Understanding Retail Investors: Evidence from China

Charles M. Jones, Donghui Shi, Xiaoyan Zhang and Xinran Zhang*

First version: September 2019
This version: October 2021

Abstract

Using comprehensive account-level data from 2016 to 2019, we examine retail investor trading behavior in the Chinese stock market. We separate millions of retail investors into five groups by their account sizes and document strong heterogeneity in their trading dynamics and performance. Retail investors with smaller account sizes cannot predict future price movements correctly, in the sense that they buy future losers and sell future winners. These investors fail to process public news and display behavioral biases such as overconfidence and gambling preferences. In sharp contrast, retail investors with larger account balances predict future returns correctly, incorporate public news in their trading, and gain more in stocks which are more attractive to investors with behavioral biases. For liquidity provision, the smaller retail investors follow daily momentum strategies, demanding immediate liquidity, while they become contrarian over weekly horizons, and they contribute positively towards firm-level liquidity. On the contrary, larger retail investors ae contrarian at daily horizons, providing immediate liquidity, but their potentially informed trades demand liquidity over longer terms.

Keywords: Retail investors, Chinese stock market, return predictability, liquidity, information content.
JEL code: G12, G14, G15

* Charles M. Jones is with Columbia Business School, Donghui Shi is with Fanhai International School of Finance, Fudan University, Xiaoyan Zhang is with PBC School of Finance, Tsinghua University, and Xinran Zhang is with School of Finance, Central University of Finance and Economics. Xiaoyan Zhang acknowledges financial support from the National Natural Science Foundation of China [Grant 71790605]. We thank Laruen Cohen, Ron Kaniel, Zhiguo He, Hao Zhou, Utpal Bhattacharya, and seminar participants at Tsinghua PBC School of Finance, Renmin University, Shanghai Jiaotong University, Fudan University, Shanghai University of Finance and Economics, and conference audiences at the 2019 CIFFP, 2021 CFRC, 2021 CICF for helpful comments and suggestions. All remaining errors are our own. Corresponding author: Xiaoyan Zhang, PBC School of Finance, 43 Chengfu Road, Beijing, China, 100083, zhangxiaoyan@pbcsf.tsinghua.edu.cn.

Understanding Retail Investors: Evidence from China

First version: September 2019
This version: October 2021

Abstract

Using comprehensive account-level data from 2016 to 2019, we examine retail investor trading behavior in the Chinese stock market. We separate millions of retail investors into five groups by their account sizes and document strong heterogeneity in their trading dynamics and performance. Retail investors with smaller account sizes cannot predict future price movements correctly, in the sense that they buy future losers and sell future winners. These investors fail to process public news and display behavioral biases such as overconfidence and gambling preferences. In sharp contrast, retail investors with larger account balances predict future returns correctly, incorporate public news in their trading, and gain more in stocks which are more attractive to investors with behavioral biases. For liquidity provision, the smaller retail investors follow daily momentum strategies, demanding immediate liquidity, while they become contrarian over weekly horizons and contribute positively to firm-level liquidity. On the contrary, larger retail investors are contrarian at daily horizons, providing immediate liquidity, but their potentially informed trades demand liquidity over longer terms.

Keywords: Retail investors, Chinese stock market, return predictability, liquidity, information content.
JEL code: G12, G14, G15

Retail investors are important participants in financial markets, and many studies are devoted to understanding their trading motives, their performance, and their roles in information and price discovery. However, these studies provide seemly conflicting results. For instance, Barber and Odean (2000, 2001, 2008) document behavioral biases exhibited by retail investors, such as over-confidence and overtrading, and as a result, retail investors make sub-optimal investment choices. Later studies, such as Kaniel et al. (2008), Kelley and Tetlock (2013), and Boehmer et al. (2021), suggest that retail investors correctly predict future stock returns and trade accordingly, which indicates that retail investors might know something about future stock price movements. Most recently, interest has shifted to a new generation of retail investors, who tend to trade at zero-commission trading platforms such as Robinhood. Barber et al. (2021), Eaton et al. (2021) and Welch (2021) find that Robinhood investors perform well, demand liquidity and engage more in attention-induced trading. How can we reconcile the conflicting results from previous studies? One possibility is that retail investors are not born equal, so the above-mentioned empirical results could be dominated by subgroups of retail investors. However, due to data limitations, few previous studies directly examine the heterogeneity of retail investors.

China's equity market, the second largest in the world, provides an ideal setting for studying retail investors and their heterogeneity. According to the annual report of the Shanghai Stock Exchange, retail investors contribute 85% of daily trading volume on the exchange, while institutional investors only contribute 15%. The dominance of retail trading in this market clearly

brings retail investors to center stage. Behind the trading volume are tens of millions of retail investors in China, accounting for the largest population of retail investors in the global market. Given the dominant role and the large population of Chinese retail investors, it is crucial for researchers, regulators and practitioners to understand retail investors' investment choices and the resulting consequences for information and price discovery as well as market quality.

With account level data from one main stock exchange, we examine the rich cross-section of retail investors, which greatly helps us to investigate their heterogeneity of retail investors and how their trading interacts with stock returns, information flows and liquidity. We obtain account-level trading and holdings data from 2016 to 2019 for over 53 million retail accounts. To comply with regulatory requirements, all Chinese retail accounts are categorized into five groups by account balances: less than 100,000 CNY (RT1), between 100,000 and 500,000 CNY (RT2), between 500,000 and 3,000,000 CNY (RT3), between 3,000,000 and 10,000,000 CNY (RT4), and greater than 10,000,000 CNY (RT5). In terms of the numbers of accounts, the above five groups account for 58.7%, 28.6%, 10.9%, 1.4% and 0.4%, respectively. We also have additional gender and age information and find the majority of Chinese retail investors are young or middle-aged males with account sizes below 500k CNY.

With this rich cross section of retail investors, we first examine whether the buy and sell activity from retail investors can predict future price movements, as well as whether some of them are better informed. If the market is perfectly efficient, and if all investors have the same

information, stock prices would follow random walks, and trading would not predict future returns. If the market is not perfectly efficient, and if some investors have value-relevant information for future stock prices, their order flows should positively predict future returns. On the other hand, if some investors are behaviorally biased, misinformed or fail to incorporate timely information into their trades, their order flows might negatively predict future returns. Using daily retail order imbalances from each retail group, we predict future stock returns at horizons ranging from one day to 60 days. The smaller retail investors, RT1-RT4, predict next-day returns with negative coefficients. That is, the prices of stocks they buy experience negative returns the next day, while the ones they sell experience positive returns. In contrast, the largest retail investors, RT5, positively predict next-day returns and buy and sell stocks in directions consistent with future price movements. When we look at longer horizons, the above-mentioned predictive patterns persist for about one month. These patterns are also quite robust when we form long-short strategies on order flow information, and for subsets of stocks with differences in size, value, liquidity and share price level.

Previous literature provides multiple explanations for the trading motives for retail investors, such as order flow persistence, liquidity provision, behavioral biases and information (dis)advantages. These explanations also naturally connect with the predictive power of retail order flows for future returns. We adopt the two-stage decomposition procedure in Boehmer et al. (2021)

to examine whether these hypotheses can explain the trading decisions of different retail investor groups, and how these decisions contribute to the predictive patterns for future returns.

Our results show that order flows from all retail investors display persistence. Order flows from smaller retail investors show momentum patterns at a daily horizon and demand immediate liquidity, while the largest retail investors always display contrarian trading patterns. The smaller retail investors also display significant behavioral biases, such as over-confidence and gambling preferences, and they fail to predict and process earnings news. On the contrary, the largest retail investors trade against the behavioral biases of the other retail groups and are capable of predicting and processing earnings news. In explaining order flow's predictive power for future returns, order persistence, daily momentum trading, behavioral biases and information disadvantages all contribute to the negative predictive power of smaller retail investors, while contrarian trading, trading against behavioral biases and information skills contribute to the positive predictive power of the largest retail investors.

Classical theoretical work on noise traders, such as Black (1986), argue that noise traders, with no information advantage, are important to the investment community by providing liquidity, making trades possible and lowering transaction costs, while informed traders, with an information advantage, actually demand liquidity and raise transaction costs (see Stoll (1978), Grossman and Miller (1988), and Campbell, Grossman, and Wang (1993)). Do Chinese retail investors help to provide liquidity to the market? Our earlier results regarding momentum vs. contrarian trading

patterns provide indirect evidence on this question. We further examine whether retail order flows affect future effective spreads, a direct measure for liquidity and transaction costs. Our empirical results strongly support the theoretical predictions in the sense that the relatively uninformed trades from smaller retail investors significantly reduce future firm-level effective spreads and thus provide future liquidity, whereas the relatively informed trades from the largest retail investors significantly increase future firm-level effective spreads and demand future liquidity.

Finally, we investigate other dimensions of the data and conduct several robustness checks. We find that male investors across all ages negatively predict returns, especially the younger ones, while some female retail investors can positively predict future returns. These findings are generally in line with Barber and Odean (2001). We also aggregate order flows within each investor group and examine whether aggregate order flows can successfully time the market, or in other words can predict future market returns. We fail to find such evidence. Results from all robustness checks are consistent with the main results.

Our study is closely related to the retail investor literature. Previous studies on retail investors mostly use data from the U.S. and other developed countries, and they mostly treat retail investors as one group. For instance, using data from a discount broker in the U.S., Barber and Odean (2000, 2001, 2008) examine the trading and investment behavior of retail investors in the U.S. and document many behavioral biases. For return predictability, Kaniel, Saar and Titman (2008), Barber, Odean, Zhu (2009), Kelley and Tetlock (2013), and Boehmer et al. (2021) use different

datasets from the U.S. and find that retail trading can positively predict the cross-section of future returns. Using data from France, Barrot, Kaniel and Sraer (2016) provide further evidence that retail investors provide liquidity, especially during market downturns. Mostly recently, Barber et al. (2021), Eaton et al. (2021) and Welch (2021) study the trading behavior of Robinhood retail investors in the U.S.

Our study is also related to studies on the rapidly growing Chinese stock market. Liu, Stambaugh and Yuan (2019), and Liu, Zhou, and Zhu (2021) construct asset pricing factors. For Chinese retail investors, An, Lou and Shi (2020) study the wealth redistribution role of financial bubbles and crashes over July 2014 and December 2015, and they document a net transfer of 250 billion CNY from the poor to ultra-wealthy retail investors over this period. Liu, Peng, Xiong and Xiong (2021) and Liao, Peng and Zhu (2021) both focus on behavioral properties of Chinese retail investors and document overconfidence, gambling preferences and extrapolative expectations in these investors. Other studies, such as Li et al. (2017), Titman et al. (2020), Hu et al. (2021) and Jiang et al. (2020)[1] focus on an earlier Chinese sample period and examine behavioral biases and reactions to corporate events.

In comparison with these earlier studies, our study makes three important contributions. First, we separate retail investors into groups based on account sizes and provide unique, direct evidence on investor heterogeneity in terms of return predictability. Second, we examine different

[1] These papers include Li, Geng, Subrahmany and Yu (2017), Chen, Gao, He, Jiang and Xiong (2019), Jiang, Liu, Peng, and Wang (2020), Titman, Wei, and Zhao (2020), and Hu, Liu and Xu (2021).

hypotheses for the return prediction patterns for different retail investor groups, and we provide clear evidence on the sources of the negative or positive predictive power of different retail investors. Finally, we provide new evidence on how retail investors affect firm-level liquidity. Our study, with its large coverage of the market for a recent sample period, is one of the most thorough and comprehensive studies for of Chinese retail investors, and it provides many important implications for regulators, practitioners, and academic researchers.

The rest of the paper is organized as follows. Section I introduces the data. We examine whether order flow measures from different investor groups can predict future returns in Section II. In Section III, we investigate alternative hypotheses to explain the trading behavior of different investors and their various predictive patterns for future returns. Section IV focuses on how retail investors affect firm-level liquidity. We conduct robustness checks in Section V, and Section VI concludes.

I. Data

A. Data on Stock Returns and Firm Characteristics

We obtain data on stock returns, volumes, and accounting information from Wind Information Inc. (WIND), the largest financial data provider in China. To be consistent with our retail data, our sample period runs from January 2016 to June 2019. We adopt the filters in Liu, Stambaugh and Yuan (2019) and exclude stocks with less than 15 days of trading records during the most recent month. Liu, Stambaugh and Yuan (2019) also eliminate stocks that have become public within the

past six months, stocks with fewer than 120 days of trading records during the past 12 months, and the smallest 30% of total firms listed in the Chinese A-share market. We do not exclude these stocks for the main results, because retail investors trade actively in small stocks and during the IPO period. We present the results with all filters from Liu et al. (2019) in our robustness checks, and our findings are almost the same with the additional filters. Starting from March 31, 2010, margin buying and short selling are allowed on Chinese stock exchanges for subsets of stocks. We include these leveraged trades in our main results, and provide additional analysis excluding leveraged trading in our robustness checks. Our sample covers over 1.1 million stock-day observations, and on each day, we have an average of around 1,200 firms.

