An Anatomy of Arbitrageurs: Evidence from Open-End Structured Funds*

Jennifer (Jie) LI

INSEAD

November 17, 2017

Abstract

This paper exploits a unique account-level dataset of structured funds to study how arbitrageurs trade during bubble periods (i.e., when large positive swings of mispricing occur in structured funds). I find that arbitrageurs can both ride bubbles during the bubble-formation periods and make arbitrage trades during the bubble-bursting periods. In particular, arbitrageurs ride bubbles more aggressively when local unsophisticated investors start to trade in the direction of fueling bubbles and quit this strategy when mispricing becomes excessive. Identification tests based on the social contagion effect among unsophisticated investors support a causal interpretation. Moreover, arbitrageurs who can ride bubbles make more trading profits than those who only conduct arbitrage trades. These results suggest that arbitrageurs do not always trade in the direction of eliminating mispricing and that local information may play a pivotal role in shaping their trading motivations.

Keywords: Structured Funds; Mispricing; Bubbles; Arbitrageurs; Unsophisticated Investors; Social Contagion

[*] Correspondence: Jennifer (Jie) Li, INSEAD, Department of Finance, 1 Ayer Rajah Avenue, Singapore, 138676. E-mail: jennifer.li@insead.edu. I would like to express my gratitude to my committee members Massimo Massa (co-chair), Hong Zhang (co-chair), Pierre Hillion and Neil Pearson for their guidance and support. I am also grateful to Simona Abis, Daniel Bens, Adrian Buss, Bernard Dumas, Federico Gavazzoni, Sergei Glebkin, Pekka Hietala, Peter Joos, John Kuong, Kevin Kaiser, James O'Donovan, Jöel Peress, Timothy Van Zandt, Theo Vermaelen, Bart Yueshen Zhou as well as seminar participants at the Paris Financial Management Conference 2016, the Wharton-INSEAD PhD Consortium 2017, INSEAD brownbags and PBC School of Finance, Tsinghua University for their helpful comments and suggestions. I am responsible for any errors.

1. Introduction

Arbitrage, the hypothetical trading strategy of arbitrageurs, is at the core of modern finance. The classical view argues that such a trading strategy allows rational arbitrageurs to help eliminate mispricing (i.e., price deviations from the fundamental value), which provides the foundation for market efficiency (Fama (1965), Friedman (1953)). The incentives and real trading strategies of arbitrageurs, however, may be more subtle than conducting pure arbitrage. There are concerns, for instance, that rational arbitrageurs may not be able, if not unwilling, to quickly trade against mispricing because of various market frictions or limits to arbitrage (De Long et al. (1990a), Pontiff (1996), Shleifer and Vishny (1997), Kyle and Xiong (2001), Gromb and Vayanos (2010), to name a few). [1] Worse, arbitrageurs may even ride bubbles in expectation of positive feedback trading (De Long et al. (1990b)) or due to synchronization risk among arbitrageurs (Abreu and Brunnermeier (2002)). Clearly, these alternative incentives and trading strategies may give rise to drastically different implications in terms of price efficiency. It is therefore important to empirically understand how arbitrageurs trade against mispricing in the real financial market.

The aim of this paper is to achieve this goal by exploring a unique account-level trading dataset of structured funds. Both mispricing and the trading behavior of arbitrageurs can be clearly identified in this dataset. Specifically, investors can invest in a base asset, a fund labeled M, which is almost identical to a standard open-end mutual fund, except that investors can choose to convert M into two structured assets at a pre-determined conversion ratio – a fixed-income asset (called asset A) and a levered-equity asset (called asset B; details will be provided in later sections). Both A and B are traded on the stock exchange and can be converted back to M by investors. Their net asset values (NAVs) are calculated from the NAV of M and announced by the fund family on a daily basis, but their trading prices often deviate from their NAVs, creating the classical scenario of mispricing in which trading price of an asset deviates from its fundamental value. Interestingly, mispricing typically concentrates on the leveraged asset B, which is consistent in spirit with Hong and Sraer (2016) and Frazzini and Pedersen (2014) in that investors with short-sale constraints and borrowing constraints may chase high-beta assets and drive up their prices.

One unique feature of the structured fund is that large swings of mispricing offer arbitrage opportunities to sophisticated investors. If the conversion-ratio-weighted average trading prices of A and B are higher than that of their NAVs (i.e., bubbles), investors can purchase M from

[1] See Gromb and Vayanos (2010) for a review of the relevant theoretical literature on limited arbitrage.

the mutual fund family at the NAV, convert it into A and B, and sell these shares on the stock market at higher prices. This scenario of structured fund mispricing is the focus of the current paper, as it allows us to examine not only arbitrage trading but also the alternative strategy of riding bubbles. [2] Moreover, because arbitrage trading strategies can be clearly identified, we can also identify arbitrageurs (unsophisticated investors) as investors who have conducted (never conducted) arbitrage trading. [3]

With the above features, the proprietary dataset I explore in this paper contains complete account-level trading information for 47,749 accounts investing in two structured funds offered by a large mutual fund family in China. The data cover investors from 31 provinces and more than 300 cities in mainland China and span the period from November 2011 to December 2015. Compared to existing empirical papers that infer arbitrageur trading from the short selling or holding data of hedge funds (Brunnermeier and Nagel (2004), Hong et al. (2012), Griffin et al. (2011)), the extensive coverage and geographic richness of these data can help not only to depict the complete trading behavior of arbitrageurs but also design tests to shed light on the type of information that arbitrageurs use, which is crucial to understand how arbitrageurs make trading decisions and exploit less-sophisticated investors.

I use this dataset to conduct three steps of empirical analysis. In the first step, I examine how arbitrageurs trade in three inter-connected dimensions: first, do arbitrageurs explore alternative trading strategies such as riding bubbles, in addition to arbitrage trading, to exploit unsophisticated investors? Second, what type of information do they use in making trading decisions? Finally, do arbitrageurs make profits from their arbitrage and alternative trading? These questions are closely related because the information filtration of arbitrageurs largely affects, if not partially determines, the optimal trading strategy of arbitrageurs. Abreu and Brunnermeier (2003), for instance, demonstrate that when the mass of arbitrageurs exceeds a critical threshold in the market in the presence of a bubble, the optimal strategy for a particular arbitrageur is to conduct arbitrage. However, if her information set suggests that the critical mass has not yet been reached, her optimal

[2] In the reverse (negative mispricing) case, investors can purchase A and B from the stock market, convert A and B back into M, and then liquidate M at its NAV. One can, of course, view this reverse case as a negative bubble. However, our analysis focuses on positive mispricing because riding a negative bubble, which requires arbitragers to short sell A and B shares, is prohibited.

[3] The arbitrage trading is called the "pair conversion mechanism" (PCM). The PCM is clearly stated in the prospectus of each fund: "On any day SZSE is open for trading prior to the termination of the structured fund, A and B can be combined to form one M which may be redeemed at the NAV of M; M may be split into A and B which can be sold in SZSE. Investors of separate components may have to pay customary brokerage fees and commissions to the SZSE".

strategy becomes riding the bubbles.

To understand the trading behavior of arbitrageurs, therefore, the critical issue is how to measure the information set of arbitrageurs. I explore a key intuition established in the current literature that local information may play a pivotal role in shaping the trading behavior of investors (Gehrig (1993), Brennan and Cao (1997), Coval and Moskowitz (1999), Coval and Moskowitz (2001), Hong et al. (2004),Coval (2003), Kaustia and Knüpfer (2012)). The geographic richness of the data allows me to construct a proxy for local information that could be particularly important to arbitrageurs: the flow of inexperienced and unsophisticated investors in the local city who start buying assets during bubble periods. The more unsophisticated investors with no experience there are to buy assets, especially the levered asset B that is vulnerable to mispricing, the more likely it is that unsophisticated investors are fueling a bubble or augmenting an existing one (Mackay (1869), Brennan (2004), Greenwood and Nagel (2009), Griffin et al. (2011), Xiong and Yu (2011) Shiller (2015), Gong et al. (2016)). Since the flow of inexperienced and unsophisticated investors in some cities may lead those in other cities, this local information presents a private signal to an arbitrageur on the likelihood of bubble formation.

However, arbitrageurs can also directly observe mispricing, which is a public signal in the setting of structured funds. The role of this public signal differs from private information. If mispricing is high, for instance, this does not necessarily mean that unsophisticated investors will further chase and augment the bubble – or the reverse (a better inference can be drawn from the flow of inexperienced and unsophisticated investors). Rather, substantial mispricing is likely to induce more arbitrageurs, including those without superior local information, to exploit this opportunity. Hence, the optimal trading strategy of an arbitrageur is likely to be determined jointly by her private signal and the public signal. For instance, if the private signal indicates that a bubble is forming whereas the public signal suggests that the mass of arbitrageurs is not yet large (i.e., mispricing is still small), riding the bubble could be the optimal strategy. By contrast, if the public signal suggests that the mass of arbitrageurs is likely to be large, her optimal strategy becomes conducting arbitrage (and ceasing her bubble-riding strategy, if any).

To test the above intuition, I examine how arbitrage flows and bubble-riding flows (both are aggregated from arbitrageurs at the city level) respond to the proxy for local information (i.e., flow of inexperienced and unsophisticated investors in the same city), the proxy for public information (i.e., existing mispricing), and their interactions. Arbitrage flows and bubble-riding flows are defined as the number of shares "partitioned" from M for arbitrage trades in a city and the number of arbitrageurs in a city conducting bubble-riding strategies, respectively. My baseline analysis adopts a panel specification with city and time fixed effects, and it focuses on a sample with positive mispricing. I find that local arbitrage flows are higher when there are more unsophisticated investors participating in the local city after controlling for the total number of unsophisticated investors participating in other cities. A one-standard-deviation increase in the number of new unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of arbitrage flows of 0.72 implies a change of 8.75% in the arbitrage flows. The interaction term between new unsophisticated investors and mispricing is significantly positively related to arbitrage flows, indicating that arbitrageurs exploit both local information and public information when making trading decisions.

However, arbitrageurs could also strategically ride bubbles during bubble-formation periods. In the baseline regression, I find that arbitrageurs ride bubbles when local unsophisticated investors start to trade in the direction of fueling the bubbles and quit this strategy when mispricing becomes substantial. The economic magnitude of this effect is that a one-standard-deviation increase in the number of new unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of the number of arbitrageurs riding bubbles of 0.04 implies a change of 17.4% in the number of arbitrageurs who buy B when mispricing is positive (the standard deviation is scaled by the bubble-riding flows). The interaction term between new unsophisticated investors and mispricing is significantly negatively related to arbitrageur bubble-riding behavior, indicating that arbitrageurs ride bubbles less when mispricing becomes high and they use both local and public information in their trading decisions. In addition to regression analysis, I graphically illustrate how arbitrageurs trade during 15 ex post bubble periods. Figure 7 clearly demonstrates that arbitrageurs buy more B before the mispricing peak, which drives up mispricing, and they sell more after the mispricing peak, which helps eliminate mispricing.

