ret[-26,-6]	0.024	8.00	0.014	7.00	2.166	6.14	2.945	2.64	1.799	7.51	1.227	5.82
Size	0.193	12.87	0.191	13.64	0.153	8.62	0.275	3.92	0.254	11.16	0.249	11.61
Lbm	-0.384	-5.05	-0.367	-4.59	-0.189	-3.68	0.082	0.33	-0.129	-2.74	-0.090	-2.36
Days	673		1193		289		566		745		1210	
Average R2	9.99%		8.29%		13.16%		11.06%		11.52%		8.98%	
Average N	265		482		125		229		302		585	

Electronic copy available at: https://ssrn.com/abstract=2822105