We present summary statistics on our sample firms in Panel A of Table 1. Daily stock returns are calculated using closing prices, which are dividend and split adjusted.[2] The average daily stock return, *Ret*, is -0.01% for Chinese stocks, while the average daily stock return is 0.04% in the U.S stock market over the same sample period. Market capitalization, *Size*, is computed as the product of the previous month's closing price and total A shares outstanding. The average Chinese firm

[2] Previous literature using the U.S. data shows that microstructure frictions can generate noise in daily return measures. For instance, Blume and Stambaugh (1986) show that daily returns computed from the end-of-day closing prices can have an upward bias due to bid-ask bounce. To assess the potential magnitude of the bias, they measure the bias as $\left(\frac{P_A-P_B}{P_A+P_B}\right)^2$, where P_A and P_B are closing ask and bid prices. Blume and Stambaugh (1986) find that the average bias for small stocks is 0.051%, and for large stocks, the bias is 0.001%, which are sizable magnitudes for daily returns averaging at less than 1%. Therefore, they recommend using closing bid-ask average prices to compute daily returns. We compute this bias measure using the closing bid and ask prices for all A-share stocks listed on the SHSE. The average bias measure is generally below 0.0002% across all stocks, which is negligible compared to the bias computed in Blume and Stambaugh (1983). Therefore, we compute daily returns using daily close prices without the Blume and Stambaugh (1983) adjustments.

capitalization is 20.1 billion CNY or 3 billion USD, about half of the cross-sectional average in the U.S stock market during the same period, which is 6.9 billion USD. The earnings to price ratio, *EP*, is computed as the ratio of the most recently reported quarterly net profit excluding non-recurrent gains/losses over last month-end's market capitalization. According to Liu, Stambaugh and Yuan (2019), the EP ratio captures the value effect. The average EP ratio is 0.0075 in China, while the average EP ratio is 0.0272 in U.S stock market. This difference may be driven by high valuations in China. Finally, monthly turnover is calculated as monthly share trading volume divided by tradable shares outstanding at the end of the previous month. The average monthly turnover in China is 48.32%, which is much larger than the monthly turnover of 22% in the U.S. during the same period.

B. Data on Retail Investors

We obtain investors' daily trading and holding data of all A-share stocks listed on the Shanghai Stock Exchange (SHSE hereafter) between January 2016 and June 2019. Out of the two stock exchanges in China, the SHSE and the Shenzhen Stock Exchange (hereafter SZSE), the SHSE accounts for 60% of the total market capitalization in China and thus is a good representation of the overall Chinese stock market.[3] Our data contains roughly 53 million accounts, and based on investor identities, they are first grouped into three major categories: retail (RT), institutional

[3] In June 2019, there are 1,471 A-share stocks listed on the SHSE, with a total market capitalization of $ 4.6 trillion. In comparison, 2,157 A-share stocks are listed on the SZSE, with a total market capitalization of $ 3 trillion. The Science and Technology Innovation Board (or STAR Market) was launched on the SHSE on July 22, 2019, and thus is not included in our study.

(INST), and corporations (CORP). Retail investors are further stratified into five groups based on their account sizes, which is the average portfolio value (including equity holdings in both SHSE and SZSE-listed firms, plus cash) over the previous twelve months. As mentioned in the introduction, there are five subgroups: below 100,000 CNY (RT1), 100,000-500,000 CNY (RT2), 500,000 - 3 million CNY (RT3), 3 million - 10 million CNY (RT4), and above 10 million CNY (RT5).

We merge the trading data and WIND data by stock ticker and present account summary statistics in Panel B of Table 1. During our sample period, the total number of active accounts for retail investors, institutions and corporations are 53.4 million, 40,000 and 47,000, respectively. Within the retail investor category, there are 31.4 million, 15.3 million, 5.8 million, 0.7 million and 0.2 million accounts for RT1 to RT5. Clearly, most of the retail investors have accounts less than 500,000 CNY. The overall trading volume on the SHSE averages 201 billion CNY per day, and retail investors, institutions and corporations account for 81%, 17% and 2% respectively of the total trading volume. Within the retail investor sector, trading volumes for RT1 to RT5 are 5%, 17%, 27%, 13% and 19% of the total trading volume, which is more evenly distributed than the numbers of accounts. For stock holdings, retail investors' holdings account for 22%, institutions 17% and corporations 62%. Within the retail investor sector, the account values for RT1 to RT5 are 1%, 4%, 6%, 3% and 7% of the total tradable market cap. Said another way, in the Chinese stock market retail investors dominate in terms of trading, while corporations dominate in holdings.

To understand firm-level trading behavior of different investor groups, we first sum up all trades on the same stock on the same day within each investor group, and then report the average daily trading volume (in shares) for each stock for each investor group. The average daily buy volumes at the individual stock level for RT1 to RT5 are 0.8 million shares, 2.7 million, 4.2 million, 2 million and 2.8 million, respectively, 2.3 million for institutions, and 0.3 million shares for corporations. On average, the daily sell volume in shares for different groups of investors are similar. To have a rough idea about holding horizons,[4] we make a simple assumption that shareholders within same investor group have identical holding horizons. Then we compute the holding period for stock *i*, type *G* investors as $1/TO_{i,G}$, where $TO_{i,G}$ is the turnover (shares traded/shares held by this type of investor) of stock *i* for type *G* investors. For example, if 1% of the shares trade each day, then it takes 100 days for the entire stock of tradable shares held by this group to turn over, and the average holding period would be 100 days. In the last row of Panel B, the average holding period for the five groups of retail investors ranges from 35 days to 50 days, reflecting their active trading and short holding horizons. Institutional holding periods in our sample are much longer at 109 trading days. Corporations barely trade in our sample period, and their estimated holding period is 6,319 trading days. In comparison, the market overall monthly turnover in the U.S. over the same period is 22%, indicating a holding period of 1/0.22 = 4.5

[4] In the U.S., there is a large amount of high frequency trading, including establishing and closing positions on the same day. China adopts the "*T+1* trading rule", which requires that if stocks are bought on day *T*, they cannot be sold on the same day. The reverse trade has to be executed on day *T+1* or later. That is, there is essentially a minimum holding period of one day.

months, which is about 90 days. These dramatic differences in holding horizons suggest that different types of Chinese investors might have quite different trading patterns and trading preferences.

To understand the relative importance of different investment groups' trading over time, we plot the time series of cross sectional means of various investors' trading activity in Figure I. Panel A presents each group's trading volume as a percentage of total trading volume. The RT3 group has the highest trading volume, accounting for about 30% of total trading. Interestingly, institutional trading gradually increases over time, from 10% in 2016 to over 20% in 2019. The corporations barely trade and account for a negligible amount of trading volume. Panel B displays the shares held percentage by each group, and the time-series patterns in holdings are quite stable.

We measure order flows from different groups of investors' using order imbalance measures, as in Chordia and Subramanyam (2004). For stock *i*, day *d*, and investor group *G*, we compute

$$Oib(i,d,G) = \frac{\sum_{j\in G} BuyVol(i,d,j) - \sum_{j\in G} SellVol(i,d,j)}{\sum_{j\in G} BuyVol(i,d,j) + \sum_{j\in G} SellVol(i,d,j)}, \quad (1)$$

where the numerator is the difference between buy and sell volumes summed up over all individual *j*'s within group *G*, and the denominator is the sum of buy and sell volumes of all individuals in group *G*. The order imbalance measure is literally an order flow measure because we actually observe each trade's direction. When a set of investors buys more than they sell, the order imbalance is positive, and vice versa. We compute the order imbalance measure for each investor

group as *OibRT*1 to *OibRT*5, *OibINST* and *OibCORP*. The overall retail order imbalance measure, *OibRT*, is calculated by summing up all trades within the five retail groups.

Table I Panel C reports summary statistics for the order imbalance measures. The average order imbalance for RT1 to RT5, institutions and corporations are -0.021, -0.011, -0.006, 0.002, 0.019, -0.003, -0.011 and -0.004, respectively.[5] The small magnitude of these average order imbalance measures indicates that most buys and sells within each investor group cancel out each other. The standard deviations of order imbalances are larger for large retail investors and institutions compared to small and medium retail investors, indicating that there is more cross-stock variation in large retail investor and institutional trading activity. The one-day autocorrelation coefficient, AR1, for these *Oib* measures are 0.243, 0.259, 0.216, 0.059 and 0.102 for RT1 to RT5, suggesting that small and medium retail order imbalances are generally more persistent than large retail imbalances.

In terms of order flow correlations across the seven groups, order flows from smaller retail investors, *OibRT*1, *OibRT*2 and *OibRT*3, are highly correlated, with correlation coefficients mostly higher than 0.60. *OibRT*4 is still positively correlated with *OibRT*1-*OibRT*3, but with a much lower correlation of0.20. The largest retail investors' order imbalance, *OibRT*5, is negatively correlated

[5] We plot the time-series of the cross-sectional mean, median and 25th and 75th percentiles of different types of investors' order imbalance in Appendix Figure I. There are no obvious time trends or structural breaks in the time series observations.

with all four other groups, with correlations around -0.15, indicating that this group of retail investors might have different trading patterns from the others. Given the large number of smaller retail accounts and their active trading, the overall retail trading, *OibRT*, is highly correlated with *OibRT*3. Institutional order imbalances are negatively correlated with all five retail groups, with correlations ranging from -0.380 to -0.188, again implying different trading patterns from retail investors, even the largest retail investors. As we saw earlier, corporations barely trade and their correlations with the rest of the investor categories are all lower than 10%.

In this data section, we always include corporations for the completeness of the summary statistics. Given that corporations are long-term investors and rarely trade, while our study focuses on trading behavior, we drop corporations from the remaining empirical results. Meanwhile, retail investors are commonly assumed to be less sophisticated investors than institutional investors. Therefore, to compare retail trading behavior with those of potentially more sophisticated investors, we keep institutions in our main empirical results as a benchmark. Finally, the overall retail trading variable, *OibRT*, is highly correlated with *OibRT*3, and thus we omit *OibRT* from our main empirical results to save space.

II. Can Retail Order Flows Predict Future Stock Returns?

Can retail investors' activity predict future stock returns in China? If they can, it is possible that these retail investors are informed about future stock price movements. We start by investigating the informational role of retail investors over the short term and the long term with

Fama-MacBeth regressions in Section II.A and II.B, respectively. In Section II.C, we examine the predictive patterns for different subsets of stocks based on firm and stock characteristics.

A. Predicting Next Day Stock Returns Using Retail Order Flows

To investigate the role different retail investors play in the price discovery process, we first examine the predictive power of various order flow variables for next-day returns using the two-stage Fama-MacBeth regression. For the first stage, we estimate the following cross-sectional regression for each day *d*,

$$Ret(i,d) = a0(d) + a1(d)Oib(i,d-1) + a2(d)'Controls(i,d-1) + u1(i,d), \qquad (2)$$

where the dependent variable $Ret(i,d)$ is the stock return for firm *i* on day *d*, and the independent variables include order imbalance measures from the previous day, $Oib(i,d-1)$, and control variables, $Controls(i,d-1)$. We follow previous literature for the choices of control variables. To control for potential momentum/reversal from past returns, we include returns from the previous day, *Ret*(-1), returns from the previous week, *Ret*(-6,-2), and returns from the previous month, *Ret*(-27, -7). For size, value and liquidity effects, we include log market size (*Size*), earnings-to-price ratio (*EP*), and turnover, all computed from the previous month-end.

From the first stage estimation, we obtain a daily time-series of coefficients, $\{a0(d), a1(d), a2(d)'\}$. For the second stage estimation, we conduct statistical inference based on the mean and standard errors of the first stage coefficients, and we compute Newey-West standard errors with 5 lags, which is the optimal lag number using a Bayesian Information Criterion (BIC).

If the order flow variable from a specific investor group predicts future returns in the right direction, in the sense that more past purchases are associated with higher future returns, and more past sales are associated with lower future returns, we expect the coefficient $a1$ to be significantly positive, and vice versa.

The estimation results for equation (2) are reported in Panel A of Table 2, which displays distinctive predictive patterns across different groups of retail investors. For the smallest retail investor group, RT1, the coefficient on retail order flow variable is -0.0093, with a significant *t*-statistic of -24.98. The negative coefficient shows that if retail investors RT1 buy more than they sell on a given day, the next day return on that stock is significantly negative. To understand the economic magnitude of the coefficients, we report the inter-quartile range for $OibRT1$ at the bottom. Multiplying the interquartile range, 0.2222, by the regression coefficient of -0.0093 generates an interquartile daily return difference of -21 basis points (more than 50% annualized!). For retail investors in groups RT2 to RT4, the predictive patterns are qualitatively similar. All coefficients are negative and statistically significant, and the daily interquartile return differences are -17, -11, and -2 basis points for RT2, RT3 and RT4, respectively. That is to say, the first four groups of investors all trade in the wrong direction vs. future price movements. Interestingly, when we move from the smaller account sizes to the larger ones, the negative coefficients become smaller, indicating that larger retail investors trade less incorrectly than smaller retail investors.

Indeed, for the largest retail investors, RT5, the coefficient on past day order imbalance is 0.0012, which is positive and significant with a *t*-statistic of 12.26. The interquartile daily return difference now is 5 basis points per day (over 12% per year). It seems that the largest retail investors' trading predicts the cross-section of future stock price movements in the correct direction.

As a comparison, the coefficient on the previous day order imbalance is 0.0016 for institutions, with a *t*-statistic of 20.34. That is to say, institutional order flows predict future stock price movements in the right direction, and the interquartile return difference is 10 basis points, about twice the magnitude of the RT5 estimate. This finding is consistent with many previous studies that institutional investors are more informed, and their trades contain more information than retail investor trades in general.

For the control variables, the coefficients on previous day return have mixed signs, while the coefficients on previous week and previous month returns are all negative and significant, indicating strong reversals over weekly and monthly horizons. Size is mostly insignificant, while the earnings-to-price variables are always positive and significant, indicating a strong value effect. The coefficients on turnover are always negative and significant, suggesting that higher turnover leads to lower returns in the future. The above findings are mostly consistent with previous studies of the Chinese stock market, such as Liu, Stambaugh and Yuan (2019). These results also confirm

that the predictive power of various order flow variables for future stock returns is not a manifestation of size, value, liquidity or momentum/reversal effect.