Thus far, I have shown that arbitrageurs exploit both arbitrage and bubble-riding strategies and that the adoption of these strategies is jointly determined by local and public information. The remaining issue is that, if riding bubbles is part of the optimal strategy, arbitrageurs should reap higher returns by doing so. To explore this question, I further separate arbitrageurs into two groups: those who conduct both bubble-riding and arbitrage strategies and those who only focus on the arbitrage strategy. [4] Arbitrageurs who can ride bubbles earn 7.39% realized returns per

[4] Arbitrageurs are labeled as riding bubbles if they ever buy B during the ex post bubble periods. Arbitrageurs are labeled as not riding bubbles if they never buy B during the ex post bubble periods. In addition, arbitrageurs obtain

trade, while those who only conduct arbitrage trades earn 3.14%.

To further explore whether there is a causal link between unsophisticated investor trading and arbitrageur trading, the second step of analysis involves one identification test based on social contagion effects among unsophisticated investors. A potential concern is that local arbitrageur and unsophisticated investor trading behavior may be spuriously correlated due to unobserved city characteristics or reverse causality. I follow the identification method of Kaustia and Knüpfer (2012) and use the social contagion effect among unsophisticated investors as a relatively exogenous shock to test how arbitrageurs respond to unsophisticated investor trading in a two-stage regression model. Kaustia and Knüpfer (2012) show that recent stock market returns that local peers experience affect an individual's stock market entry decision, and they rule out alternative explanations such as market returns, media coverage and short sales constraints. Their identification strategy of using large geographical variation in stock returns while controlling for zip code and time fixed effects and using clustering with province-level standard errors helps to rule out common unobservables and argues for a causal relationship between peer performance and stock market entry in the same neighborhood. As in the account-level trading data, zip code information is available for both arbitrageurs and unsophisticated investors.

In the first stage, I aggregate the number of unsophisticated investors who experience positive returns or negative returns and new unsophisticated investors at the city level and find that the new entry of unsophisticated investors only responds to the number of unsophisticated investors who experience positive returns but not to the number who experience negative returns. This pattern is consistent with selective communication: people are more likely to talk about favorable experiences. The second stage explores the influence of predicted new entry of unsophisticated investors on arbitrageur flows and bubble-riding flows. The predicted new entry is the fitted part of the number of unsophisticated investors who experience positive returns in the first stage and estimated over the entire sample period. By regressing the predicted new entry of unsophisticated investors on proxies for arbitrageur trading behavior, I find compelling evidence that the results are consistent with the baseline regression.

The final step of the empirical analysis examines whether the results are robust to an additional test and robustness checks. An additional test of migrant arbitrageurs supports the finding that arbitrageurs respond to local information. Migrant arbitrageurs are defined as those arbitrageurs whose

higher realized returns than unsophisticated investors, which validates the classification of arbitrageurs.

city of residence differs from their hometown. I use the National Identity Numbers of investors to trace their city of birth (i.e., their city of origin) and apply the previous tests to investors whose trading locations differ from their city of origin (i.e., migrants). If arbitrageurs indeed respond to local information, their trading behavior should be more sensitive to local unsophisticated investor participation than hometown unsophisticated investor participation. Consistent with the baseline regression, the newly entering unsophisticated investors in the local city have more explanatory power for arbitrageur trading behavior than the average new entry of unsophisticated investors in hometowns.

The main empirical results are robust to four sets of robustness checks. In the first set of robustness checks, I use alternative definitions of arbitragers and apply the baseline analysis. [5] The second set of tests use an alternative definition of location and apply the baseline analysis. [6] The third set of tests address concerns related to the weekend effect. As numerous empirical papers indicate that the distribution of stock returns varies by the day of the week, it is reasonable to suspect that the weekend effect may affect the mispricing of the structured funds and thus affect the baseline regression conducted at a daily frequency (Lakonishok and Levi (1982), French (1980)). To resolve the weekend effect concern, key variables are aggregated at a weekly frequency and applied to the baseline regression. The fourth set of tests concern the subsamples of mispricing. The main tests are based on positive mispricing when there is an ex ante potential bubble. The results from the subsample tests show that the main results are robust to various mispricing samples.

To the best of my knowledge, this paper is the first to clearly identify mispricing, arbitrageurs and unsophisticated investors in a clean setting. By studying the trading behavior of arbitrageurs using unique account-level data, this paper complements existing empirical papers that study arbitrageur trading behavior using quarterly institutional holding data (Baker and Savaşoglu (2002), Brunnermeier and Nagel (2004), Griffin et al. (2011)). As noted by Puckett and Yan (2011), changes in quarterly holdings data do not capture intra-quarter transactions when funds purchase and sell the same stock. By exploiting account-level trading data, this paper provides empirical evidence consistent with the synchronization risk model of Abreu and Brunnermeier (2003) and positive feedback trading model of De Long et al. (1990b).

This paper contributes to the empirical literature on whether rational arbitrageurs improve price

[5] Two new definitions of arbitrageurs are used: arbitrageurs who engage in arbitrage trading at least twice and ex ante identified arbitrageurs.

[6] The alternative definition of location is when I define the local area as the province.

efficiency or stabilize financial markets. As arbitrageurs could use short selling to drive over-valued assets back to fundamentals, a series of papers use short selling data to show that they can improve price efficiency and stabilize the market (Akbas et al. (2013), Hwang and Liu (2014), Wu and Zhang (2015)). As hedge funds emerged as institutionalized arbitrageurs and data on their stock holdings became available, a number of papers have used hedge fund holding data to show that hedge funds are informed and reduce mispricing in the market (Agarwal et al. (2009), Akbas et al. (2015), Cao et al. (2014), Kokkonen and Suominen (2015), Sias et al. (2015)). Another group of studies documents that arbitrageurs do not stabilize financial markets. Hong et al. (2012) use short selling interest data to show that arbitrageurs amplify economic shocks by short covering. Hedge funds ride tech bubbles instead of exerting a corrective force on stock prices during technology bubbles (Brunnermeier and Nagel (2004), Griffin et al. (2011)). The empirical finding that rational arbitrageurs may not always trade in the direction of eliminating mispricing has important normative implications in the financial market.

This paper also contributes to the literature related to the role of local information in shaping the trading behavior of investors. A number of existing studies have shown that geography plays a very important role in the economy (Lerner (1995), Audretsch and Feldman (1996), Audretsch and Stephan (1996), Jaffe et al. (1993)) and local investors have preferences for local assets because of informational advantages (Gehrig (1993), Brennan and Cao (1997), Coval and Moskowitz (1999), Coval and Moskowitz (2001),Coval (2003)). The results in this paper extend the current literature on the local information advantage by showing that arbitrageurs may also exploit local information when making their trading decisions.

Finally, this paper speaks to the theories of bubbles and misvaluation. The structured fund setting provides investors with relatively low-cost arbitrage opportunities. My contribution is to show that positive feedback trading and the social contagion effect are two factors that could cause mispricing even when arbitrage is relatively easy. Positive feedback trading is broadly consistent with disagreement models such as Scheinkman and Xiong (2003) when investors' belief updates are driven by the realized returns on their own past trading. The social contagion effect among unsophisticated investors are consistent with extrapolative models in Barberis et al. (2016), in which new investors enter after observing positive past returns.

The paper is organized as follows. Section 2 provides an introduction to China's structured fund market and similar financial products. Section 3 describes the data and constructs the variables. I report the empirical evidence in section 4 and robustness checks in section 5. I conclude the paper in section 6.

2. Institutional Setting

2.1. Open-end Structured Fund Setting

Structured funds originated in the United Kingdom in the 1960s and appeared in the United States in the 1980s. Since the 1990s, structured funds have become one of the main types of mutual funds in U.S. capital markets.

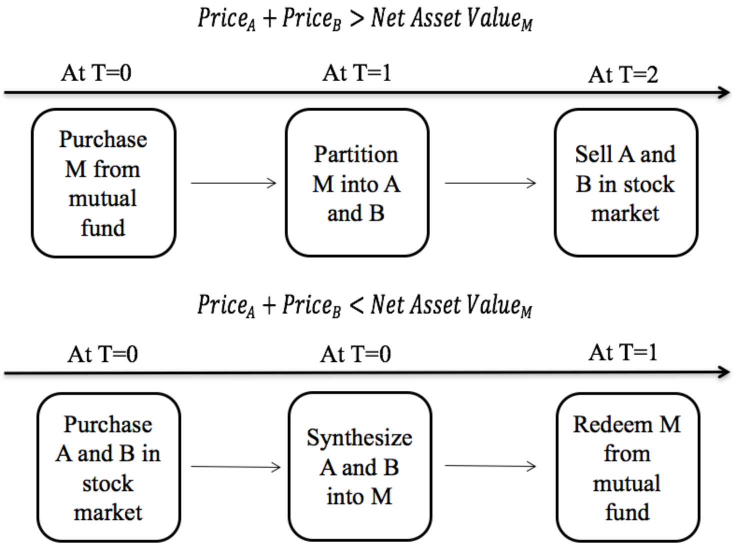
With the development of the country's capital market, new financial products are being introduced into the Chinese financial market. In 2007, UBS SDIC Fund Management Company established the first closed-end structured fund in China. In 2009, Changsheng Fund Management Company launched the first open-end structured fund. In 2009, UBS SDIC issued the first fund that had an arbitrage trading mechanism that provided investors with low-cost arbitrage opportunities between M, A and B. As of the end of 2015, of 220 dual-purpose funds in China, 150 open-end funds had implemented the arbitrage trading strategy, including 139 equity funds and 11 bond funds.

One challenge in identifying mispricing is that we do not really know the fundamental value of an asset. The structured fund setting overcomes the challenge by providing both trading prices and fundamental values of the same asset. A unit of the structured fund is labeled M, and if M is traded on the stock exchange, it will be split into A and B according to a fixed conversion ratio. Investors in A are entitled to a pre-specified minimum annual interest rate. Investors in B are entitled to any residual value after A investors are paid their interest. A can be viewed as a fixed-income security, while B can be viewed as a levered security. Their NAVs are calculated from the NAV of M and announced by the fund family on a daily basis, but their trading prices often deviate from their NAVs, creating the classical scenario of mispricing in which the trading price of an asset deviates from its fundamental value. [7]

Figure 1 describes how the arbitrage strategy works. If the conversion-ratio-weighted average trading prices of A and B ($P_A + P_B$) are higher than that of their NAVs (NAV_M), investors can purchase M from the mutual fund family at the NAV, convert it into A and B, and sell these shares

[7] A receives a pre-specified annual interest rate, and its NAV is determined by $NAV_t^A = 1 + \frac{R*t}{365}$, where R is the pre-specified annual interest rate, and t is the number of trading days passed during the year. B receives any residual value after A. The NAV of B can be backed out of the equation $NAV_t^A + NAV_t^B = NAV_t^M$, where $NAV_t^A$is the NAV of A at time t and $NAV_t^B$is the NAV of B at time t.

Figure 1: Arbitrage Trading Mechanism



$Price_A + Price_B > Net\ Asset\ Value_M$
At T=0
At T=1
At T=2
Purchase M from mutual fund
Partition M into A and B
Sell A and B in stock market
$Price_A + Price_B < Net\ Asset\ Value_M$
At T=0
At T=0
At T=1
Purchase A and B in stock market
Synthesize A and B into M
Redeem M from mutual fund

on the stock market at higher prices. If the conversion-ratio-weighted average trading prices of A and B ($P_A + P_B$) are smaller than those of their NAVs (NAV_M), investors can purchase A and B from the stock market, synthesize them into M, and sell M back to the mutual fund family at NAV. [8]

< **Insert Figure 1 here** >

In the Shenzhen Stock Exchange (SZSE), trading is subject to the "T+1" rule, which requires investors to hold their shares for one day before selling. Investors thus bear the risks in the market because they have to wait one day for the discount arbitrage deal to be settled and two days for premium arbitrage to be settled.

Despite the rapid development of the Chinese financial market, the central government has been very cautious about introducing new financial products, especially financial derivatives. The China Securities Regulatory Commission (CSRC) issued "structured fund review guidelines" to regulate structured funds. Note also that investors are legally prohibited from short selling mutual fund shares in China.[9]

[8] Based on the NAV of M and the trading prices of A and B, the structured fund mispricing for fund i at time t is defined as $Mispricing_{it} = \frac{P^A_{it}+P^B_{it}-NAV^M_{it}}{NAV^M_{it}}$, where $P^A_{it} + P^B_{it}$ are the synthetic trading prices of A and B in the stock market based on the prices of A and B.

[9] The CSRC started to allow shorting of a selected set of stocks only in 2010 but allows none for mutual funds.

2.2. Similar Financial Products

2.2.1. Primes and Scores

Primes and scores are financial products that are most similar to the structured funds in this paper. Primes and scores split the cash flow of a stock into dividend and capital gain component respectively. Jarrow and O'Hara (1989) investigate the mispricing of primes and scores and establish a nonparametric statistical model for the prices of primes and scores. Barber (1994) uses the misperceptions of noise traders to explain the time-series variation of the premium pricing of primes and scores.

2.2.2. Dual-purpose Fund

Dual-purpose funds are diversified closed-end funds that are capitalized with two classes of shares and have a fixed termination date. The two classes are capital shares which pay no dividends and are redeemable at their NAVs at the maturity of the fund and income shares which have the rights to any dividends or income that the fund may earn and are to be redeemed at a pre-determined price at the maturity of the fund.

In 1967 and 1968, seven dual-purpose funds were successfully underwritten. Most literature about dual-purpose funds was from 1970s to 1980s. Litzenberger and Sosin (1977) derive the guidelines for the structure and management of dual-purpose funds. In the paper, they talk about the economic incentives, expenses and performance, short sale restrictions, patterns of discounts and premiums of the dual-purpose funds. They conclude that it is better to structure dual-purpose funds as open-end funds than closed-end funds because the closed-end fund discounts would disappear if funds were organized as open-end trusts.

Ingersoll (1976) formulates a dual-purpose fund pricing function based on the option studies of Black-Scholes and Merton. He finds that as other closed-end funds do, capital shares are sold at a price below their NAVs.

3. Data and Variable Construction

This section provides descriptions of the data used in this paper, identifies two groups of investors and defines key variables used in the empirical analysis.

3.1. Market Data

The structured fund data were collected from GTA database. [10] The database includes 150 open-end structured funds in Chinese market. In the 150 open-end structured funds, 139 funds are equity funds and 11 are bond funds. The sample period is from 2009 to April 2016 because the first open-end structured fund was launched in 2009.

The market data includes the following information:

1) Basic information about M, A and B: for M, the database covers the fund ID, fund name, fund ID for the corresponding A and B, establishment date, original leverage ratio for A and B and fund style; for A, the database includes fund ID, fund name, the corresponding M, listing date, code of stock exchange and minimum annual interest rate; for B, the database includes fund ID, fund name, the corresponding M, listing date, code of stock exchange.

2) Net asset value about M , A and B for each day: net asset values and accumulated net asset values are calculated each day for M , A and B respectively.

3) Market trading price for A and B for each day: the database contains fund ID, fund name, trading date and the trading prices for A and B such as opening price, closing price, high price, low price, trading volume and trading amount measured in RMB.

4) Premiums or discounts of A and B for each day: for A, the database includes fund ID, fund name, trading date, premiums or discounts of trading prices relative to net asset values, minimum annual interest rate and internal rate of return; for B, the database includes fund ID, fund name, trading date, premiums or discounts of trading price relative to net asset value, leverage ratio based on net asset value and leverage ratio based on trading price.

Information about the stock market index and underlying stocks are from the China Stock Market and Accounting Research (CSMAR) database provided by the Wharton Research Data Services (WRDS).

3.2. Investor Trading Data

The unique account-level trading data come from a confidential mutual fund family in China. The mutual fund family is located in Shanghai. It ranks in the top 30 in China, both in terms of the number of mutual funds offered and in terms of total net assets (TNA) under management, with investors from all 31 provinces and around 300 cities in Mainland China. The fund family allows

[10] The company supplies the Chinese financial stock market data to the WRDS database, which is the same as China Stock Market and Accounting Research (CSMAR) and is commonly used in the finance academic community.

investors to open investment accounts either directly online or indirectly through brokerage firms or bank branches. Each investor is allowed to open only one account through these channels, which is registered under his or her National Identity Number (at any given time, each citizen in China has a unique National Identity Number). After opening the account, investors can buy shares of any fund offered by this family and/or redeem their existing shares. The investment rules on the operations side of mutual fund investment are identical to those in the U.S.

For each account, the database allows us to retrieve information about a) the investor profile, b) trading history, and c) dividend distributions. The investor profile contains an investors personal information, including his or her unique National Identity Number, date of birth, gender, contemporaneous postcode and distribution channel. For each transaction, the trading file provides the name of the mutual fund involved, the total number of shares purchased or redeemed, the total value of the purchase or redemption, the total transaction fees related to these transactions, and the total number of shares after the transaction. Finally, the dividend file provides information about the type and total amount of dividends distributed to each investor based on his/her shareholdings in the specific mutual fund.

The trades file includes the records of all trades for 47,749 accounts including 47,303 individual investors and 446 institutional investors for two structured funds from November 2011 to December 2015. The file documents all investor transactions in history including the "businflag" of the transactions on M: purchases, redemptions, exchanges between funds, Automatic Investment Plan and so forth. Investor arbitrage activities can be clearly identified by the "businflag". The file also includes the quantity traded for M, A and B on each day and different types of transaction costs in the particular trading activities.

Table 1 presents descriptive statistics of investor trading data in the two structured mutual funds. The summary statistics are based on a sample of mutual fund investors who trade in a top-30 mutual fund family in the period from September 2011 to December 2015. Panel A reports the total number of investors trading in M, A and B. Panel B presents the distribution of individual investor age in the full sample. Panel C shows the number of transactions in premium arbitrage and discount arbitrage, in trading A, B and M respectively. From panel A in this table, we can see that the distribution of investors trading different types of shares varies across equity and bond funds. For the equity fund, investors trade B more aggressively than A, which is reasonable because investors prefer high-beta assets to lever up when they face borrowing constraints in Hong and Sraer (2016) and Frazzini and Pedersen (2014). The distribution of the number of trades in panel

C shows consistent statistics that B is traded significantly more frequent than A for the equity fund.

< Insert Table 1 here >

Table 2 presents the summary statistics of mispricing in the structured mutual funds. Panel A reports numbers of observations, means, medians and standard deviations, along with the minimum, maximum, 25%, and 75% quintile values of the absolute value of mispricing for 150 open-end structured funds in the entire Chinese market. Panel B presents the distribution of the absolute value of mispricing for the two funds with investor trading information. Panel C reports the distribution of the absolute value of mispricing during bubble periods. The bubble period in panel C is defined based on the mispricing of M shares. I rank the time series mispricing of each fund and define the top-15 premium peaks as bubble peaks, then define the 7 days before each bubble peak as the "bubble-formation period" and 7 days after each bubble peak as "bubble-bursting period".

< Insert Table 2 here >

3.3. Identify Two Groups of Investors

By exploiting the arbitrage trading mechanism (i.e.,"PCM"), I identify two groups of investors: arbitrageurs and unsophisticated investors. Arbitrageurs are defined as those investors who engage in arbitrage trading at least once when arbitrage opportunities arise at any point in their trading history with the fund. Unsophisticated investors are those investors who never conduct arbitrage trades even when they are presented with profitable arbitrage opportunities because they are not sophisticated enough to exploit those opportunities.

To demonstrate the validity of the classification, table 3 tabulates the characteristics of the two groups of investors: arbitrageurs and unsophisticated investors for the equity fund with investor trading data. Panel A reports numbers of observations, means and standard deviations, along with the 5%, 25%, 75% and 95% quintile values of differences in age, wealth (RMB value) and trading experience for the two groups of investors. The average age of arbitrageurs is 45, which is higher than the 43 observed for unsophisticated investors. Wealth represents the average RMB value of each purchase across all trading accounts. From the average level of wealth, we can predict that arbitrageurs are, on average, wealthier than unsophisticated investors in the sample. Experience is constructed based on the number of days between the first trading date and the last trading date. The average number of trading days is 261 days for arbitrageurs and 161 days for unsophisticated investors, meaning that arbitrageurs are more experienced than unsophisticated investors. Panel

B presents the trading characteristics of the two groups of investors in the equity fund. Trading size is the daily average RMB value of total purchases and sales across all trading accounts for each group of investors. Trading volume is the daily average number of shares bought and sold across all trading accounts for each group of investors. From the mean of trading volume, we can conclude that arbitrageurs trade at higher volume on average. Profitability measures the average realized return per transaction across all trading assets and trading accounts for each group of investors. The profitability measure provides evidence that, overall, arbitrageurs gain positive returns and unsophisticated investors gain negative returns. Panel C reports the average realized return per transaction in trading other mutual funds in the same mutual fund family for the two groups of investors. The profitability of arbitrageurs in trading other funds is also higher than that of unsophisticated investors. This is a out-of-sample test to show that the classification of investors is indeed valid.