B. Predicting Long Term Stock Returns Using Retail Order Flows

The exercise in Section II.A focuses on next-day return prediction. It is natural to ask whether the predictive patterns carry on for longer terms. If the predictive pattern quickly vanishes or reverses, what we observe might be driven by short-term noise. If the predictive pattern persists over longer horizons, it is more likely the return predictability is linked to firm fundamentals or persistent biases. Therefore, we extend equation (2) to longer horizons up to 12 weeks:

$$Ret(i,w) = b0(w) + b1(w)Oib(i,d-1) + b2(w)'Controls(i,d-1) + u2(i,w). \qquad (3)$$

That is, we use previous day order imbalance, $Oib(i,d-1)$, to predict *w*-week ahead returns. To precisely observe the decay rate of the predictive power of order imbalance measures, we choose the dependent variable, $Ret(i,w)$, to be a weekly return for a specific 5-day period in week *w*, rather than a cumulative return from day *d* to the week *w*. In our empirical estimation, *w* ranges from one to 12 weeks. If order imbalances have only short-lived predictive power for future returns that then vanishes, we should observe the coefficient $b1$ decrease to zero quickly. Alternatively, if the specified retail order imbalance has longer predictive power, the coefficient $b1$ should remain statistically significant for a longer period.

We present the estimates of coefficient $b1$ in equation (3) in Table II Panel B. To save space, we only report the coefficients and the statistical significance level by asterisks, with ***, ** and

* indicating significance at 1%, 5% and 10% level, respectively. For the smallest retail investors RT1, the coefficient on *OibRT*1 monotonically decreases from -0.0226 at week one to -0.0005 at week 12, while the coefficients are statistically significant up to week nine. The negative predictive power of *OibRT*2, *OibRT*3, and *OibRT*4 also gradually decreases and becomes statistically insignificant around week seven. The positive predictive power of *OibRT*5 and *OibINST* diminishes a bit faster, and their statistical significance disappears at week four and five, respectively. Interestingly, while the magnitude and significance of the predictive coefficients decreases over the longer term, we observe little in the way of reversal patterns. The persistence of cross-sectional predictability indicates that the predictive power is rooted in information related to fundamentals or from persistent noise trading or behavioral biases.

C. Predicting Patterns Across Firms with Different Characteristics

Previous studies show that stock returns can be significantly affected by firm and stock characteristics, such as size, EP ratio and liquidity. Do predictive patterns of retail order flows differ across firms with different characteristics? To answer this question, we modify the specification in equation (2) by allowing different coefficients for firms with different characteristics,

$$Ret(i,d) = c0(d) + [c1(d)Dummy1(i,d-1) + c2(d)Dummy2(i,d-1) + c3(d)Dummy3(i,d-1)]Oib(i,d-1) + c4(d)'Controls(i,d-1) + u3(i,d). \quad (4)$$

Take size as an example. We first separate all firms on day *d* into three groups, based on previous month-end firm market capitalization. The dummy variable, $Dummy1(i, d-1)$, takes value 1 if firm *i* belongs to the smallest 1/3 of firms, zero otherwise; $Dummy2(i, d-1)$ takes value 1 if firm *i* belongs to the medium 1/3 of firms, zero otherwise; and $Dummy3(i, d-1)$ takes value 1 if firm i belongs to the largest 1/3 of firms, zero otherwise. The coefficients $c1, c2\ and\ c3$ provide information on whether the predictive pattern changes for firms with different sizes.

Estimation results for equation (4) are reported in Table III. In the first three rows, we separate firms by their market capitalization. The negative predictive pattern of order flow from RT1-RT4 for next day return, as observed in Table II, is quite robust for firms with different sizes. But it is interesting to notice that the magnitudes generally decrease from the smallest firms to the largest firms, indicating that the negative predictive pattern is the strongest for smaller firms. For the large retail investors, RT5, the positive predictive pattern remains for the small and medium-sized firms, but not for large firms, indicating that their information advantage, if any, might be concentrated in smaller firms. As a comparison, order flows from institutions significantly predict next day returns in all three rows, and more so for the large firms, suggesting that their information advantage, if any, might be more prominent for larger firms.

When we separate firms by EP, turnover and stock price, we observe similar interesting patterns. That is, the predictive patterns in Table II are generally robust across firms with different characteristics, and the negative (positive) predictive power of smaller (larger) retail investors is

stronger for small, low EP, and higher turnover firms, while the positive predictive power of institutional investors is stronger for large and high EP firms.

III. What Drives the Order Imbalance Predictive Power for Future Returns?

Previous literature provides several hypotheses for explaining investor order flows, and these might help to explain the heterogeneous predictive patterns from different investor groups for future returns. In Section III.A, we introduce a two-stage decomposition for the order flow's predictive power for future returns. We present the empirical results for the decomposition in Section III.B. We take a closer look at the information channel using event days in Section III.C.

A. A Two-Stage Decomposition to Explain Order Imbalance's Predictive Power

We consider four hypotheses for explaining the order flow dynamics and their predictive power for future stock returns. First, Chordia and Subrahmanyam (2004) state that order flows tend to be persistent, and persistent buying/selling pressure could lead directly to the predictability of future returns. Second, Kaniel, Saar, and Titman (2008) argue that retail traders in the U.S. are mostly contrarian, which provides liquidity to the market, so contrarian trades might positively predict future returns. Following this logic, if the retail trades tend to be momentum strategies, which demand liquidity, then it is possible that the momentum trades might negatively predict future returns. Third, Liu et al. (2021) connect retail trading motives to behavioral biases, and they find that over-confidence about information advantage and gambling preferences are the two dominant behavioral biases that affect trades of Chinese retail investors. Finally, Kelley and Tetlock (2013)

find that retail investors, especially the aggressive ones, may have valuable information about fundamental firm news, and thus their trading could correctly predict the direction of future returns. The above hypotheses are not mutually exclusive.

To find out whether the above hypotheses help to explain the trading behavior of different retail investor groups, and their predictive power for future stock returns, we follow the two-stage decomposition method as in Boehmer et al. (2021). For the first stage, we use the above hypothesis to explain the retail flow measures to find out which ones are important drivers for the order flows. This step also helps to decompose the retail order flows into hypothesis-implied components for each hypothesis. For the second stage, we investigate which of the hypothesis-implied components contributes to the predictive pattern of different investor order flow measures.

To estimate the two-stage decomposition, we first identify proxies for each hypothesis. The proxies for the first two hypotheses are relatively easy to construct. For the order-persistence hypothesis, we adopt the previous day order imbalance measure, *Oib*(-1), as the proxy. For the liquidity provision hypothesis, since it is directly linked to previous contrarian/momentum trading, we use returns from the previous day, week and month as proxies. For the overconfidence behavioral bias, we follow Barber et al. (2008) and Liu et al. (2021) and proxy it with turnover, computed as average daily turnover from the previous 20 days. For gambling preferences, we

follow Bail et al. (2011) and compute the maximum daily returns from the previous 20 days as the proxy.[6]

For the information hypothesis, the most influential information at the firm level is earnings surprise, hence we follow Kelley and Tetlock (2013) and measure firm-level information by the cumulative abnormal returns (CAR) over the earnings announcement period. To be more specific, assuming the earnings announcement day is day 0, we compute the cumulative returns over day -1 and day 0, and subtract the market returns over the same period to obtain cumulative abnormal returns. However, unlike the proxies for order persistence, liquidity provision and behavioral biases, which can be computed for each stock on each day, the news proxies are only available for 1.58% of stock-days, which would render our two-stage estimation imprecise. To cope with this missing data issue for news hypothesis, in this section we only consider the order persistence, liquidity provision and behavioral bias hypotheses, and we focus on the news hypothesis using an event-day approach in Section III.C.

At the first stage, for each day *d*, we estimate a cross-sectional specification,

$$Oib(i,d) = d0(d) + d1(d)Oib(i,d-1) + d2(d)'Ret(i,d-1) + d3(d)Overconf(i,d-1) + d4(d)Gamble(i,d-1) + u4(i,d). \quad (5)$$

[6] For gambling preference proxy, Liu et al. (2021) rely on events when the stock return hits 10% price limit, which is only available for 0.07% of the total sample. Therefore, we choose the maximum daily return from previous 20 days as an alternative for more data coverage. We also consider alternative proxies for gambling preferences, such as idiosyncratic volatility and skewness. The results are similar to those using maximum daily return, and are available on request.

After we obtain the time-series of coefficients, $\{\widehat{d0}(d), \widehat{d1}(d), \widehat{d2}(d)', \widehat{d3}(d), \widehat{d4}(d)\}$, we conduct statistical inference using the time-series means and standard errors, which are adjusted using Newey-West with five lags, in order to understand whether each hypothesis contributes to retail order flows. Meanwhile, the first stage estimation allows us to decompose *Oib(i,d)* into five components:

$$Oib(i,d) = \widehat{Oib}_{i,d}^{persistence} + \widehat{Oib}_{i,d}^{liquidity} + \widehat{Oib}_{i,d}^{overconf} + \widehat{Oib}_{i,d}^{gamble} + \widehat{Oib}_{i,d}^{other}, \quad (6)$$

with $\widehat{Oib}_{i,d}^{persistence} = \widehat{d1}(d)Oib(i,d-1), \widehat{Oib}_{i,d}^{liquidity} = \widehat{d2}(d)'Ret(i,d-1), \widehat{Oib}_{i,d}^{overconf} = \widehat{d3}(d)Overconf(i,d-1), \widehat{Oib}_{i,d}^{gamble} = \widehat{d4}(d)Gamble(i,d-1)$ *and* $\widehat{Oib}_{i,d}^{other} = \widehat{u4}(i,d-1) + \widehat{d0}(d-1)$. That is, the "persistence" part is related to the order persistence hypothesis, the "liquidity" part is related to the liquidity provision hypothesis, the "overconf" and "gamble" are both related to behavioral biases, and the "other" component is the residual component, which potentially contains other relevant information about future returns.

For the second stage of the decomposition, we relate future returns to each individual component of order flow by estimating the following specification using the Fama-MacBeth methodology:

$$Ret(i,d+1) = e0(d+1) + e1(d+1)\widehat{Oib}_{i,d}^{persistence} + e2(d+1)\widehat{Oib}_{i,d}^{liquidity} + e3(d+1)\widehat{Oib}_{i,d}^{overconf} + e4(d+1)\widehat{Oib}_{i,d}^{gamble} + e5(d+1)\widehat{Oib}_{i,d}^{other} + e6(d+1)'Controls(i,d) + u5(i,d+1). \quad (7)$$

With the decomposition in equation (6), the coefficient estimates in equation (7) show how each component of various order flows helps to predict future stock returns.

According to Boehmer et al. (2021), the advantage of the two-stage decomposition approach is that it includes components of $Oib(i,d)$ from alternative hypotheses in a unified and internally consistent empirical framework. In terms of caveats, we need to make empirical assumptions on proxies for different hypotheses. These empirical assumptions seem to us to be reasonable, but the interpretation of the results depends on the validity of our empirical assumptions.

B. Estimation Results for the Two-Stage Decomposition

We report first-stage estimation results in Table IV Panel A. In the first row, the coefficients on lagged order flow variables are always positive and significant, strongly supporting the order persistence hypothesis. For the next three rows, we connect order flows with returns from previous day, week and month, and the patterns are quite interesting. The order imbalances of RT1, RT2, and RT3 load positively and significantly on the previous day return, indicating that these investors buy more if the previous day return is positive, and sell more if the previous day return is negative. This corresponds to a daily momentum trading strategy, which demands immediate liquidity. For larger retail investors in RT4 and RT5, order imbalances load negatively and significantly on returns from the previous day, indicating that they are contrarian investors, buying low and selling high, and possibly providing immediate liquidity. If we extend the horizon to one week or one

month, then the coefficients on all returns are negative and significant, indicating that all retail investors follow contrarian strategies, buying losers and selling winners over the longer term.[7]

The next two rows present results on how behavioral biases are related to order flows. The coefficients on the overconfidence proxy are all positive and significant for RT1-RT4, indicating that overconfidence might be a strong driver for these retail investors' trading. Intriguingly, the magnitude of the coefficients gradually decreases from 0.0792 for RT1 to 0.0177 for RT4, implying a decreasing impact of overconfidence for retail investors as their account sizes increase. For the largest retail group, RT5, the coefficient becomes -0.0590 with a significant *t*-stat of -5.20. That is, the largest retail investors trade against this overconfidence behavioral bias, possibly because they don't have this bias, or because they provide liquidity to those with overconfidence. In terms of the gambling preference, for RT1-RT4 the coefficients are always positive and significant, indicating these retail investors like to buy stocks with lottery features. Interestingly, the coefficients gradually increase from 0.0467 for RT1 to 0.2583 for RT4, suggesting that larger retail investors have stronger gambling preferences. When we move on to RT5, the coefficient is -0.0863 with a significant *t*-stat of -3.97, which again means the largest retail investors trade

[7] Our finding that large retail investors are contrarian and smaller ones are momentum traders over daily horizon is quite interesting and different from some previous studies. For instance, contrarian patterns have been documented in Kaniel, Saar and Titman (2008) using monthly horizons in the U.S., and Barrot, Kaniel and Sraer (2016) using daily and weekly horizons in France. Using U.S. data, Kelley and Tetlock (2013) and Boehmer et al. (2021) both find that retail trades follow momentum over daily horizons, but are contrarian at weekly horizons. In our setting, we find the trading patterns from investors with smaller account sizes are similar to those in Kelley and Tetlock (2013) and Boehmer et al. (2021), while the investors with the largest account sizes behave similarly to the patterns documented in Kaniel et al. (2008) and Barrot et al. (2016).

against the gambling preference, possibly because they don't have gambling preferences, or because they provide liquidity to those with gambling preferences.