< Insert Table 3 here >

Table 4 provides realized return analysis for the two groups of investors: arbitrageurs and unsophisticated investors. The statistical results confirm the validity of my classification: arbitrageurs on average gain significantly higher returns than unsophisticated investors both in the full sample period and in bubble periods. The transactions can be separated into normal trades and arbitrage trades. Arbitrageurs conduct normal trades, such as trading A and B independently, and they also engage in arbitrage trading. However, unsophisticated investors only trade M, A and B independently in the stock market. First, I calculate the purchasing cost of each selling transaction using the FIFO (first in first out) convention based on the entire trading history of the investors and calculate realized return based on their historical purchasing cost. Then, realized returns are aggregated across all investors in each group for date t using the equal-weighted and value-weighted methods. Finally, I conduct a t-test for the time series of average realized returns. Panel A shows the equal-weighted average realized return after transaction costs, and panel B shows the value-weighted average return after transaction costs. Panel A1 presents the equal-weighted average realized return during bubble periods, and panel B1 presents the equal-weighted average realized return during bubble periods. As in table 2, I rank the time series mispricing of the fund and define the top-15 premiums as bubble peaks, then define the 7 days before each bubble peak as the "bubble-formation period" and the 7 days after each bubble peak as the "bubble-bursting period".

< Insert Table 4 here >

3.4. Variable Construction

This section describes the construction of variables for the empirical regressions. First, structured fund mispricing and its two components are defined according to the literature Barber (1994) and Lee et al. (1991) as follows:

$$Mispricing_{it} = \frac{(P^A_{it} + P^B_{it}) - NAV^M_{it}}{NAV^M_{it}} \tag{1}$$

$$Mispricing^A_{it} = \frac{P^A_{it} - NAV^A_{it}}{NAV^A_{it}} \tag{2}$$

$$Mispricing^B_{it} = \frac{P^B_{it} - NAV^B_{it}}{NAV^B_{it}} \tag{3}$$

where $Mispricing_{it}$ is the structured fund mispricing for fund i at time t, and $Mispricing^A_{it}$, $Mispricing^B_{it}$ are two components of the structured fund mispricing. P^A_{it} is the closing price of A for fund i on day t and P^B_{it} is the closing price of B for fund i on day t. NAV^M_{it} is the NAV of M for fund i on day t. NAV^A_{it} is the NAV of A for fund i on day t, which is calculated as $NAV^A_{it} = 1 + \frac{R*t}{365}$, where R is the pre-determined interest rate of A. NAV^B_{it} is the net asset value of B for fund i on day t, which is calculated as $NAV^B_{it} = NAV^M_{it} - NAV^A_{it}$.

I then turn to describe the measurement of the behavior of arbitrageurs and unsophisticated investors. To better link investor behavior to local information, investor trading activities with the equity structured fund are aggregated at the city level based on the residential address of each investor. It is intuitive from the features of A and B that B is the primary driver of structured fund mispricing, which is also consistent with the view that investors prefer high-beta assets to lever up when they face borrowing constraints in Hong and Sraer (2016) and Frazzini and Pedersen (2014). I focus primarily on the behavior of two groups of investors in trading B. The premium arbitrage flow for each city is the logarithm of aggregated number of shares "partitioned" from M for premium arbitrage across all arbitrageurs in the same city, which can be directly observed from the trading data. Arbitrageurs who ride bubbles are those arbitrageurs who buy B, thus helping drive up the price of B in the market when mispricing is positive (when there is ex ante potential bubble formation). A new unsophisticated investor is defined as the first time an unsophisticated investor enters the market. The new entry of unsophisticated investors for a city is the logarithm of the aggregate number of new unsophisticated investors to buy B.

4. Empirical Evidence

4.1. Existence of Mispricing of the Structured Funds

4.1.1. Value-weighted Average Index of Mispricing

Following Lee et al. (1991), I construct the value-weighted average index of mispricing for the structured funds in the market at daily and monthly levels:

$$VWP_t = \sum_{i=1}^{n_t} W_t * Mispricing_{it} \tag{4}$$

where $W_i = \frac{NAV_{it}}{\sum_{i=1}^{n_t} NAV_{it}}$, and NAV_{it} is equal to the NAV of the base asset M for fund i at the end of period t, $Mispricing_{it} = \frac{(P_{it}^A + P_{it}^B) - NAV_{it}}{NAV_{it}} * 100$, $P_{it}^A + P_{it}^B$ is the synthetic market price of M in the stock market based on the price of A and B, and n_t is the number of funds with available data at the end of period t. Following the method used by Barber (1994), the market price of fund i $(P_{it}^A + P_{it}^B)$ equals the synthetic closing prices of A and B because M is not traded on the stock exchange by itself. [11]

Using the market data of 150 open-end structured funds, I calculate the value-weighted average mispricing for each trading day. The first graph in figure 2 shows the value-weighted average mispricing across 150 open-end structured funds from January 2010 to April 2016, with the red line plotting the trend of the Shanghai Shenzhen CSI 300 Index. [12] The cross-sectional daily average premium is 1.6%, including 0.6% to 1.2% in transaction costs, and the daily average discount is -1.1%, including 0.5% to 1% in transaction costs.

< Insert Figure 2 here >

Figure 3 plots the historical daily mispricing for the two funds with investor trading data. Significant mispricing has already been discovered based on the market data of 150 funds. This demonstrates that the two funds that I use exhibit the same pattern of mispricing as the other structured funds in the market.

< Insert Figure 3 here >

[11] The same method is used by Barber (1994) because the units (the combined Prime and Score) are thinly traded in the AMEX.

[12] The CSI 300 is a capitalization-weighted stock market index designed to replicate the performance of 300 stocks traded on the Shanghai and Shenzhen stock exchanges and is one of the most commonly used stock market indexes in China.

4.1.2. Define Bubble Periods

Griffin et al. (2011) note that the "tech bubble" and the more recent credit and real estate bubbles pose challenges for efficient market theories and are not well understood. Traditionally, a bubble is usually defined as an "upward price movement over an extended range that then implodes," as in Kindleberger (1978), or as "price levels [having been] too high to be explained by reasonable expectations of future cash flows" in Ofek and Richardson (2002). Previous bubbles such as the "tech bubble" or "tulip bubble" exhibit persistently growing prices for a certain period of time, but we do not really know the fundamental value in these cases. In Pástor and Veronesi (2006), they actually argue that the existence of a Nasdaq bubble in the late 1990s should not be taken for granted because uncertainty over average profitability, which increases the fundamental value of a firm, was unusually high in the late 1990s. Using the structured fund setting, we can directly identify deviations from the fundamental value and thus define a "bubble" according to mispricing.

As mentioned previously, a structured fund allows investors to conduct arbitrage between the mutual fund M and A and B in the stock market. Structured fund mispricing is defined as the difference between the synthetic market price of A and B and the NAV of M. The two components of structured fund mispricing (i.e., mispricing of A and mispricing of B) are defined according to the assets' features, namely, that A is similar to a fixed-income security and B is similar to a levered security that claims the residual value of the underlying portfolio.

The dynamics of bubbles are shown in figure 4 in the event time window for the equity fund with investor trading data. I define 15 bubbles based on the top-15 instances of structured fund mispricing. The event time $t = 0$ is the mispricing peak. The "bubble period" is 7 days before and after the mispricing peak. These 15 bubbles are independent in the sense that the bubble periods do not overlap one another. For each event time t, I take the average of the 15 mispricings at each event date t. This figure shows that the mispricing of a structured fund is almost 20% relative to its fundamental value. The area in between the red dotted lines is when profitable arbitrage opportunities exist.[13]

< Insert Figure 4 here >

[13] Profitable arbitrage opportunities exist when the mispricing is higher than the estimated transaction costs in PCM. The average transaction cost for the two funds can be estimated from the trading data: approximately 0.873% (0.485%) for the premium (discount) arbitrage of the equity fund and 0.638% (0.07%) for the premium (discount) arbitrage of the bond fund.

4.2. How Do Arbitrageurs Trade: Baseline Analysis

In this section, I examine how arbitrageurs trade in three inter-connected dimensions: first, do arbitrageurs explore arbitrage trading or other alternative trading strategies such as riding bubbles? Second, what type of information do they use when making trading decisions? Finally, do arbitrageurs make profits from their arbitrage and alternative trading?

4.2.1. Arbitrageurs Trade against Unsophisticated Investors

The following multivariate specification is to further verify the relationship between the participation of unsophisticated investors and arbitrageur behavior and examines the results of the following daily panel regressions with time and city fixed effects:

$$Arbitrage\,flow_{c,t} = \alpha_0 + \alpha_1 * New\,entry^U_{c,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^U_{c,t-1} + \alpha_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_6 * M_{c,t-1} + \epsilon_{c,t} \quad (5)$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in city c on day t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Other\,entry^U_{c,t-1}$ is the logarithm of the total number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1. $Mispricing_{t-1}$ is the positive mispricing of a structured fund for fund f on day t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles). The vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs who buy B at t-1 and mispricing for fund f on day t. I include time and city fixed effects in all specifications. Time fixed effects remove the time trend, and city fixed effects control for the time-invariant unobservable heterogeneity across different cities. Please refer to the Internet Appendix for variable definitions.

The regression results are reported in table 5, panel A. Models 1-6 examine the relationship between the premium arbitrage flow and the new entry of unsophisticated investors. In the six specifications, time and city fixed effects are included to remove the potential influence of time invariant city-level characteristics, while in Model 6, I further include control variables such as arbitrage flows on day t-1 because arbitrageurs may influence one another in time series trading. All models exhibit a significant positive relationship between arbitrage flows and the new entry of unsophisticated investors to buy B – adding control variables such as contemporaneous mispricing neither affects this relationship nor changes its level of significance. To interpret the economic magnitude of this test in the most comprehensive Model 6, a one-standard-deviation increase in the number of new entries of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of premium arbitrage flows of 0.72, implies an increase of 8.75% in premium arbitrage flows. [14]

In addition to controlling city and time fixed effects, it is natural to control for the number of new unsophisticated investors entering the stock market to buy B in other cities to argue for a local impact of unsophisticated investors on arbitrage behavior. As expected, the new entry of unsophisticated investors in other cities does not have explanatory power in the local city. One would expect structured fund mispricing to drive the premium arbitrage flow, and thus, I control for structured fund mispricing in the regression. Here, the variable $Mispricing_{t-1}$ represents the days when mispricing is positive (when there are ex ante potential bubbles). Surprisingly, mispricing does not have explanatory power for the premium arbitrage flow at the city level. In Appendix B, a vector autoregressive model indicates that in time series, the aggregate arbitrage flow is significantly positively related to the absolute value of mispricing on day t-1. The lack of significance may be explained by the relative importance of local information and public information. Models 4-6 introduce interaction terms between the new entry of unsophisticated investors and market mispricing. The coefficient is significant and positive (0.055, with a t-statistic of 3.09), suggesting that mispricing enhances the sensitivity of arbitrage flows to the new entry of unsophisticated investors. Model 6 controls for the lagged premium arbitrage flow because there may also be a contagion effect among arbitrageurs, and the statistics verify this prediction. As I control for contemporaneous mispricing, it is negatively correlated with the arbitrage flow, which is reasonable because the mispricing is calculated using the daily closing price. Arbitrageurs would not know the mispricing at the end of the day when they trade during the day, and thus, the arbitrage flow during the day will drive the premium down at the end of the day.

[14] For instance, in Model 6, the regression coefficient of the logarithm of premium arbitrage flows on the logarithm of the number of new entries of unsophisticated investors is 0.525. A one-standard-deviation increase in the number of new entries of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of premium arbitrage flows of 0.72, implies a change of 0.525*0.12/0.72=8.75% in the premium arbitrage flows.

< Insert Table 5 here >

4.2.2. Arbitrageurs Ride Bubbles

An unconventional position predicts that arbitrageurs may actually drive bubbles. Arbitrageurs may contribute to price movements based on the expectation that positive-feedback traders will purchase the securities later at even higher prices (De Long et al. (1990b)) or that riding the bubble can maximize profits (Abreu and Brunnermeier (2002)). Based on these theoretical contributions, Abreu and Brunnermeier (2002) and Griffin et al. (2011) show that arbitrageurs rode the tech bubble using hedge fund holding data. However, quarterly institutional holding data may miss intermediate trading information for these arbitrageurs.

I examine the relationship between the participation of unsophisticated investors and arbitrageurs riding bubbles in the following daily panel regressions with time and city fixed effects:

$$Riding\,bubble_{c,t} = \alpha_0 + \alpha_1 * New\,entry^U_{c,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^U_{c,t-1} + \alpha_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_6 * M_{c,t-1} + \epsilon_{c,t} \quad (6)$$

where the subscripts c and t refer to city and day, respectively. $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Other\,entry^U_{c,t-1}$ is the logarithm of the total number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1. $Mispricing_{t-1}$ is the structured fund mispricing for fund f on day t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles), and the vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs who buy B at t-1 and the mispricing for fund f on day t. Please refer to the Internet Appendix for variable definitions.

Panel B in table 5 shows the results. Models 1-6 examine the relationship between arbitrageurs riding bubbles and the new entry of unsophisticated investors. In the six specifications, time and city fixed effects are included to remove the potential influence of time-invariant city-level characteristics, while in Model 6, I further include control variables such as the number of arbitrageurs riding a bubble on day t-1 because arbitrageurs may influence one another in time series trading. All models exhibit a significant positive relationship between arbitrageurs driving up bubbles and the new entry of unsophisticated investors to buy B – adding control variables such as contemporaneous mispricing neither affects this relationship nor changes its level of significance. To interpret the economic magnitude of this test in the most comprehensive Model 6, a one-standard-deviation increase in number of new entries of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of the total number of arbitrageurs riding bubbles of 0.04, implies an increase of 17.4% in the number of arbitrageurs riding bubbles. [15]

In addition to controlling for city and time fixed effects, it is natural to control for the number of new unsophisticated investors entering the stock market to buy B in other cities when arguing for a local impact of unsophisticated investors on arbitrage behavior. As expected, the new entry of unsophisticated investors in other cities does not have explanatory power in the local city. Models 4-6 introduce the interaction terms between the new entry of unsophisticated investors and market mispricing. The coefficient is significant and negative (0.005, with a t-statistic of negative 3.77), suggesting that arbitrageurs will decide to quit the bubble-riding strategy when mispricing becomes excessive, namely the probability of a bubble bursting is high. Model 6 controls for the lagged proxy for arbitrageurs riding bubbles because there may also be a contagion effect among arbitrageurs, and the statistics verify this prediction. In this regression, market mispricing does not appear to be a significant predictor of arbitrageurs riding bubbles, indicating that arbitrageurs may focus more on local information than public information to decide whether to ride bubbles.

< Insert Table 5 here >

In addition to the regression analysis, I also document different trading behaviors of the two groups of investors during ex post defined bubble periods (see figure 5, figure 6 and figure 7).

Figure 5 presents the number of new investors entering the market to buy B during bubble periods. During a bubble period, new arbitrageurs continue entering the market to buy B, and there is a large drop at the price peak. Then, the number of new arbitrageurs is much lower than before the price peak. As the figure demonstrates, unsophisticated investors continue to enter after the

[15] For instance, in Model 6, the regression coefficient of the logarithm of the total number of arbitrageurs riding bubbles on the logarithm of the number of new entry of unsophisticated investors is 0.058. A one-standard-deviation increase in the number of new entries of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of the number of arbitrageurs riding bubbles of 0.04, implies a change of 0.058*0.12/0.04=17.4% in the number of arbitrageurs riding bubbles.

price peak and stop entering when the price has already dropped.

< Insert Figure 5 here >

Figure 6 shows total number of shares of B bought by arbitrageurs and unsophisticated investors. During the bubble period, arbitrageurs buy heavily when the bubble starts to grow and decrease their buying when the bubble approaches the peak, and they stop buying after the price peak. However, unsophisticated investors increase the number of shares they buy, driving up the price of B, and buy even more after the price peak.

< Insert Figure 6 here >

Figure 7 documents the net purchases of B for both arbitrageurs and unsophisticated investors during bubble periods. As arbitrageurs may buy and sell B in a short period, in this figure, I examine the net purchases of B in the market. The net flow of B during a bubble period is the total number of shares bought minus the total number of shares sold during the bubble period. It is clear that for arbitrageurs, before the bubble peaks, net purchases are positive, while after the bubble peaks, the net purchases are negative and there is a large drop around the price peak. This suggests that arbitrageurs are in general riding the bubble before the price peak and arbitrage away mispricing after the price peak. However, regarding the net purchasing of unsophisticated investors, they are generally buying regardless of whether it is before or after the bubble peaks.

< Insert Figure 7 here >

4.3. Two Identification Tests

4.3.1. Two-stage regression

This section investigates the general relationship between arbitrageur behavior and unsophisticated investor participation in two steps. The first step exploits social contagion among unsophisticated investors, and the second step investigates how the predicted new entry of unsophisticated investors affects arbitrageur behavior.

Various papers have studied the influence of peer actions in the stock market. For instance, Hong et al. (2004) and Brown et al. (2008) provide evidence that individuals are more likely to participate in the stock market when their geographically proximate peers participate. Kaustia and Knüpfer (2012) explains the entry decisions of individual investors by the stock market outcomes of peers. They attempt to distinguish between two plausible channels through which stock market outcomes of peers could influence entry decisions. In the first channel, investors use peer outcomes to update beliefs about long-term fundamentals, such as the equity premium. In the second channel, peer outcomes are not directly observable, and investors rely on "word-of-mouth" verbal accounts. Verbal accounts are likely be biased toward reporting positive outcomes, as investors are unlikely to benefit from discussing their negative outcomes with their peers. The authors find that the lagged average return affects entry decisions when it is positive but is unrelated to entry decisions when it is negative. This is consistent with the second channel: selective reporting and peer returns affecting entry via word-of-mouth communication.

Kaustia and Knüpfer (2012) show that the recent stock returns that local peers experience affect an individual's stock market entry decision, and they rule out alternative explanations such as market returns, media coverage, and short sales constraints. I follow the identification method of Kaustia and Knüpfer (2012) and use the social contagion effect among unsophisticated investors as a relatively exogenous shock to test how arbitrageurs respond to unsophisticated investor trading. Their identification strategy of controlling for zip code and time fixed effects and clustering standard errors at the province level helps to rule out reverse causality and common unobservables and leads them to argue for a causal relationship between peer performance and stock market entry. As my account-level trading data are similar to those used in Kaustia and Knüpfer (2012), zip code information is available for both unsophisticated investors and arbitrageurs.

This two-stage least-squares specification estimates the effect of participation by unsophisticated investors on arbitrageur behavior in the period from 2011 to 2015. In the first stage, I predict the estimated number of new entries of unsophisticated investors using the following panel regression:

$$Newentry^{U}_{c,t} = \beta_0 + \beta_1 Posreturn\,num^{U}_{c,t-1} + \beta_2 Posreturn\,mean^{U}_{c,t-1} + \beta_3 Negreturn\,num^{U}_{c,t} + \beta_4 Negreturn\,mean^{U}_{c,t-1} + \beta_5 CEFD_{t-1} + \epsilon_{c,t} \tag{7}$$

where subscripts c and t refer to city and week, respectively. $Newentry^{U}_{c,t}$ is the number of new unsophisticated investors entering the stock market to buy B in city c and week t-1; New unsophisticated investor is defined as the first time such an investor enters the market to trade; $Posreturn\,num^{U}_{c,t-1}$ is the number of unsophisticated investors who experience positive returns since their purchase in city c and week t-1; $Posreturn\,mean^{U}_{c,t-1}$ is the average positive returns of those investors in city c and week t-1; $Negreturn\, num^{U}_{c,t}$ is the number of unsophisticated investors who experience negative returns since their purchase in city c and week t-1; and $Negreturn\, mean^{U}_{c,t-1}$ is the average negative returns of those investors in city c and week t-1. $CEFD_{t-1}$ is the closed-end fund discount used as a proxy to control for investor sentiment in the market.

In the second stage, I regress the arbitrage flow and bubble-riding proxies on the predicted number of new entries of unsophisticated investors in the following regression:

$$\begin{aligned} Arbitrage\, flow_{c,t}/Riding\, bubble_{c,t} = \alpha_0 + \alpha_1 * Predicted\, new\, entry^{U}_{c,t-1} + \\ \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\, entry^{U}_{c,t-1} + \alpha_4 * New\, entry^{U}_{c,t-1} * Mispricing_{t-1} + \\ \alpha_5 * Other\, entry^{U}_{c,t-1} * Mispricing_{t-1} + \epsilon_{c,t} \end{aligned} \tag{8}$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\, flow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in city c on day t; $Riding\, bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $Predicted\, new\, entry^{U}_{c,t-1}$ is the logarithm of the estimated new entries of unsophisticated investors, which is the fitted part of $\beta_1 Posreturn\, num^{U}_{c,t-1}$ in the first-stage regression; and $Other\, entry^{U}_{c,t-1}$ is the logarithm of the total number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1. $Mispricing_{t-1}$ is the structured fund mispricing for fund f on day t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles).

< Insert Table 6 here >

In addition, I further explore why arbitrageurs would like to ride bubbles. Table 7 provides one of the basic reasons that arbitrageurs would like to ride bubbles before the price peak. In this analysis, arbitrageurs are separated into two groups: arbitrageurs riding bubbles and arbitrageurs not riding bubbles. Arbitrageurs are labeled as riding bubbles if they ever buy B during the bubble periods when the structured fund mispricing already exceeds the transaction costs, meaning that arbitrageurs are supposed to buy M, partition it into A and B and sell them in the market. Arbitrageurs are labeled as not riding bubbles if they never buy B during the bubble period when the market premium already exceeds the transaction costs, meaning that profitable arbitrage opportunities exist. As shown in table 7, arbitrageurs who are riding bubbles gain higher realized returns on B before the price peak and arbitrage away mispricing after the price peak.