We report the second stage of the decomposition results in Panel B of Table IV. We take the first retail group, RT1, as an example. The coefficient estimate on *Oib(Persistence)* is -0.0338, with a *t*-statistic of -16.16, which implies that price pressure significantly and negatively contributes to the predictive power of RT1 trading flow. The coefficient estimate on *Oib(Liquidity)* is -0.0093, with a *t*-statistic of -2.63, which implies that momentum trading probably significantly and negatively contributes to the predictive power of RT1 trading flow. The coefficient of *Oib(Overconfidence)* is -0.1128, with a *t*-statistic of -2.86, and the coefficient for *Oib(Gamble)* is insignificant. For the *Oib(Other)* component, the coefficient is -0.0084, with a significant *t*-statistic of -27.58, indicating that there is other information, other than those incorporated in the three hypotheses, that contributes to RT1's negative predictive pattern for future returns. In terms of economic magnitude, we compute the interquartile range of all five components of the order imbalance measure. For the smallest retail group RT1, if we move from the 25th percentile to the 75th percentile in the distribution, the interquartile differences in future one-day stock return, for the *Oib(Persistence)*, *Oib(Liquidity)*, *Oib(O*verconf*)*, *Oib(Gamble)* and *Oib(Other)*, are -0.1179%, -0.0287%, -0.0162%, -0.0401%, -0.1782%, respectively. That is to say, order persistence, liquidity demand, overconfidence, and gambling preferences all contribute to the negative predictive power

of RT1 for next day returns. Similar patterns are observed for other smaller retail investor groups RT2-RT4.

If we turn our attention to the largest retail investors, RT5, the patterns are quite different. In terms of coefficient estimates, we find the persistence hypothesis and trading against over-confidence are both positive and significant. For the interquartile returns, all three hypotheses contribute to RT5's positive predictive pattern for future returns.

Overall, our decomposition exercise shows that a substantial part of the negative predictive power of the retail investors with smaller account sizes comes from order persistence, liquidity demand, and behavioral biases, while the positive predictive power of the retail investors with larger account balances comes from order persistence and trading against overconfidence and gambling preferences. The significance and the large magnitude of the "other" component indicates that existing hypotheses cannot fully explain the trading behaviors and their predictive power for returns. So what does "other" stand for? One possibility is information, which we take a close look at in the next subsection.

C. A Close Look at the Information Channel

It is important to understand how various retail investors participate in the information discovery process. However, earnings news only happens quarterly rather than daily, so the daily Fama-MacBeth estimation we adopt for the two stage estimation might not be proper for understanding how Chinese retail investors process information. As an alternative, in this section,

we focus on event days to study this issue. To capture each retail investor groups' participation in the information discovery process, we proceed in three steps: we first examine whether different retail investors can predict earnings news the next day, then we check whether they can process contemporaneous earnings news, and finally whether their predictive power for future returns improves or deteriorates on earnings event days.[8]

For this first step, to find out whether retail order flows can predict earnings news, we adopt the two stage Fama-MacBeth estimation. For the first stage, we estimate the follow specification for each quarter q:

$$CAR(i, d-1, d) = f0(q) + f1(q)Oib(i, d-2) + f2(q)'Controls(i, q-1) + u6(i, q). \quad (8)$$

Here we predict cumulative abnormal returns from day d-1 to day d, with day d being the earnings announcement day, using previous order imbalance measures from day d-2. We continue to use the Fama-MacBeth approach to estimate equation (8) and make inferences. Notice that each firm only has one earnings day each quarter, and equation (8) is estimated for each quarter. The second stage inference is based on the quarterly time-series of the estimated coefficients, and standard errors are computed using Newey-West with 4 lags. If retail order flows can predict earnings surprises in the right direction, the coefficient $f1$ should be significantly positive, and vice versa. Intuitively, for investors to be able to predict future returns around earnings announcements, they probably need access to private information.

[8] As an alternative to earnings news, we consider public news from CFND dataset in Section V.F. The results are qualitatively similar.

We present the estimation results in Panel A of Table V. For retail investors RT1-RT3, the coefficients $f1$ are -0.0251, -0.0234, and -0.0166, respectively, all with highly significant *t*-statistics. These negative and significant coefficients indicate that these investors incorrectly predict earnings surprises. In contrast, the coefficients $f1$ for RT5 and institutional investors are 0.0023, and 0.0034, both positive and statistically significant, implying that these investors are able to correctly predict future earnings surprises. In between these two extreme cases, the coefficient $f1$ for RT4 is close to zero and insignificant.

For the second step, we examine whether different retail groups can process contemporaneous news. Here the dependent variable is retail order flow, *Oib(i,d)*, and we connect it to contemporaneous earnings news, *CAR(i,d-1,d)*. The specification is similar to equation (8), except the timeline is different:

$$Oib(i,d) = g0(q) + g1(q)CAR(i,d-1,d) + g2(q)'Controls(i,d-2) + u7(i,q). \quad (9)$$

If a particular type of retail order imbalance can process contemporaneous earnings announcement news in the right direction, we expect the associated coefficient $g1$ to be significantly positive, and vice versa. Unlike the predictive specification in equation (8), equation (9) focuses on whether the investors under investigation have the ability to process public information rather than having access to private information.

Panel B of Table V report the estimation results. For retail investors RT1-RT4, the coefficients $g1$ are -1.9225, -1.8291, -1.4349, -0.8781 respectively, all with highly significant *t*-statistics. These

negative and significant coefficients indicate that these retail investor groups process the contemporaneous earnings announcement news in the wrong direction. In contrast, the coefficient $g1$ for RT5 is 0.1583, though insignificant, while the coefficient $g1$ for institutional investors is 2.7228 and significant, implying that these investors are able to correctly process contemporaneous earnings announcement news.

For the third step, we examine whether retail order flows' predictive power for future returns improves or deteriorates on event days to understand how much the information hypothesis helps to explain the return predictive patterns we observe in Section II. We estimate a modified version of equation (2), by adding the event day dummy and an interaction term:

$$Ret(i,d) = h0(d) + [h1(d) + h2(d)Event(i,d-1)]Oib(i,d-1) + \quad (10)$$
$$h3(d)Event(i,d-1) + h4'(d)Controls(i,d-1) + u8(i,d) \ .$$

Here the event dummy Event(i, d-1), is equal to one if the firm i has news on day d-1, and zero otherwise. For non-news days, the predictive power of retail trades is measured by coefficient $h1$; for news days, the predictive power is measured by ($h1$+$h2$). If coefficient $h2$ is significantly different from zero, that group of retail investors anticipates future stock returns differently on these news days. In the U.S., firm earnings announcements are chosen by firms and scattered throughout the year. In China, all firms are required to report their financial statements to regulators before four preset deadline dates each year. As a result, firms mostly announce their earnings within a short period before these deadline dates, and there would be zero announcements outside

of these short periods. To make sure that we have enough observations to estimate the Fama-Macbeth coefficients in equation (10), we only include days with at least 5% of total number of firms with earnings announcements, which gives us 68 days, or 8% of the total days in our sample.

The results are presented in Table V Panel C. Here we take the smallest retail investors, RT1, as an example. The coefficients on order imbalance, $h1$, is -0.0079 and is statistically significant, indicating that on average the trades from RT1 negatively predict future returns. When there is earnings announcement news, the coefficient on the interaction of event dummy and the order imbalance is -0.0080, with a significant *t*-statistic of -3.20, implying that the negative prediction of RT1 for future stock returns becomes significantly larger on earnings news days. This is consistent with our earlier finding that the smaller retail investors fail to predict and process the earnings news, which leads to more negative prediction for returns on event days. We observe similar patterns for RT2, RT3 and RT4.

For the largest retail investors, RT5, the coefficients $h1$ and $h2$ are 0.0005 and 0.0014, both statistically significant. That is to say, the large retail investors' predictive power for future returns is much stronger on earnings news days, possibly because these retail investors can correctly predict and process the earnings news, which enhances their ability to predict future stock returns. The pattern for institutional investors is quite similar to that of RT5.

Overall, our results reveal interesting heterogeneous patterns of how retail investors predict and process public information. On one hand, smaller retail investors are unable to predict future

news and lack skills to correctly process public news, while the largest retail investors and institutions are able to correctly process future earnings news and incorporate the contemporaneous news into their trading. The differences in information-processing abilities of different retail investors clearly contribute to the differences in their predictive powers for future returns.

IV. How Do Different Retail Order Flows Affect Future Liquidity?

The smaller retail investors, who predict future returns with negative signs, with poor information processing abilities and behavioral biases, behave similarly to the noise traders in classic works, such as Black (1986). Black (1986) argues that noise traders are important participants for a well-functioning capital market, because they provide liquidity and lower transaction costs for other investors. If all participants in the market are informed traders, then there probably would be no trades. Do Chinese retail investors provide liquidity and reduce transaction costs? Our earlier results show that smaller retail investors are momentum traders over daily horizons, and probably demand immediate liquidity, while the largest retail investors are contrarian at daily horizons and likely provide immediate liquidity. These trading patterns, momentum or contrarian, are indirect measures for contemporaneous liquidity provision. In this section, we directly examine how retail flows affect future firm level transaction costs and liquidity, over both the short and longer term.

In choosing the right measure for firm level transaction costs and liquidity, we follow Chen (2014) and Zhang et al. (2013), which examine Chinese stock market liquidity, and choose the relative effective spread. We compute the relative effective spread for each trade *k* as twice the distance between the trade price P_{ik} of stock *i* for trade *k* and the prevailing quote midpoint price, M_{ik}, scaled by the midpoint price, as

$$RES_{i,k} = 2|P_{ik} - M_{ik}|/M_{ik.} \quad (11)$$

The daily relative effective spread, $RES_{i,d}$ for stock *i* on day *d*, is computed as the average of effective spread for all trades during the day. Higher values for effective spread indicate higher transaction costs and lower liquidity and vice versa. To understand how retail flows affect future effective spreads, we adopt the empirical specifications from equation (2), and we use previous retail order flows to predict the next day's relative effective spread.

We report the estimation results for next day liquidity in Table VI Panel A. The coefficient on *OibRT*1 is -0.0099 with a highly significant *t*-statistic of -8.68, indicating that order flows from the smallest retail investors help to reduce next day effective spreads and increase firm level liquidity. Similar patterns exist for RT2 through RT4. For the largest retail investors, the coefficient is 0.0019 with a significant *t*-statistic of 6.18, indicating that informed trades from these investors actually reduce future short-term liquidity. A similar pattern is observed for institutions. That is, the smaller retail investors, with the wrong expectations for future price movements, help to increase future liquidity over next day; while the largest retail investors and institutional investors, being informed,

decrease liquidity over next day. The control variables all have consistent signs with previous studies in the literature.

In Panel B, we adopt the specification in equation (3) and examine whether retail order flows have an impact on longer-term liquidity. The coefficients on *OibRT*1 are negative and significant for week one to six, indicating that these smallest retail investors provide liquidity for the next month. Similar patterns are observed for all retail investors, suggesting that they all provide liquidity at the firm level within a month. This is consistent with our earlier finding that all retail investors are contrarian over weekly and monthly horizons and potentially provide liquidity. Interestingly, order flows from the largest retail investors, RT5, always have positive and significant coefficients for predicting future relative effective spreads for next 12-weeks. Notice that our earlier findings in Table II Panel B show that the largest retail investors can predict returns over the next 9 weeks, potentially demanding long-term liquidity, while the results in Table IV Panel A show that they are contrarian over monthly horizons, potentially supplying long-term liquidity. Combined, the positive coefficients here suggest that for the largest retail investors, their informed trading actually increases future transaction costs. For institutional investors, the coefficients are mostly positive but become insignificant at longer horizons.

To summarize, we provide direct evidence that different retail investors play different roles for future liquidity provision. The order flows from small retail investors decrease future

transaction costs within the month, indicating liquidity provision, while order flows from large retail investors and institutions are likely more informed and worsen future liquidity.

V. Further Discussions and Robustness

A. Ages and Genders

In this section, we examine heterogeneity through demographic differences, such as gender and age, of retail investors. According to Barber and Odean (2001), male investors could be more susceptible to behavioral biases, such as overconfidence and lack of attention. Due to the limited access to data, we only have a three-month sample period from January 2019 to March 2019 on investor gender and age. We first present summary statistics on age and gender in Table VII Panel A. Male investors contribute 66% of trading volume on average, and females account for 34%. Within the male group, the trading volume (%) across age groups below 35, between 35 to 45, between 45 to 55 and above 55 is 10%, 19%, 24% and 14% (summing to the 66% male total), while the trading volume (%) for the same age groups for females is 5%, 9%, 11% and 9% (summing to the 34% of volume traded by females). That is, across all gender-age groups, younger male investors trade the most.

Next, we examine the determinants of return prediction for each gender-age group specified in equation (2). The results are reported in Table VII Panel B. For return predictions, we find male investors across all ages negatively predict returns, with middle-aged males losing the most, while the youngest and oldest female retail investors (age less than 35 or above 55) can positively predict

future returns. These interesting patterns across age and gender provide further evidence regarding heterogeneity of retail investors.

B. Can Retail Order Flows Predict Market Returns?

So far we have focused on the relation between retail order flows at firm level and how they affect firm -level returns and liquidity. Given the dominance of retail investors in Chinese equity market, we are curious to understand how retail trading from each group is related to market-level conditions. In another word, if we aggregate the trades within each retail group, do they have material information about future market-level returns?