< **Insert Table 7 here** >

An additional test explores who among the arbitrageurs is riding bubbles by examining the results of the following daily panel regressions with time and city fixed effects:

$$\begin{aligned} Riding\,bubble_{c,t} = \lambda_0 + \lambda_1 * New\,entry^{U}_{c,t-1} + \lambda_2 * Mispricing_{t-1} + \lambda_3 * Non\,inventory_{c,t-1} + \\ \lambda_4 * New\,entry^{U}_{c,t-1} * Mispricing_{t-1} + \lambda_5 * Non\,inventory_{c,t-1} * Mispricing_{t-1} + \\ \lambda_6 * New\,entry^{U}_{c,t-1} * Mispricing_{t-1} * Non\,inventory_{c,t-1} + \epsilon_{c,t} \end{aligned} \tag{9}$$

where the subscripts c and t refer to city and day, respectively. $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B at time t; $New\,entry^{U}_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c at time t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Non\,inventory_{c,t-1}$ is the logarithm of the number of arbitrageurs who do not have inventory of B in their account in city c at time t-1. $Mispricing_{t-1}$ is the structured fund mispricing for fund f at t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles). Please refer to the Internet Appendix for variable definitions.

< **Insert Table 8 here** >

4.3.2. Migrant Arbitrageur

In this section, I use the identification numbers of arbitrageurs to trace their region of birth and apply the previous tests to migrant arbitrageurs, namely, those whose trading locations differ from their region of birth, who are contained in my sample. Since each investor can only register one account with the fund family, in most cases, the trading location is the residence of the investor. National Identity Numbers allow me to trace the region of birth for each investor.

If arbitrageurs indeed respond to local information, their trading behavior should be more sensitive to local unsophisticated investor participation than hometown unsophisticated investor participation. The following daily panel regressions with time and city fixed effects test the relationship between the behavior of migrant arbitrageurs and the participation of unsophisticated investors in their city of residence vs. their hometown:

$$\begin{aligned} Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \gamma_0 + \gamma_1 * New\,entry^{U}_{c,t-1} + \gamma_2 * Mispricing_{t-1} + \\ \gamma_3 * Hometown\,entry^{U}_{c,t-1} + \gamma_4 * New\,entry^{U}_{c,t-1} * Mispricing_{t-1} + \\ \gamma_5 * Hometown\,entry^{U}_{c,t-1} * Mispricing_{t-1} + \gamma_6 * M_{c,t-1} + \epsilon_{c,t} \end{aligned} \tag{10}$$

where the subscripts c and t refer to city and day, respectively. Migrant arbitrageurs are defined as those arbitrageurs whose hometown province and province of residence are different.[16] $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage for migrant arbitrageurs in city c on day t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of migrant arbitrageurs who buy B on day t; $New\,entry^{U}_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Hometown\,entry^{U}_{c,t-1}$ is the logarithm of the average number of new unsophisticated investors entering the stock market to buy B for the hometown cities of all migrant arbitrageurs in city c on day t-1. $Mispricing_{t-1}$ is the structured fund mispricing for fund f at t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles), and the vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs who buy B at t-1 and the mispricing for fund f on day t.

The main results of this system of equations are tabulated in table 9. Models 1-3 report the results when the dependent variable is the logarithm of the premium arbitrage flow, and Models 4-6 present the results when the dependent variable is the logarithm of the number of arbitrageurs riding bubbles. In all specifications, there is a significant positive relationship between arbitrageur behavior and the new entry of unsophisticated investors. More important to this analysis, the variable $Hometown\,entry^{U}_{c,t-1}$ has little explanatory power for local arbitrageur behavior, indicating that arbitrageurs indeed respond to local information.

< Insert Table 9 here >

[16] Province is a larger administrative unit than city, thus arbitrageurs who move from one province to another must also change their city.

4.4. *What Causes Mispricing?*

4.4.1. *Positive Feedback Trading*

Positive feedback trading is another important factor in Shiller's feedback loop theory of bubbles, and it means that investors are more likely to trade again if their past returns are positive. In testing positive feedback trading, an investor might use multiple buys to accumulate a position and then liquidate the position using multiple sell orders. This raises the issue of how to treat sets of transactions in which multiple buys or sells are used to accumulate or liquidate a position. I define a transaction cycle to resolve this issue. Starting from a holding of zero shares of fund f, a transaction cycle begins with a purchase of some non-zero amount. It continues through possibly multiple purchases and sales, until the investor's position returns to zero. This is a single transaction cycle.

The proportional hazards model accounts for the time that has elapsed since an investor completed the last transaction cycle. Specifically, consider an investor A who had a large positive return yesterday and another investor B who had a large positive return a few months ago but has not yet traded again. Investor A is more likely to trade on date t than investor B, who has probably left the market and is unlikely to trade on date t.

The proportional hazards model specifies that $\lambda_i, f, t(\tau)$ is the hazard function of starting a new transaction cycle for an existing investor i in fund f on day t τ trading days after the end of the investor's last transaction cycle, and it takes the following form:

$$\lambda_{i,f,t}(\tau) = \lambda(\tau) * e^{x_{i,f,t} * \beta} \tag{11}$$

where $\lambda(\tau)$ is the baseline hazard rate and $x_{i,f,t}$ is a vector of covariates that proportionally shift the baseline hazard. For investors who have previously completed one transaction cycle $x_{i,f,t} * \beta$ is given by

$$x_{i,f,t} * \beta = a1 * Returnlag1_{i,f,t} + b1 * (I(Returnlag1_{i,f,t} > 0)) + controls + u1_{i,f,t} \tag{12}$$

where $Returnlag1_{i,f,t}$ is the return on the most recent transaction cycle of investor i in B of fund i before date t. The dummy variable $I(Returnlag1_{i,f,t} > 0)$ takes value one if $Returnlag1_{i,f,t} > 0$ and zero otherwise. I also control for fund and time fixed effects.

The results of the regression model are reported in table D.18 and, in general, show that positive returns on previous transaction cycles predict a higher probability that investors open a new transaction cycle for one-cycle investors and two-cycle investors. The coefficients on $I(Returnlag1_{i,f,t} > 0)$ for one-cycle investors and two-cycle investors are large and highly significant, with p-values less than 0.0001. In the sample, 75% of the investors are one-cycle investors. These results indicate some presence of positive feedback trading among unsophisticated investors.

< Insert Table D.18 here >

4.4.2. The Impact of New Entry and Repeat Entry

In this section, I document the impact of the new entry of unsophisticated investors due to social contagion and the repeated entry of unsophisticated investors due to positive feedback trading on both mispricing and arbitrageur behavior.

First of all, this panel regression tests the impact of the new entry of unsophisticated investors due to social contagion and feedback trading on mispricing of the equity structured funds:

$$\begin{aligned} Mispricing_{f,t} = \phi_0 + \phi_1 New\,entry^U_{f,t-1} + \phi_2 Repeat\,entry^U_{f,t-1} + \phi_3 B\,volume_{f,t-1} + \\ \phi_4 A\,volume_{f,t-1} + \phi_5 A\,volatility_{f,t-1} + \phi_6 B\,volatility_{f,t-1} + \epsilon_{f,t} \end{aligned} \tag{13}$$

where $Mispricing_{f,t}$ is the structured fund mispricing calculated as the difference between the synthetic trading price of A, B and the NAV of M for fund f at time t. $New\,entry^U_{f,t-1}$ is the total number of new unsophisticated investors aggregated at fund level, which is due to social contagion effects among unsophisticated investors. $Repeat\,entry^U_{f,t-1}$ is the total number of repeated entry of unsophisticated investors aggregated at fund level, which is due to the positive feedback trading for each unsophisticated investor herself.

Regression results in table 10 show that proxies for social contagion and positive feedback trading are significantly positively related to the structured fund mispricing. When new entry and repeat entry are put together in model 5 and model 6, new entry has more explanatory power on mispricing than repeated entry, indicating that the new entry of unsophisticated investors is a better proxy than the repeated entry.

< Insert Table 10 here >

The impact of new entry and repeat entry on arbitrageur behavior is examined in the following daily panel regressions with time and city fixed effects:

$$\begin{aligned} Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \delta_0 + \delta_1 * New\,entry^U_{c,t-1} + \delta_2 * Mispricing_{t-1} + \\ \delta_3 * Repeat\,entry^U_{c,t-1} + \delta_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \\ \delta_5 * Repeat\,entry^U_{c,t-1} * Mispricing_{t-1} + \delta_6 * M_{c,t-1} + \epsilon_{c,t} \end{aligned} \tag{14}$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares partitioned from M for the premium arbitrage in city c on day t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1, which is mostly due to social contagion effect among unsophisticated investors. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Repeat\,entry^U_{c,t-1}$ is the logarithm of the number of existing unsophisticated investors entering the stock market to buy B after the first trade in B in city c on day t, which is mostly due to positive feedback trading effect. $Mispricing_{t-1}$ is the structured fund mispricing for fund f at t-1 which is calculated as the difference between synthetic trading price in the market and the NAV of M; here it represents the days when mispricing is positive (when there is ex ante potential bubbles). The vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs riding bubbles at t-1 and mispricing for fund f on day t.

The results are reported in table 11. Models 1-3 report the results when the dependent variable is the logarithm of premium arbitrage flow and models 4-6 present the results when the dependent variable is the logarithm of the number of arbitrageurs riding bubbles. In all specifications, there exists a significant positive relationship between arbitrageur trading and unsophisticated investor trading. The repeated entry of unsophisticated investors due to feedback trading exhibits some explanatory power on arbitrage flows and riding-bubble flows. To interpret the economic magnitude of both new entry and repeated entry in the most comprehensive model 3, a one-standard-deviation increase in the number of new entries of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of premium arbitrage flows of 0.72, implies an increase of 6.73% in the premium arbitrage flows. [17] As to the economic significance of repeated entry of

[17] For instance, in Model 3, the regression coefficient of the logarithm of the premium arbitrage flows on the logarithm of the number of new entries of unsophisticated investors is 0.404. A one-standard-deviation increase in the

unsophisticated investors, a one-standard-deviation increase in the number of repeated entry of unsophisticated investors, which is 0.16 in the sample, compared to the standard deviation of premium arbitrage flows of 0.72, implies an increase of 1.72% in the premium arbitrage flows. [18]

To interpret the economic magnitude of both new entry and repeated entry in the most comprehensive Model 6, a one-standard-deviation increase in the number of new entries of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of the number of arbitrageurs riding bubbles of 0.04, implies an increase of 13.2% in the number of arbitrageurs riding bubbles. [19] Regarding the economic significance of the repeated entry of unsophisticated investors, a one-standard-deviation increase in the number of repeated entry of unsophisticated investors, which is 0.16 in the sample, compared to the standard deviation of the number of arbitrageurs riding bubbles of 0.04, implies an increase of 9.2% in the number of arbitrageurs riding bubbles. [20]

< Insert Table 11 here >

5. Robustness Check

In this section, I conduct four sets of robustness checks to further validate the previous results.

In the first set of robustness checks, I use alternative definitions of arbitrageurs and apply the baseline analysis in table 5. First, arbitrageurs are defined as those investors who conduct twice arbitrage trading at least twice. In addition, ex ante identified arbitrageurs are used in the regressions. Ex ante identified arbitrageurs are those investors who conduct arbitrage trading in their first year, and I examine their trading behavior in the rest of the sample periods. The main findings

number of new entry of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of premium arbitrage flows of 0.72, implies a change of 0.404*0.12/0.72=6.73% in the premium arbitrage flows.

[18] For instance, in Model 3, the regression coefficient of logarithm of premium arbitrage flows on the logarithm of number of repeated entries of unsophisticated investors is 0.085. A one-standard-deviation increase in the number of repeated entries of unsophisticated investors, which is 0.16 in the sample, compared to the standard deviation of premium arbitrage flows of 0.79, implies a change of 0.085*0.16/0.72=1.89% in the premium arbitrage flows.

[19] For instance, in Model 6, the regression coefficient of the logarithm of number of arbitrageurs riding bubbles on the logarithm of the number of new entries of unsophisticated investors is 0.044. A one-standard-deviation increase in the number of new entries of unsophisticated investors, which is 0.12 in the sample, compared to the standard deviation of the number of arbitrageurs riding bubbles of 0.04, implies a change of 0.044*0.12/0.04=13.2% in the number of arbitrageurs riding bubbles.

[20] For instance, in Model 6, the regression coefficient of the logarithm of number of arbitrageurs riding bubbles on the logarithm of the number of repeated entries of unsophisticated investors is 0.023. A one-standard-deviation increase in the number of repeated entries of unsophisticated investors, which is 0.16 in the sample, compared to the standard deviation of the number of arbitrageurs riding bubbles of 0.04, implies a change of 0.023*0.16/0.04=9.2% in the number of arbitrageurs riding bubbles.

hold: local arbitrage flows are higher when there are more unsophisticated investors participating in the same local financial market, and arbitrageurs are strategically riding bubbles to earn higher returns and exiting when the probability of a bubble bursting increases. Taken together, this table 12 and the previous results indicate that the impact of local unsophisticated investors on arbitrageur behavior is significant regardless of the definition of arbitrageurs.

The second set of tests use an alternative definition of location and apply the baseline analysis in table 5. While the cross-section in the main test is at the city level, province is used as the new definition of location here. The data set covers 31 provinces across China, and the results are reported in table 13. Model 1 to Model 3 present the empirical results for arbitrage flows, and Model 4 to Model 6 present the results for arbitrageurs riding bubbles and are consistent with the main findings in table 5.

The third set of tests address the concern regarding the weekend effect. As numerous empirical papers indicate that the distribution of stock returns varies by the day of the week, it is reasonable to suspect that the weekend effect may affect the mispricing of the structured funds and thus affect the baseline regression that is conducted at a daily frequency (Lakonishok and Levi (1982), French (1980)). To resolve the weekend effect concern, I aggregate the key variables at a weekly frequency and apply the baseline analysis in table 5. Table 14 reports the empirical results of the panel regression with city and week fixed effects, and they are consistent with the two main findings.

The fourth set of tests concern the subsamples of structured fund mispricing. The main tests are based on positive mispricing because the arbitrage flows and bubble-riding behavior of arbitrageurs are only related to positive mispricing in my setting. In panel A of table 15, I apply the baseline test in all samples of mispricing and in the subsample when mispricing is above 0.5%. The main findings hold: local arbitrage flows are higher when there are more unsophisticated investors participating in the same local financial market, and arbitrageurs are strategically riding bubbles to earn higher returns and exiting when the probability of a bubble bursting is higher.

6. Conclusion

This paper empirically tests how arbitrageurs trade against mispricing in the real financial market in the setting of structured funds where both mispricing and the trading behavior of arbitrageurs and unsophisticated investors can be clearly identified.

Based on unique account-level trading data that contain complete trading information on 47,749 accounts and more than 300 cities, I find that arbitrageurs can both ride bubbles during bubble-formation periods and make arbitrage trades during bubble-bursting periods. In particular, arbitrageurs ride bubbles more aggressively when local unsophisticated investors start to trade in the direction of fueling bubbles and quit this strategy when mispricing becomes excessive. In addition, arbitrageurs who ride bubbles earn higher realized returns than those arbitrageurs who only conduct arbitrage trades. Furthermore, the results show that arbitrageurs exploit both public and local information when making trading decisions. Finally, an identification test based on a social contagion effect among unsophisticated investors suggests a causal relationship between unsophisticated investor trading and arbitrageur trading.

The results have important normative implications regarding the role of arbitrageurs in the financial market. This paper suggests that arbitrageurs do not always trade in the direction of eliminating mispricing. Local information may play a pivotal role in shaping arbitrageurs trading motivations.

References

Abreu, D., Brunnermeier, M., 2002. Synchronization risk and delayed arbitrage. Journal of Financial Economics 66, 341–360.

Abreu, D., Brunnermeier, M.K., 2003. Bubbles and crashes. Econometrica 71, 173–204.

Agarwal, V., Daniel, N.D., Naik, N.Y., 2009. Role of managerial incentives and discretion in hedge fund performance. Journal of Finance 64, 2221–2256.

Akbas, F., Armstrong, W.J., Sorescu, S., Subrahmanyam, A., 2015. Smart money, dumb money, and capital market anomalies. Journal of Financial Economics 118, 355–382.

Akbas, F., Boehmer, E., Erturk, B., Sorescu, S.M., 2013. Short interest, returns, and fundamentals .

Audretsch, D.B., Feldman, M.P., 1996. R&d spillovers and the geography of innovation and production. American Economic Review 86, 630–640.

Audretsch, D.B., Stephan, P.E., 1996. Company-scientist locational links: the case of biotechnology. American Economic Review 86, 641–652.

Baker, M., Savaşoglu, S., 2002. Limited arbitrage in mergers and acquisitions. Journal of Financial Economics 64, 91–115.

Barber, B.M., 1994. Noise trading and prime and score premiums. Journal of Empirical Finance 1, 251–278.

Barberis, N., Greenwood, R., Jin, L., Shleifer, A., 2016. Extrapolation and bubbles. Technical Report. National Bureau of Economic Research.

Brennan, M.J., 2004. How did it happen? Economic Notes 33, 3–22.

Brennan, M.J., Cao, H.H., 1997. International portfolio investment flows. Journal of Finance 52, 1851–1880.

Brown, J.R., Ivković, Z., Smith, P.A., Weisbenner, S., 2008. Neighbors matter: causal community effects and stock market participation. Journal of Finance 63, 1509–1531.

Brunnermeier, M., Nagel, S., 2004. Hedge funds and the technology bubble. Journal of Finance 59, 2013–2040.

Cao, C., Liang, B., Lo, A.W., Petrasek, L., 2014. Hedge fund holdings and stock market efficiency. Review of Asset Pricing Studies .

Coval, J., 2003. International capital flows when investors have local information. Division of Research, Harvard Business School.

Coval, J.D., Moskowitz, T.J., 1999. Home bias at home: local equity preference in domestic portfolios. Journal of Finance 54, 2045–2073.

Coval, J.D., Moskowitz, T.J., 2001. The geography of investment: informed trading and asset prices. Journal of Political Economy 109, 811–841.

De Long, J.B., Shleifer, A., Summers, L.H., Waldmann, R.J., 1990a. Noise trader risk in financial markets. Journal of Political Economy , 703–738.

De Long, J.B., Shleifer, A., Summers, L.H., Waldmann, R.J., 1990b. Positive feedback investment strategies and destabilizing rational speculation. Journal of Finance 45, 379–395.

Fama, E.F., 1965. The behavior of stock-market prices. Journal of Business 38, 34–105.

Frazzini, A., Pedersen, L.H., 2014. Betting against beta. Journal of Financial Economics 111, 1–25.

French, K.R., 1980. Stock returns and the weekend effect. Journal of Financial Economics 8, 55–69.

Friedman, M., 1953. The methodology of positive economics .

Gehrig, T., 1993. An information based explanation of the domestic bias in international equity investment. The Scandinavian Journal of Economics , 97–109.

Gong, B., Pan, D., Shi, D., 2016. New investors and bubbles: an analysis of the baosteel call warrant bubble. Management Science .

Greenwood, R., Nagel, S., 2009. Inexperienced investors and bubbles. Journal of Financial Economics 93, 239–258.

Griffin, J.M., Harris, J.H., Shu, T., Topaloglu, S., 2011. Who drove and burst the tech bubble? Journal of Finance 66, 1251–1290.

Gromb, D., Vayanos, D., 2010. Limits of arbitrage. Annu. Rev. Financ. Econ. 2, 251–275.

Hong, H., Kubik, J.D., Fishman, T., 2012. Do arbitrageurs amplify economic shocks? Journal of Financial Economics 103, 454–470.

Hong, H., Kubik, J.D., Stein, J.C., 2004. Social interaction and stock-market participation. Journal of Finance 59, 137–163.

Hong, H., Sraer, D.A., 2016. Speculative betas. Journal of Finance 71, 2095–2144.

Ingersoll, J.E., 1976. A theoretical and empirical investigation of the dual purpose funds: an application of contingent-claims analysis. Journal of Financial Economics 3, 83–123.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Quarterly Journal of Economics 108, 577–598.

Jarrow, R.A., O'Hara, M., 1989. Primes and scores: an essay on market imperfections. Journal of Finance 44, 1263–1287.

Kaustia, M., Knüpfer, S., 2012. Peer performance and stock market entry. Journal of Financial Economics 104, 321–338.

Kindleberger, C.P., 1978. Economic response: comparative studies in trade, finance, and growth. Harvard University Press.

Kokkonen, J., Suominen, M., 2015. Hedge funds and stock market efficiency. Management Science 61, 2890–2904.

Kyle, A.S., Xiong, W., 2001. Contagion as a wealth effect. Journal of Finance 56, 1401–1440.

Lakonishok, J., Levi, M., 1982. Weekend effects on stock returns: a note. Journal of Finance 37, 883–889.

Lee, C., Shleifer, A., Thaler, R.H., 1991. Investor sentiment and the closed-end fund puzzle. Journal of Finance 46, 75–109.

Lerner, J., 1995. Venture capitalists and the oversight of private firms. Journal of Finance 50, 301–318.

Litzenberger, R.H., Sosin, H.B., 1977. The structure and management of dual purpose funds. Journal of Financial Economics 4, 203–230.

Mackay, C., 1869. Memoirs of extraordinary popular delusions and the madness of crowds. George Routledge and Sons.

Ofek, E., Richardson, M., 2002. The valuation and market rationality of internet stock prices. Oxford Review of Economic Policy 18, 265–287.

Pástor, L., Veronesi, P., 2006. Was there a nasdaq bubble in the late 1990s? Journal of Financial Economics 81, 61–100.

Pontiff, J., 1996. Costly arbitrage: evidence from closed-end funds. Quarterly Journal of Economics , 1135–1151.

Puckett, A., Yan, X.S., 2011. The interim trading skills of institutional investors. Journal of Finance 66, 601–633.

Scheinkman, J.A., Xiong, W., 2003. Overconfidence and speculative bubbles. Journal of political Economy 111, 1183–1220.