To answer this question, we first compute the aggregate order flows within each group *G* of investors on day *d* as $Aoib(d,G) = \frac{\sum_i BuyDollarVol(i,d,G) - \sum_i SellDollarVol(i,d,G)}{\sum_i BuyDollarVol(i,d,G) + \sum_i SellDollarVol(i,d,G)}$. Notice here we use dollar volumes, which are the products of share volumes and closing prices, rather than share volumes asin equation (1). The original firm level oib measure in equation (1) doesn't involve closing prices, because it is only for one firm. Both numerator and denominator are measured in shares and are directly comparable. Here we aggregate trading volume across different stocks, and we need to make the numbers comparable across firms by using dollar volumes.

To find out whether aggregate order flows are related to future market conditions, we estimate the following time-series specifications,

$$Mkt(d+k) = m0 + m1 \times Aoib(d,G) + m2 \times Y(d) + u9(d+k). \qquad (12)$$

That is, we use aggregate order imbalance for investor group G on day d, $Aoib(d,G)$, to predict k-days ahead market returns, k=1, …, 5. Here we compute market return, *Mkt*, as the daily value-weighted average of firm-level returns.

We report the results in Table VII Panel C. For the smallest retail investors, RT1, the coefficients on *AoibRT*1 are mostly negative and insignificant, showing no evidence of predictive power. Similar patterns are observed for RT2. For RT3 and RT4, it is interesting to find they both have negative and significant coefficients over a one-day horizon, indicating that they predict the market return in the wrong direction. The largest retail investors, RT5, also fail to exhibit significant predictive power for market returns. In contrast, the coefficient for institutional investors at one-day horizon is 0.0132 with a significant *t*-statistic of 2.84, suggesting some predictive power, but the coefficient quickly turns negative and insignificant after three days.

C. Applying stricter filters from Liu, Stambaugh and Yuan (2019)

In this study, we apply a filter from Liu et al. (2019) and discard stocks with less than 15 days of trading during the most recent month. In addition, Liu et al. (2019) also eliminate stocks that have become public within the past six months, stocks with less than 120 days of trading during the past 12 months, and the smallest 30% of firms listed in SHSE and SZSE. We add all these additional filters and check the robustness of our results.

In Table VII Panel D, the order imbalance prediction directions are similar to the results in Table II. The first four groups of retail investors tend to trade in the wrong direction for future price

movements, while the largest retail investor group RT5 and institutions trade in the same direction as the cross-section of future stock returns. The economic magnitudes for the first four type of retail investors are quantitatively similar, while RT5's economic magnitude is only half as large when adding these additional filters, perhaps because RT5's positive return mainly comes from small stocks. The economic magnitude for institutions is still large. In conclusion, our main results are robust to the stricter filters from Liu, Stambaugh and Yuan (2019).

D. Leveraged positions

Our trade level data also identify investors' margin buys, short sales and collateral trades. Leveraged trading may be different from non-leverage trading. On each day, margin buys account for 10% of the trading volume, short sales account for 0.2% and collateral trading accounts for 15% during our sample period. We exclude the leverage trades and re-estimate equation (2).

Results are reported in Table VII Panel E. The order imbalance prediction directions are similar to the results in Table II. The first four groups of retail investors trade in the wrong direction of future price movements, while the largest retail investor group RT5 and institutions trade in a way that positively predicts the cross-section of future stock returns. The economic magnitudes are quantitatively similar. In conclusion, our results are robust to whether or not we include these leverage trades.

E. Forming Portfolios Using Retail Order Flows

Our main results in previous sections are based on regressions, which assumes linear relations between the future returns and order flow variables. In this section, we adopt an alternative portfolio approach and examine the robustness of the previous results. To be more specific, we sort firms into five groups, based on previous day's order imbalance from a particular investor group, buy the 20% of stocks with the highest order imbalance measures, and short the 20% of stocks with the lowest order imbalance measures. We report the risk adjusted returns (alphas) on this long-short strategy for one to 60 days, where we conduct risk adjustment by using the Liu, Stambaugh and Yuan (2019) three factor model.

From Table VII Panel F, the one-day long-short portfolio alpha, using the previous day order imbalance from RT1, is -0.0042, and highly significant. The weekly (5-day) alpha for the long-short portfolio is -0.0089. When we increase the holding horizon to 60 days, the average alpha becomes -1.83%, and still significant. The general pattern is that cumulative holding-period alphas and returns continue to grow in magnitude in general. We observe no evidence of a reversal in return predictability. Similar patterns exist for RT2 and RT3. For larger retail investors in RT4, the one-day alpha is negative at -0.0007, but it becomes insignificant for horizons longer than one day. For the largest retail investors in RT5, the one-day alpha is 0.0017, positive but insignificant, while the longer horizon alphas are also positive and significant, indicating that RT5 trades have positive returns and do not reverse in the long run. For comparison, the long-short strategy following institutional order imbalance generates positive and significant returns for the next one to 60 days,

ranging between 25 basis points and 1.37%. We also observe no evidence of return reversals for these long-short portfolios, implying the information in institutional trading is persistent, consistent with previous literature that institutional trading is more informed.

F. News from CFNDS

Our earlier results show that the smaller retail investors lack skills to predict or process public earnings news, while the largest retail investors and institutions are able to correctly predict and process future earnings news and incorporate the contemporaneous news into their trading. In this section, we use an alternative public news dataset to investigate whether the results from earnings news can be extended to other news. We obtain news data from the Financial News Database of Chinese Listed Companies (CFND), which includes news on all A-share stocks from over 400 internet media and over 600 newspapers. In comparison with earnings news, the data coverage is more substantial, but the news content is more diverse.

We estimate equation (10) and report the results in Table VII Panel G. For the smallest retail investors, RT1, the coefficient on order flow is -0.0075 and highly significant, confirming that their order flows predict returns negatively. The coefficient on the interaction between order flow and the event dummy is -0.0045, again highly significant, suggesting the negative predictive power is significantly stronger on news days, which is consistent with our results in Section III.C. Similar patterns are observed for RT2-RT4. For the largest retail investors, RT5, the coefficient on order flow is 0.0008, and on the interaction is 0.0011, both highly significant, indicating that the RT5

order flow on average predicts future returns in the correct direction, and their prediction becomes much stronger on news days. To summarize, we confirm with an alternative news dataset that smaller retail investors lack skills to process public earnings news, and their negative predictions for future returns are worse on news days, while the largest retail investors and institutions are able to correctly process future earnings news and enhance their predictive power for future returns.

IV. Conclusion

Using comprehensive account-level trading and holding data from 2016 to 2019, we separate tens of millions of retail investors into five groups by their account sizes, and examine heterogeneity in retail investors' return predictabilities, sources of the return predictabilities and influence on market liquidity.

We provide strong and direct evidence on retail investors' heterogeneity. Retail investors with account sizes less than 3mil CNY buy and sell stocks in the wrong directions. The prices of stocks they buy experience negative returns the next day, while the ones they sell experience positive returns. For retail investors with large account balances, their trading predicts returns in the correct direction. In tracing their differences in predicting future returns, we provide evidence that the negative predictive power of the retail investors with smaller account sizes are mostly related to their order persistence, daily momentum trading, behavioral biases and failures in processing earnings news. In contrast, the positive predictive power of the large retail investors is mostly associated with order persistence, contrarian trading, trading against behavioral biases and

advantages in processing earnings news. We further find the order flows from smaller retail investors significantly reduce future transaction costs and improve liquidity.

Our results on the heterogeneity of retail investors help to understand the conflicting empirical results in the previous literature regarding retail investors. In addition, it is interesting that the exchange itself acknowledges the heterogeneity in retail investors and is focused on adopting policies on suitability that restrict some kinds of trading for the smallest accounts. For example, the Shanghai Stock Exchange requires a retail investor to have at least 500k CNY holdings of stocks for at least 20 trading days to open a leverage trading account or to trade on the riskier Science and Technology Innovation Board (or STAR Market). These policies effectively exclude the smallest retail investors from leverage trading and trading on riskier start-ups, which could help protect these small retail investors from even worse losses.

References

An, Li, Dong Lou, and Donghui Shi, 2021, Wealth redistribution in bubbles and crashes, Working paper, Tsinghua University.

Bali, Turan G., Nusret Cakici, and Robert F. Whitelaw, 2011, Maxing out: Stocks as lotteries and the cross-section of expected returns, *Journal of Financial Economics*, 99(2), 427-446.

Barber, Brad M., Xing Huang, Terrance Odean, and Christopher Schwarz, 2021, Attention induced trading and returns: Evidence from Robinhood users, Working paper, University of California.

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, 2001, Boys will be boys: Gender, overconfidence, and common stock investment, *The Quarterly Journal of Economics*, *116*(1), 261-292.

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.

Black, Fischer, 1986, Noise, *Journal of Finance*, 41(3), 528-543.

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, Xiaoyan Zhang, and Xinran Zhang, 2021, Tracking retail investor activity, *Journal of Finance*, forthcoming.

Campbell, John Y., Sanford J. Grossman, and Jiang Wang, 1993, Trading volume and serial correlation in stock returns, *Quarterly Journal of Economics*, 108, 905–939.

Chen, Hui, 2014, Comparison and application of two kinds of bid-ask spreads from daily data (Translated from Mandarin), *Chinese Review of Financial Studies*, 3, 80-90.

Chen, Ting, Zhenyu Gao, Jibao He, Wenxi Jiang, and Wei Xiong, 2019, Daily price limits and destructive market behavior, *Journal of Econometrics*, 208(1), 249-264.

Chordia, Tarun, and Avanidhar Subrahmanyam, 2004, Order imbalance and stock returns: Theory and evidence, *Journal of Financial Economics*, 72, 485−518.

Eaton, Gregory W., T. Clifton Green, Brian Roseman, and Yanbin Wu, 2021, Zero-commission individual investors, high frequency traders, and stock market quality, Working paper, Oklahoma State University.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy*, 81, 607−636.

Grossman, Sanford J., and Merton H. Miller, 1988, Liquidity and market structure, *Journal of Finance*, 43, 617–633.

Hu, Conghui, Yu-Jane Liu, and Xin Xu, 2021, The valuation effect of stock dividends or splits: Evidence from a catering perspective, *Journal of Empirical Finance*, 61, 163-179.

Jiang, Lei, Jinyu Liu, Lin Peng, and Baolian Wang, 2019, Investor attention and asset pricing anomalies, Working paper, Tsinghua University.

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.

Li, Xindan, Ziyang Geng, Avanidhar Subrahmanyam, Honghai Yu, 2017, Do wealthy investors have an informational advantage? Evidence based on account classifications of individual investors, *Journal of Empirical Finance* 44, 1-18.

Liao, Jingchi, Cameron Peng, and Ning Zhu, 2021, Extrapolative bubbles and trading volume, *Review of Financial Studies*, forthcoming.

Liu, Jianan, Robert F. Stambaugh, and Yu Yuan, 2019, Size and value in China, *Journal of Financial Economics*, 134(1), 48-69.

Liu, Hongqi, Cameron Peng, Wei A. Xiong, and Wei Xiong, 2021, Taming the bias zoo, *Journal of Financial Economics*, forthcoming.

Liu, Yang, Guofu Zhou, and Yingzi Zhu, 2021, Trend factor in China: The role of large individual trading, Working paper, Tsinghua University.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, 55, 703−708.

Stoll, Hans R., 1978, The supply of dealer services in securities markets, *Journal of Finance*, 33, 1133–1151.

Titman, Sheridan, Chishen Wei, and Bin Zhao, 2020, Stock price manipulation: Corporate actions and the exploitation of retail investors in China, Working paper, University of Texas at Austin.

Welch, Ivo, 2020, The wisdom of the Robinhood crowd, *Journal of Finance*, forthcoming.

Zhang, Zheng, Yizong Li, Yulong Zhang, Xiang Liu, 2013, A test on indirect liquidity measures in China stock market: An empirical analysis of the direct and indirect measure of bid-ask spread (Translated from Mandarin), *China Economic Quarterly*, 13(1), 233-262.

Table I. Summary statistics

This table reports summary statistics for stock characteristics, trading and holdings by different investor groups. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trades during the previous month. Panel A reports the time series average of the cross-sectional distribution of stock level characteristics, daily stock return (*Ret*), market capitalization (*Size*), earnings to price ratio (*EP*), and monthly turnover (*Turnover*). Panel B shows the number of accounts, aggregate trading and holdings, and average stock-level buy and sell volume by different investor groups. The holding horizon is the shares held by each type of investor divided by shares traded by this type of investor and captures how many days on average this type of investor takes to turn over a position. Panel C reports the time series average of the cross-sectional mean, standard deviation, autocorrelation (AR1), and cross correlations of order imbalances by different investor groups. Order imbalances (*Oib*) are computed as the buy share volume minus sell share volume divided by buy share volume plus sell share volume for each investor group, as specified in equation (1).

Panel A. Stock characteristics

	Variable description	Mean	Std	P25	P50	P75
Ret	Daily Stock Return	-0.01%	2.17%	-1.09%	-0.22%	0.77%
Size	Market Capitalization (Billion CNY)	20.1	80.3	2.9	5.6	12.1
EP	Earnings to Price Ratio	0.0075	0.0155	0.0018	0.0060	0.0122
Turnover	Monthly Turnover (of tradable A shares)	48.32%	72.48%	14.09%	25.40%	49.97%

Panel B. Number of accounts, trading and holdings by different types of investors

Account value	RT1 <100K CNY	RT2 (100K,500K) CNY	RT3 (500K,3M) CNY	RT4 (3M,10M) CNY	RT5 >10M CNY	INST	CORP
Number of Accounts (thousands)	31,410	15,282	5,827	735	235	40	47
Aggregate Trading Volume (Bil. CNY)	9	35	54	27	37	35	3
Aggregate Trading Volume (% of total)	5%	17%	27%	13%	19%	17%	2%
Aggregate Holdings Value (Bil CNY)	336	951	1,566	840	1,794	4,201	15,547

Aggregate Holdings Value (% of total)	1%	4%	6%	3%	7%	17%	62%
Stock level Buy Share Volume (Mil.)	0.801	2.709	4.180	2.052	2.777	2.344	0.255
Stock level Sell Share Volume (Mil.)	0.799	2.698	4.179	2.051	2.777	2.366	0.259
Holding Horizon (Days)	50	36	35	35	49	109	6,319

Panel C. Order imbalance in the cross section by investor group

	Mean	Std	AR1	Correlations							
				OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibRT	OibINST	OibCORP
OibRT1	-0.021	0.187	0.243	1							
OibRT2	-0.011	0.171	0.259	0.802	1						
OibRT3	-0.006	0.166	0.216	0.610	0.710	1					
OibRT4	0.002	0.250	0.059	0.194	0.244	0.256	1				
OibRT5	0.019	0.352	0.102	-0.151	-0.158	-0.163	-0.091	1			
OibRT	-0.003	0.113	0.272	0.512	0.604	0.642	0.447	0.342	1		
OibINST	-0.011	0.455	0.205	-0.315	-0.365	-0.380	-0.263	-0.188	-0.615	1	
OibCORP	-0.004	0.720	0.088	0.022	0.029	0.021	-0.007	-0.043	-0.032	-0.044	1

Table II. Predicting Future Stock Returns Using Order Imbalances from Different Investor Groups

This table reports estimation results on whether trading activity by different investor groups can predict the cross section of one-day-ahead returns and returns over the returns over the next 12 weeks. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trading during the previous month. In Panel A, we present coefficient estimates from Fama-MacBeth (1973) regressions as specified in equation (2). Panel B reports coefficient estimates from Fama-MacBeth (1973) regressions specified in equation (3). The independent variables are the previous day order imbalance *Oib(-1)*, and the control variables are the previous day return *Ret(-1)*, the previous week return *Ret(-6,-2)* and the previous month return *Ret(-27,-7)*, previous month log market cap (*Size*), earnings to price ratio (*EP*) and monthly turnover (*Turnover*). For each regression in Panel A, we also provide the interquartile range for the relevant explanatory order imbalance to compute the difference in predicted future returns for the interquartile range (*Interquartile return*). To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 5 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. Predict next day return

Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1)	Estimate	-0.0093***	-0.0091***	-0.0065***	-0.0009***	0.0012***	0.0016***
	[*t*-stat]	[-24.98]	[-22.58]	[-18.50]	[-7.21]	[12.26]	[20.34]
Ret(-1)	Estimate	-0.0027	-0.0091**	0.0006	0.0189***	0.0190***	0.0132***
	[*t*-stat]	[-0.62]	[-2.07]	[0.13]	[4.06]	[4.13]	[2.79]
Ret(-6,-2)	Estimate	-0.0149***	-0.0132***	-0.0124***	-0.0120***	-0.0115***	-0.0113***
	[*t*-stat]	[-8.06]	[-7.07]	[-6.62]	[-6.37]	[-6.13]	[-6.04]
Ret(-27,-7)	Estimate	-0.0039***	-0.0036***	-0.0034***	-0.0033***	-0.0032***	-0.0034***
	[*t*-stat]	[-4.36]	[-4.04]	[-3.86]	[-3.72]	[-3.62]	[-3.85]
Size	Estimate	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	[*t*-stat]	[0.36]	[0.17]	[-0.16]	[-0.32]	[-0.18]	[-0.21]
EP	Estimate	0.0147***	0.0150***	0.0145***	0.0144***	0.0146***	0.0140***
	[*t*-stat]	[3.54]	[3.57]	[3.41]	[3.42]	[3.47]	[3.34]

Turnover	Estimate	-0.0007***	-0.0007***	-0.0007***	-0.0007***	-0.0007***	-0.0007***
	[*t*-stat]	[-3.47]	[-3.63]	[-3.69]	[-3.83]	[-3.83]	[-3.83]
Intercept	Estimate	-0.0012	-0.0006	0.0002	0.0005	0.0001	0.0002
	[*t*-stat]	[-0.48]	[-0.26]	[0.06]	[0.19]	[0.06]	[0.10]
Adj.R2		8.83%	8.68%	8.43%	8.10%	8.11%	8.25%
Interquartile		0.2222	0.1827	0.1678	0.2868	0.4536	0.6740
Interquartile return		-0.2062%	-0.1668%	-0.1089%	-0.0247%	0.0523%	0.1046%

Panel B. Predict returns over the next 12 weeks

Week	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
w=1	-0.0226***	-0.0220***	-0.0144***	-0.0019***	0.0027***	0.0044***
w=2	-0.0065***	-0.0060***	-0.0037***	0.0001	0.0010***	0.0012***
w=3	-0.0038***	-0.0031***	-0.0015**	0.0001	0.0007***	0.0007***
w=4	-0.0024***	-0.0021***	-0.0012*	-0.0001	0.0007***	0.0005**
w=5	-0.0014***	-0.0014***	-0.0011**	-0.0005**	0.0004**	0.0002
w=6	-0.0029***	-0.0024***	-0.0016***	-0.0003	0.0001	0.0005***
w=7	-0.0027***	-0.0025***	-0.0018***	-0.0002	0.0001	0.0007***
w=8	-0.0015***	-0.0010*	-0.0007	-0.0002	0.0003**	0.0004***
w=9	-0.0010**	-0.0005	-0.0004	0.0002	0.0004**	0.0001
w=10	-0.0007	-0.0004	-0.0004	0.0002	-0.0001	0.0002
w=11	-0.0006	-0.0007	-0.0001	0.0000	-0.0002	0.0002
w=12	-0.0005	-0.0001	0.0005	0.0001	0.0001	0.0000

Table III. Predicting Next Day Stock Returns for Different Subgroups of Firms

This table reports estimation results on whether trading activity by different investor groups can predict the cross section of one-day-ahead returns for firms with different characteristics. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trading during the previous month. We present coefficient estimates from Fama-MacBeth (1973) regressions in equation (4). The dependent variable is the return on day *d*. The independent variables are the previous day order imbalance *Oib(-1)*, and the control variables are the previous day return *Ret(-1)*, the previous week return *Ret(-6,-2)* and the previous month return *Ret(-27,-7)*, previous month log market cap (*Size*), earnings to price ratio (*EP*) and monthly turnover (*Turnover*). To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 5 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Oib.var	Coefficients	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Small Size	$c1$	-0.0119***	-0.0131***	-0.0090***	-0.0006***	0.0020***	0.0014***
Medium Size	$c2$	-0.0096***	-0.0093***	-0.0066***	-0.0011***	0.0009***	0.0015***
Large Size	$c3$	-0.0067***	-0.0061***	-0.0044***	-0.0012***	0.0001	0.0021***
Low EP	$c1$	-0.0122***	-0.0131***	-0.0095***	-0.0010***	0.0020***	0.0015***
Medium EP	$c2$	-0.0097***	-0.0098***	-0.0069***	-0.0009***	0.0012***	0.0015***
High EP	$c3$	-0.0062***	-0.0056***	-0.0039***	-0.0007***	0.0001	0.0017***
Low Turnover	$c1$	-0.0061***	-0.0057***	-0.0041***	-0.0007***	0.0003***	0.0013***
Medium Turnover	$c2$	-0.0091***	-0.0092***	-0.0066***	-0.0010***	0.0011***	0.0014***
High Turnover	$c3$	-0.0159***	-0.0176***	-0.0128***	-0.0009***	0.0026***	0.0019***
Low Price	$c1$	-0.0088***	-0.0086***	-0.0067***	-0.0012***	0.0009***	0.0014***
Medium Price	$c2$	-0.0098***	-0.0095***	-0.0068***	-0.0007***	0.0013***	0.0016***
High Price	$c3$	-0.0094***	-0.0094***	-0.0060***	-0.0007***	0.0014***	0.0016***

Table IV. Two Stage Decomposition for Understanding the Predictive Patterns of Order Flow Variables

This table reports estimation results on a decomposition of the predictive power of different investor groups' order imbalance for the cross-section of future stock returns. Our sample period covers January 2016 to June 2019, and the sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 trading days in the previous month. We estimate two-stage Fama-MacBeth (1973) regressions. Panel A reports the first-stage estimation results, where the order imbalance measures are decomposed into five components as specified in equation (5). Panel B reports the second-stage decomposition of order imbalance's predictive power, as specified in equations (6) to (7). The dependent variable is the return on day *d*. As independent variables, the variable *Oib(*-1,*Persistence)* is estimated in the first stage using past order imbalance and reflects price pressure. The variable *Oib(*-1,*Liquidity)* is estimated in the first stage using past returns over different horizons and is connected to the liquidity provision or liquidity demand hypothesis. The variable *Oib(*-1,*Overconfidence)* is estimated in the first stage using stock turnover from the previous 20 days and reflects overconfidence. The variable *Oib*(-1,*Gamble*) is estimated in the first stage using maximum daily returns from previous 20 days and reflects a preference for gambling. The residual part of the previous day order imbalance from the first-stage estimation is denoted "other," which can be attributed to private information about future returns. As control variables, we include previous day return, *Ret*(-1), previous week return, *Ret*(-6,-2), and previous month return, *Ret*(-27, -7), previous month log market cap (*Size*), earnings to price ratio (*EP*) and monthly turnover (*Turnover*). 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 report the difference in predicted day-ahead returns for observations at the two ends of the interquartile range (*Interquartile return*) in Panel B. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 5 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. First stage of projecting order imbalance on persistence, past returns, overconfidence and gambling proxies

		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1)	Estimate	0.1867***	0.1965***	0.1711***	0.0499***	0.1037***	0.2220***
	[*t*-stat]	[40.80]	[48.27]	[49.37]	[21.67]	[39.62]	[65.06]
Ret(-1)	Estimate	0.5131***	0.7244***	0.4469***	-0.2210***	-1.2907***	-2.1369***
	[*t*-stat]	[15.62]	[28.96]	[21.74]	[-9.79]	[-39.26]	[-31.23]
Ret(-6,-2)	Estimate	-0.4186***	-0.2186***	-0.1047***	-0.0788***	-0.0513***	0.1322***
	[*t*-stat]	[-28.53]	[-19.63]	[-10.78]	[-6.79]	[-3.97]	[4.43]
Ret(-27,-7)	Estimate	-0.0325***	-0.0171***	-0.0205***	-0.0377***	-0.0231***	0.0271***
	[*t*-stat]	[-7.13]	[-4.67]	[-6.83]	[-10.73]	[-5.00]	[2.78]
Overconf(-1)	Estimate	0.0792***	0.0366***	0.0318***	0.0177*	-0.0590***	0.1599***
	[*t*-stat]	[5.16]	[3.34]	[4.24]	[1.87]	[-5.20]	[8.35]
Gamble(-1)	Estimate	0.0467***	0.1025***	0.1991***	0.2583***	-0.0863***	-0.6651***
	[*t*-stat]	[2.61]	[7.48]	[16.00]	[15.70]	[-3.97]	[-13.94]
Intercept	Estimate	-0.0216***	-0.0121***	-0.0115***	-0.0078***	0.0232***	0.0073
	[*t*-stat]	[-7.05]	[-5.90]	[-8.98]	[-5.18]	[12.06]	[1.32]
Adj.R2		7.08%	5.59%	3.88%	0.74%	2.02%	7.36%

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

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1,Persistence)	Estimate	-0.0338***	-0.0277***	-0.0228***	0.0032	0.0054***	0.0056***
	[*t*-stat]	[-16.16]	[-15.50]	[-13.49]	[0.57]	[5.72]	[17.09]
Oib(-1,Liquidity)	Estimate	-0.0093***	-0.0188***	-0.0233***	-0.0130	-0.0007	0.0061***
	[*t*-stat]	[-2.63]	[-4.59]	[-3.37]	[-1.48]	[-0.24]	[3.53]
Oib(-1,Overconf)	Estimate	-0.1128***	-0.0724	-0.2876***	-0.1014	0.0492*	0.0001
	[*t*-stat]	[-2.86]	[-1.28]	[-3.00]	[-1.70]	[1.94]	[0.02]
Oib(-1,Gamble)	Estimate	0.0064	-0.0960***	-0.0038	-0.0299	-0.0094	0.0235**
	[*t*-stat]	[0.32]	[-3.31]	[-0.11]	[-1.44]	[-0.44]	[2.54]
Oib(-1,Other)	Estimate	-0.0084***	-0.0082***	-0.0058***	-0.0007***	0.0010***	0.0013***
	[*t*-stat]	[-27.58]	[-24.59]	[-21.58]	[-7.67]	[13.13]	[21.18]
Adj.R2		10.44%	10.28%	9.99%	9.54%	9.42%	9.46%
Interquartile return							
Oib(-1,Persistence)		-0.1179%	-0.0964%	-0.0598%	-0.0131%	0.0281%	0.0759%
Oib(-1,Liquidity)		-0.0287%	-0.0347%	-0.0257%	0.0099%	0.0095%	0.0171%
Oib(-1,Overconf)		-0.0162%	-0.0321%	-0.0379%	-0.0383%	0.0233%	-0.0008%
Oib(-1,Gamble)		-0.0401%	-0.0347%	-0.0342%	-0.0355%	0.0448%	0.0492%
Oib(-1,Other)		-0.1782%	-0.1495%	-0.1009%	-0.0231%	0.0492%	0.0835%

Table V. A Closer Look at the Relation between Investor Order Flows and Public News

This table reports estimation results on the relation of different investor groups order flow and public news in the form of earnings announcements. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trades in the previous month. Panel A reports whether different investor groups' trading activity can predict earnings surprises. For each quarter, we estimate Fama-MacBeth regressions as specified in equation (8) to measure stock returns around the earnings announcement for firm *i* on day *d*. The dependent variable, earnings surprise, is proxied by the cumulative abnormal return from day *d*-1 to day *d*, *CAR[-1,0]*. As independent variables, we use order imbalance measures from day *d*-2, *Oib(-2)*, to avoid overlapping with the CAR calculation. Other control variables are same as those in Table III. Panel B reports whether trades from different retail groups can process contemporaneous news. For each quarter, we estimate Fama-MacBeth regressions as specified in equation (9) to measure investor trading on the earnings announcement day. The dependent variables are order imbalance measures *Oib(0)*. As independent variables, we use the cumulative abnormal return from day *d*-1 to day *d*, *CAR[-1,0]*. Other control variables are same as those in Table III. Panel C reports how earnings news days affect the return predictability of different investor group trades. We estimate Fama MacBeth regressions, as specified in equation (10). The dependent variable is the return on day *d*. The independent variables are the previous day's order imbalance *Oib(-1)*, the news dummies *Event(-1)* and the interaction terms *Oib(-1)*Event(-1)*. The *Event(-1)* dummy is equal to 1 if there is earnings announcement for that firm-day and zero otherwise. Other control variables are the same as those in Table III; those coefficients are not reported. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 4 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. Investor order flow predicting future earnings announcement news events

Dep.var		CAR[-1,0]	CAR[-1,0]	CAR[-1,0]	CAR[-1,0]	CAR[-1,0]	CAR[-1,0]
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-2)	Estimate	-0.0251***	-0.0234***	-0.0166***	-0.0003	0.0023***	0.0034***
	[*t*-stat]	[-7.33]	[-5.38]	[-4.40]	[-0.29]	[3.50]	[5.34]
Adj.R2		6.33%	5.98%	5.57%	5.19%	5.15%	5.39%

Panel B. Investor order flow regressed on contemporaneous earnings announcement news events

		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
CAR[-1,0]	Estimate	-1.9225***	-1.8291***	-1.4349***	-0.8781***	0.1583	2.7228***
	[*t*-stat]	[-17.79]	[-16.00]	[-14.34]	[-8.34]	[1.56]	[7.56]
Adj.R2		13.66%	14.60%	10.14%	1.85%	0.60%	5.92%

Panel C. Return predictive power of investor order flow interacted with earnings announcement news events

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret
		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1)	Estimate	-0.0079***	-0.0070***	-0.0038***	-0.0008*	0.0005**	0.0014***
	[*t*-stat]	[-8.27]	[-7.04]	[-3.55]	[-1.89]	[2.04]	[5.27]
Oib(-1)*Event(-1)	Estimate	-0.0080***	-0.0093***	-0.0071***	-0.0007	0.0014**	0.0020***
	[*t*-stat]	[-3.20]	[-3.15]	[-3.64]	[-0.75]	[2.26]	[3.16]
Event(-1)	Estimate	0.0011**	0.0008*	0.0006	0.0005	0.0005	0.0007*
	[*t*-stat]	[2.57]	［1.89]	[1.59]	[1.41]	[1.50]	[1.91]
Adj.R2		7.36%	7.16%	6.82%	6.38%	6.35%	6.60%

Table VI. Retail Order Flows and Future Firm Level Liquidity

This table reports estimation results on whether trading activity by different investor groups can predict the cross section of one-day-ahead liquidity. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trading during the previous month. We estimate Fama-MacBeth (1973) regressions similar to those in equation (2) and equation (3). The dependent variable is the next day effective spread in Panel A and next 12-week effective spread in Panel B. The relative effective spread (*RES*) is the (proportional) distance between the trade price P_{ik} in stock *i* at trade *k* and the prevailing quote midpoint M_{ik}, as specified in equation (11). As independent variables, we use the previous day order imbalance *Oib(-1)*, the previous month log market cap (*Size*), earnings to price ratio (*EP*) and monthly turnover (*Turnover*) . To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 5 lags. ***, ** and * indicate significance at the1%, 5% and 10% level.

Panel A. Predict next-day effective spread

Dep.var		RES	RES	RES	RES	RES	RES
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1)	Estimate	-0.0099***	-0.0054***	-0.0057***	-0.0019***	0.0019***	0.0010**
	[*t*-stat]	[-8.68]	[-4.92]	[-6.49]	[-4.70]	[6.18]	[2.35]
Size	Estimate	-0.0115***	-0.0115***	-0.0115***	-0.0115***	-0.0115***	-0.0114***
	[*t*-stat]	[-29.65]	[-29.80]	[-29.64]	[-29.66]	[-29.72]	[-29.65]
EP	Estimate	-0.6144***	-0.6159***	-0.6167***	-0.6154***	-0.6136***	-0.6127***
	[*t*-stat]	[-43.94]	[-44.03]	[-44.04]	[-43.99]	[-43.95]	[-44.12]
Turnover	Estimate	-0.0366***	-0.0368***	-0.0369***	-0.0369***	-0.0369***	-0.0367***
	[*t*-stat]	[-59.70]	[-59.90]	[-59.91]	[-59.95]	[-59.89]	[-60.15]
Intercept	Estimate	0.4342***	0.4347***	0.4341***	0.4349***	0.4347***	0.4321***
	[*t*-stat]	[42.19]	[42.47]	[42.36]	[42.36]	[42.42]	[42.44]
Adj.R2		12.75%	12.57%	12.41%	12.31%	12.34%	12.51%
Interquartile		0.2222	0.1827	0.1678	0.2868	0.4536	0.6740
Interquartile spread		-0.2200%	-0.0987%	-0.0956%	-0.0545%	0.0862%	0.0674%

Panel B. Predict longer horizon effective spread, coefficient *b*1 in equation (3)

Week	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
w=1	-0.0084***	-0.0038**	-0.0045**	-0.0017***	0.0015***	0.0008
w=2	-0.0062***	-0.0015	-0.0025**	-0.0014***	0.0012***	0.0005
w=3	-0.0051***	-0.0002	-0.0013	-0.0011**	0.0012***	0.0004
w=4	-0.0044**	0.0007	-0.0007	-0.0009*	0.0011***	0.0002
w=5	-0.0035*	0.0015	-0.0001	-0.0008*	0.0011***	0.0001
w=6	-0.0034*	0.0016	0.0002	-0.0007	0.0010***	0.0000
w=7	-0.0028	0.0019	0.0006	-0.0007	0.0010***	-0.0001
w=8	-0.0024	0.0024	0.0008	-0.0004	0.0009**	-0.0002
w=9	-0.0024	0.0023	0.0007	-0.0006	0.0009**	-0.0002
w=10	-0.0023	0.0026	0.0011	-0.0005	0.0008**	0.0000
w=11	-0.0023	0.0028	0.0016	-0.0005	0.0009**	0.0000
w=12	-0.0017	0.0032	0.0017	-0.0005	0.0008**	-0.0002

Table VII. Further Discussion and Robustness

This table reports robustness results. Panel A and Panel B report return prediction across different age and gender investor groups. The sample period covers January 2019 to March 2019. The sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trades during the previous month. Since age and gender are applicable only for retail investors, we only include retail investors. Panel A reports the summary statistics of trading volume across different age and gender groups. Panel B reports return predictions across different gender and age groups. Panel C shows the results of predicting market return using aggregate order imbalances by different investor groups, as specified in equation (12). Panel D and Panel E report return predictions by adding more data filters. The sample period covers January 2016 to June 2019, and the sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 trading days in the previous month. Panel D reports return predictions by including all filters in Liu, Stambaugh and Yuan (2019). Panel E report return predictions by excluding leveraged trading, which consists of margin buys, short sales and collateral trading. Panel F reports the Liu, Stambaugh and Yuan (2019) three-factor adjusted alphas for long-short portfolios formed on order imbalances, with holding periods from 1 day to 60 days. Panel G reports how news from CFNDS affects the return predictability of different investor group trades, as specified in equation (10). ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. Summary Statistics of gender and age groups

Gender	Trading Volume (% of total for that gender)			
Age	<35	35-45	45-55	>55
Male	10%	19%	24%	14%
Female	5%	9%	11%	9%

Panel B. Cross-sectional return predictions, by different gender and age groups

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret	Ret	Ret
Gender		Male	Male	Male	Male	Female	Female	Female	Female
Age		<35	35-45	45-55	>55	<35	35-45	45-55	>55
Oib(-1)	Estimate	-0.0015***	-0.0032***	-0.0054***	-0.0026***	0.0007**	-0.0004	-0.0021**	0.0007*
	[*t*-stat]	[-3.31]	[-4.82]	[-5.11]	[-3.20]	[2.16]	[-0.80]	[-2.57]	[1.74]
	Interquartile return	-0.04%	-0.07%	-0.12%	-0.08%	0.03%	-0.01%	-0.06%	0.02%

Panel C. Predicting market return using aggregate order imbalances by different investor groups

k days ahead	AoibRT1	AoibRT2	AoibRT3	AoibRT4	AoibRT5	AoibINST
k=1	-0.0096	-0.0195*	-0.0448***	-0.0354**	0.0126	0.0132***
k=2	-0.0097	-0.0145	-0.0238	-0.0138	-0.0019	0.0046
k=3	-0.0103	-0.0163	-0.0244	-0.0182	-0.0016	0.0111**
k=4	0.0025	0.0066	0.0177	0.0139	0.0025	-0.0046
k=5	0.013*	0.0183*	0.0331**	0.0272*	-0.0021	-0.0117**

Panel D. Cross-sectional return predictions, including all filters from Liu Stambaugh and Yuan (2019)

	Dep.var	Ret	Ret	Ret	Ret	Ret	Ret
	Oib.var	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibInst
Oib(-1)	Estimate	-0.0075***	-0.0071***	-0.0052***	-0.0011***	0.0004***	0.0017***
	[*t*-stat]	[-24.16]	[-20.81]	[-16.40]	[-7.91]	[4.49]	[17.88]
	Interquartile return	-0.17%	-0.14%	-0.09%	-0.03%	0.02%	0.11%

Panel E. Cross-sectional return predictions, excluding leveraged trading

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibInst
Oib(-1)	Estimate	-0.0092***	-0.0088***	-0.0056***	-0.0003***	0.0008***	0.0016***
	[*t*-stat]	[-24.62]	[-21.97]	[-17.20]	[-3.43]	[11.25]	[20.21]
	Interquartile return	-0.20%	-0.16%	-0.11%	-0.01%	0.05%	0.11%

Panel F. Risk adjusted alphas for long-short portfolios formed on order imbalances over different holding periods

Holding Period (days)	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
1	-0.0042***	-0.0036***	-0.0027***	-0.0007***	0.0017***	0.0025***
5	-0.0089***	-0.0068***	-0.0038***	-0.0004	0.0034***	0.0056***
10	-0.0115***	-0.0091***	-0.0048***	-0.0001	0.0046***	0.0075***
20	-0.0130***	-0.0104***	-0.0052***	0.0003	0.0056***	0.0098***
30	-0.0148***	-0.011***	-0.0056***	0.0005	0.0064***	0.0103***
40	-0.0183***	-0.0136***	-0.0069***	0.0001	0.0063***	0.0128***
50	-0.0187***	-0.0136***	-0.0063***	0.0005	0.0065***	0.0133***
60	-0.0183***	-0.0139***	-0.0072***	-0.0010	0.0057***	0.0137***

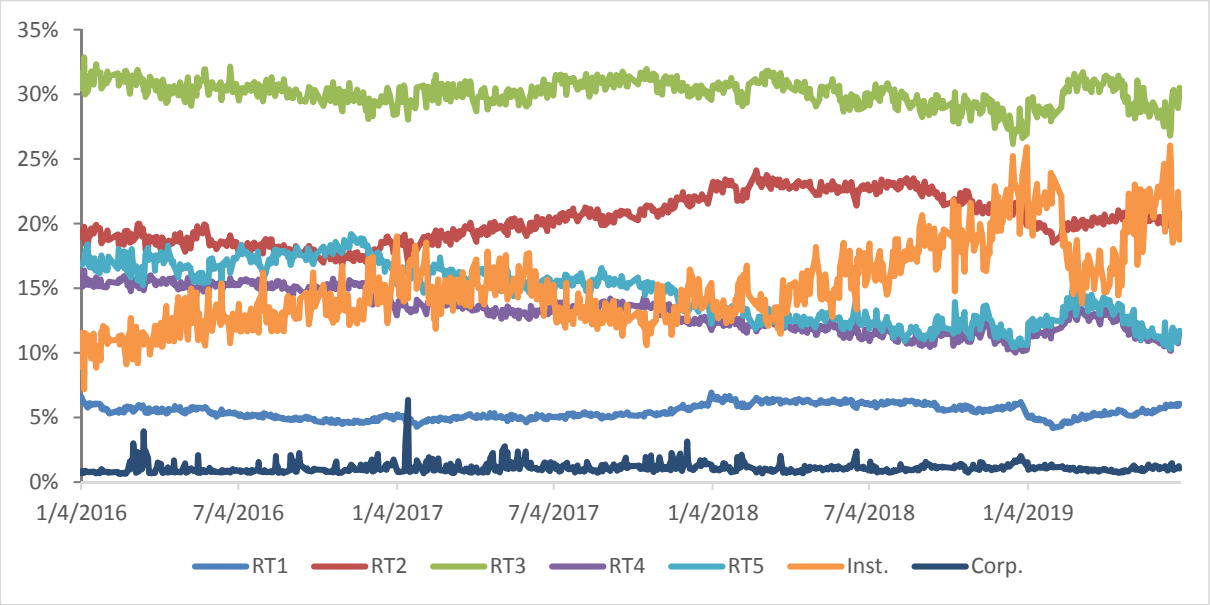
Panel G. Return predictive power of investor order flows interacted with CNFD news events

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1)	Estimate	-0.0075***	-0.0070***	-0.0047***	-0.0005***	0.0008***	0.0014***
	[*t*-stat]	[-23.61]	[-21.16]	[-16.21]	[-3.81]	[9.82]	[17.27]
Oib(-1)*Event(-1)	Estimate	-0.0045***	-0.0053***	-0.0048***	-0.0013***	0.0011***	0.0007***
	[*t*-stat]	[-11.02]	[-11.92]	[-10.79]	[-6.28]	[6.45]	[6.32]
Event(-1)	Estimate	0.0002**	0.0001*	0.0001	0.0001	0.0000	0.0001
	[*t*-stat]	[2.24]	[1.83]	[1.17]	[0.72]	[0.37]	[1.20]
Adj.R2		9.01%	8.86%	8.58%	8.19%	8.21%	8.34%

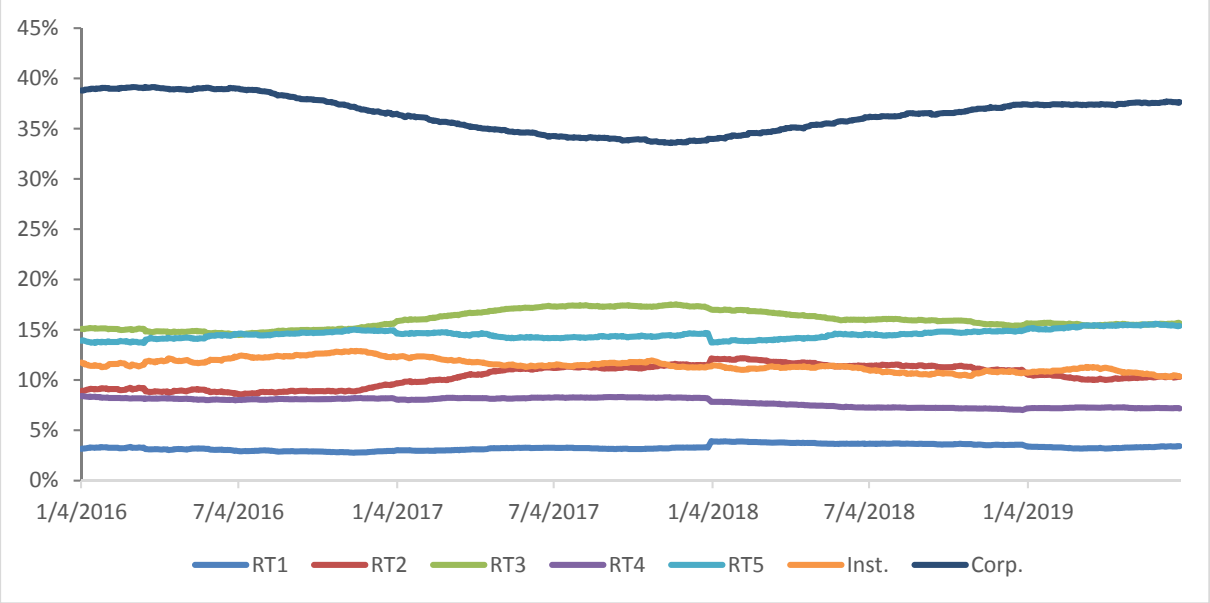
Figure I. Different Investor Type Order Flows between Jan 2016 and Jun 2019

There figures report the time series plot of the cross sectional mean for different types of investor trading activity from January 2016 to June 2019. Our sample firms are A-share stocks listed on the Shanghai Stock Exchange. In Panel A, we present the volume percentage by each type of investor. In Panel B, we show the shares held by each type of investor.

Panel A. Share volume (%), Cross Sectional Mean for Different Investor Types

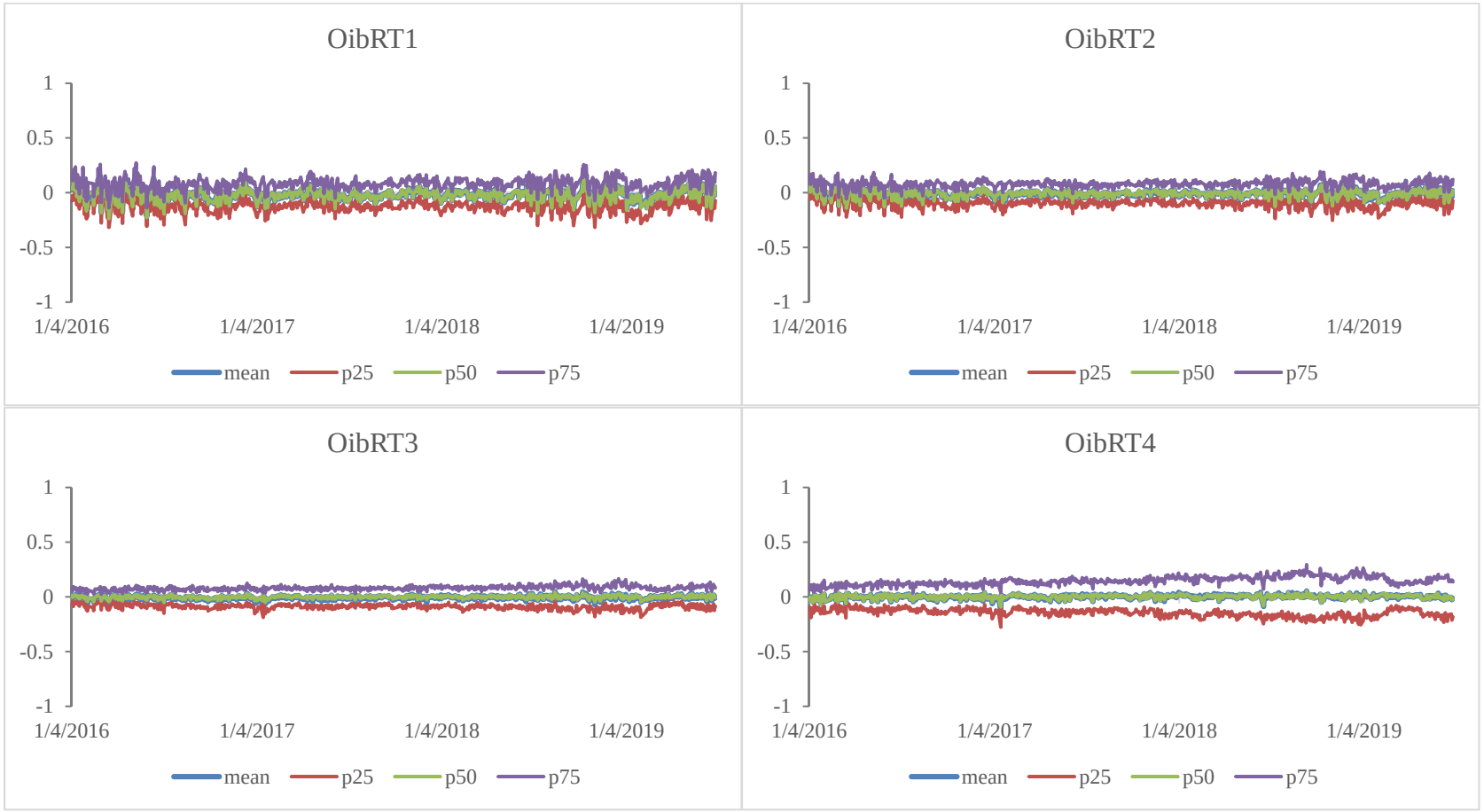


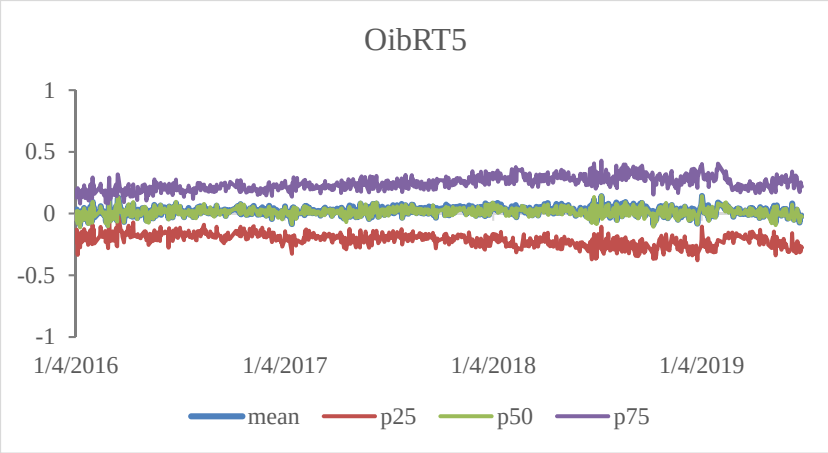
Panel B. Shares Held (%), Cross Sectional Mean for Different Investor Types

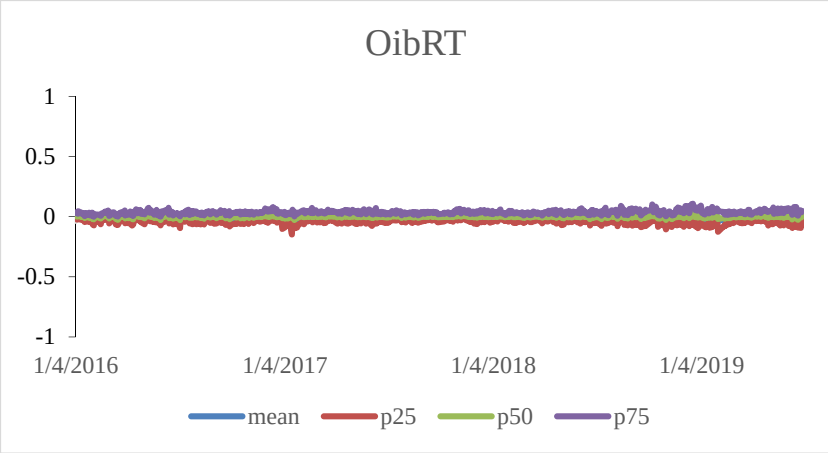


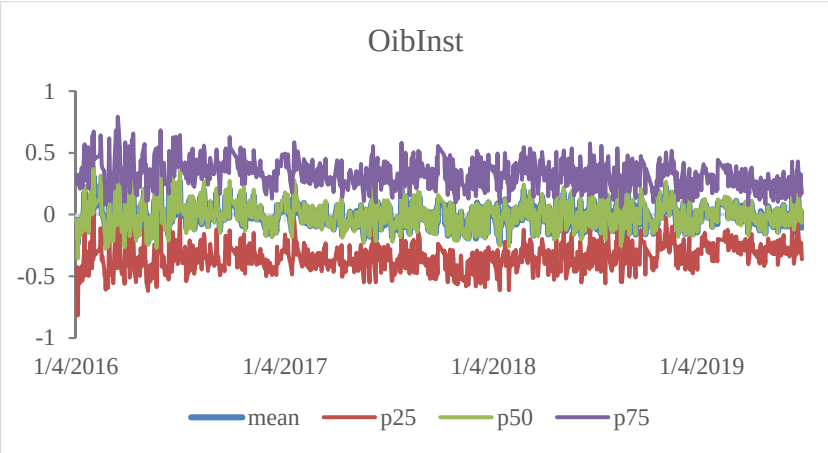
Appendix Figure I. Time Series of Different Types of Investor Order Imbalance

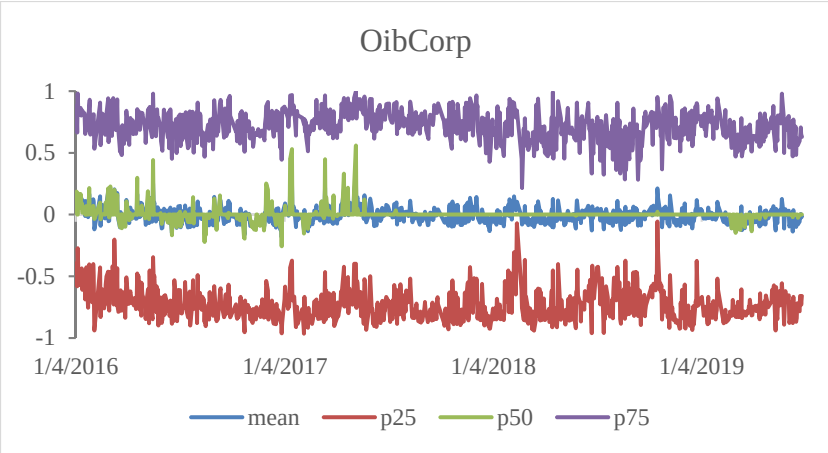
These figures reports time series of different types of investor trading activity. Our sample period covers January 2016 to June 2019, and our firms are A-share stocks listed on the Shanghai Stock Exchange. We present the cross-sectional mean, median, 25th percentiles and 75th percentiles of scaled daily order imbalances by each investor group each day. Order imbalance is computed as the buy share volume minus sell share volume divided by buy share volume plus sell share volume for each investor group, specified in Equation (1).




OibRT5
1
0.5
0
-0.5
-1
1/4/2016
1/4/2017
1/4/2018
1/4/2019
mean
p25
p50
p75


OibRT
1
0.5
0
-0.5
-1
1/4/2016
1/4/2017
1/4/2018
1/4/2019
mean
p25
p50
p75


OibInst
1
0.5
0
-0.5
-1
1/4/2016
1/4/2017
1/4/2018
1/4/2019
mean
p25
p50
p75


OibCorp
1
0.5
0
-0.5
-1
1/4/2016
1/4/2017
1/4/2018
1/4/2019
mean
p25
p50
p75