Shiller, R.J., 2015. Irrational exuberance. Princeton university press.

Shleifer, A., Vishny, R.W., 1997. The limits of arbitrage. Journal of Finance 52, 35–55.

Sias, R., Turtle, H., Zykaj, B., 2015. Hedge fund crowds and mispricing. Management Science 62, 764–784.

Wu, J.J., Zhang, A.J., 2015. Have short sellers become more sophisticated? evidence from market anomalies .

Xiong, W., Yu, J., 2011. The chinese warrants bubble. American Economic Review 101, 2723–2753.

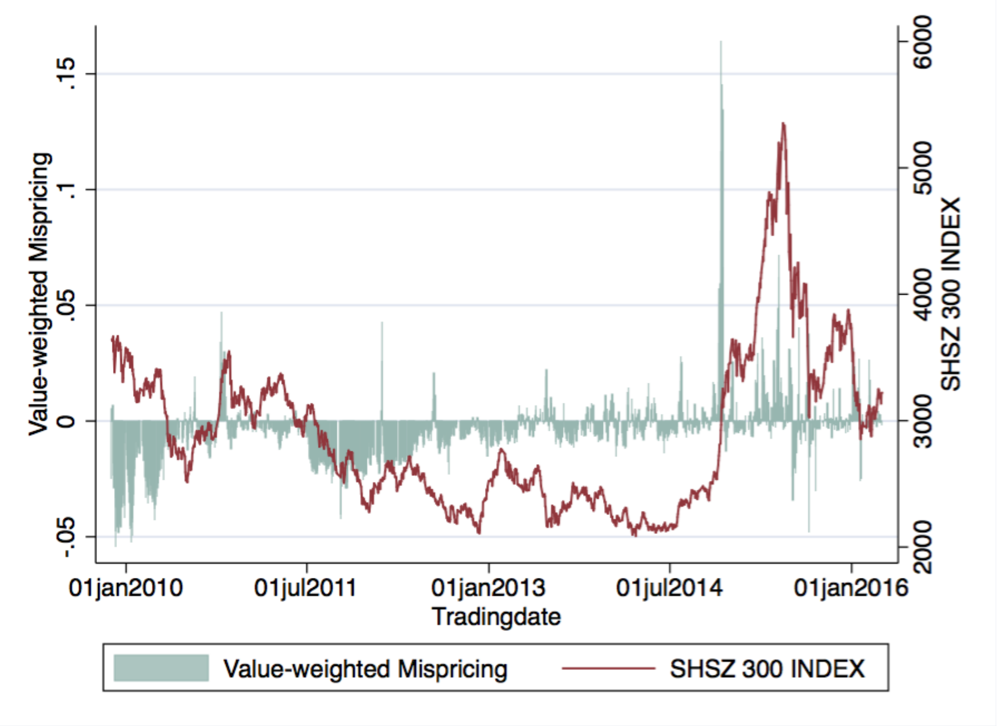


Figure 2: Value-weighted Mispricing across 150 Funds Marketwide. This figure presents the value-weighted average structured fund mispricing across 150 open-end structured funds in the market. The mispricing is calculated following Lee et al. (1991) using equation $VWP_t = \sum_{i=1}^{n_t} W_t * Mispricing_{it}$, where $W_i = \frac{NAV_{it}}{\sum_{i=1}^{nt} NAV_{it}}$ and NAV_{it} is equal to the net asset value of fund i at the end of period t, $Mispricing_{it} = \frac{(P_{it}^A + P_{it}^B) - NAV_{it}^M}{NAV_{it}^M} * 100$, $P_{it}^A + P_{it}^B$ is the synthetic market price of M in the stock market based on the prices of A and B. and n_t is the number of funds with available data at the end of period t.

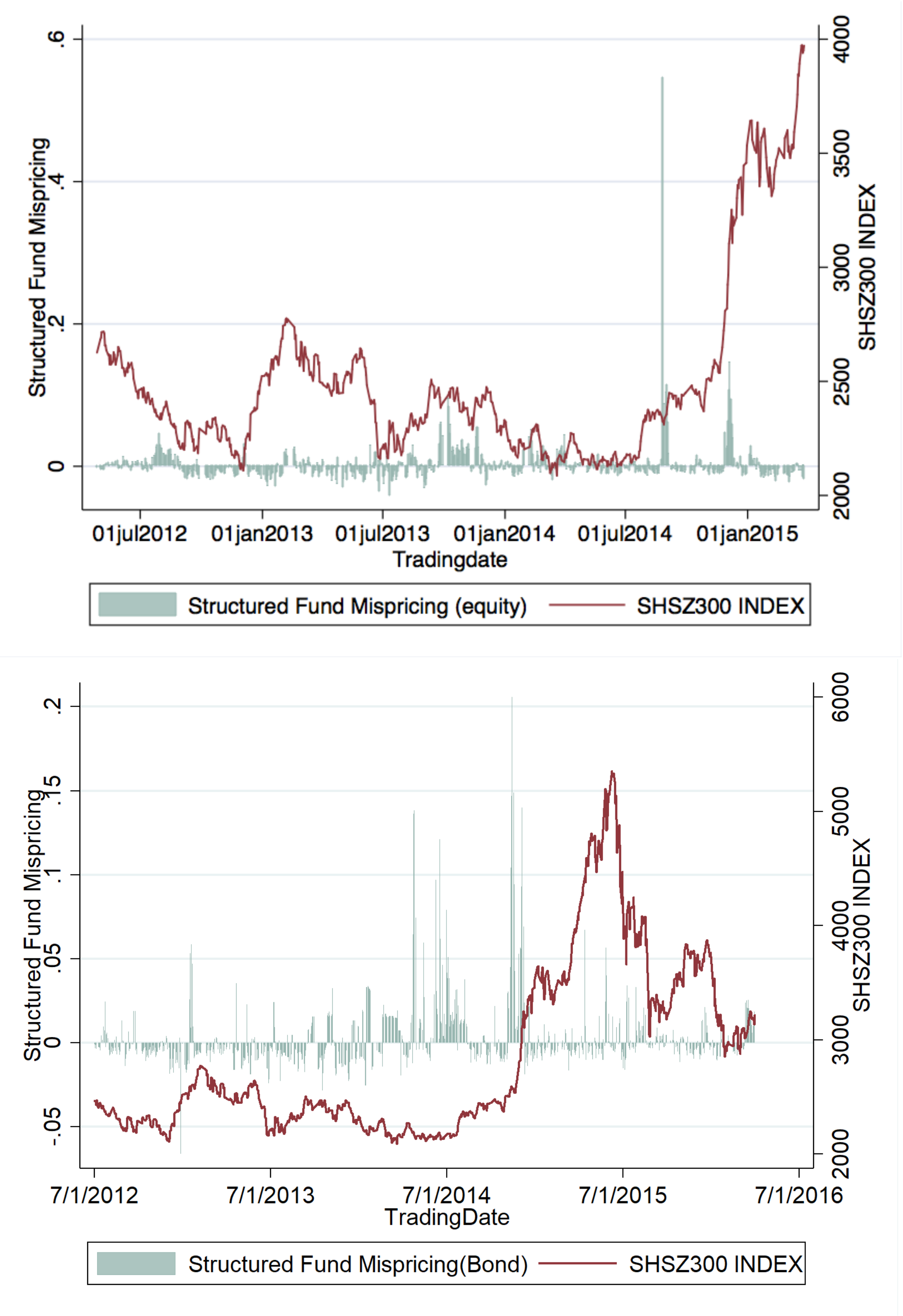


Figure 3: Mispricing for Two Funds with Investor Trading Data. This figure plots the historical daily mispricing for the two funds with investor trading data. Significant mispricing has already been discovered based on the market data of 150 funds. This demonstrates that the two funds that I use exhibit the same pattern of mispricing as the other structured funds in the market. Structured fund mispricing is calculated by $Mispricing_{it} = \frac{(P^A_{it}+P^B_{it})-NAV^M_{it}}{NAV^M_{it}}$, where $P^A_{it} + P^B_{it}$ is the synthetic market price of M in the stock market based on the prices of A and B.

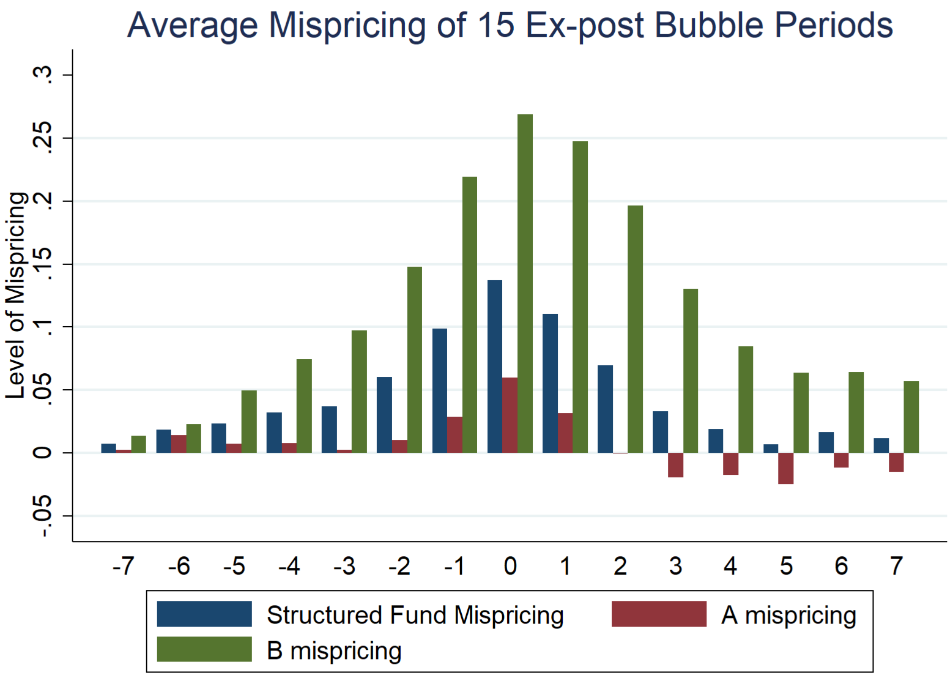


Figure 4: Define Bubble Periods. This figure presents the dynamics of structured fund mispricing and its two components (i.e., mispricing of A and mispricing of B) in event time. I define 15 bubbles based on the top-15 instances of structured fund mispricing. The event time $t = 0$ is the mispricing peak. The "bubble period" is 7 days before and after the premium peak. These 15 bubbles are independent in the sense that the bubble periods do not overlap one another. Based on the NAV of M and trading prices of A and B, $Mispricing_{it} = \frac{(P^A_{it}+P^B_{it})-NAV^M_{it}}{NAV^M_{it}}$, $Mispricing^A_{it} = \frac{P^A_{it}-NAV^A_{it}}{NAV^A_{it}}$,$Mispricing^B_{it} = \frac{P^B_{it}-NAV^B_{it}}{NAV^B_{it}}$.

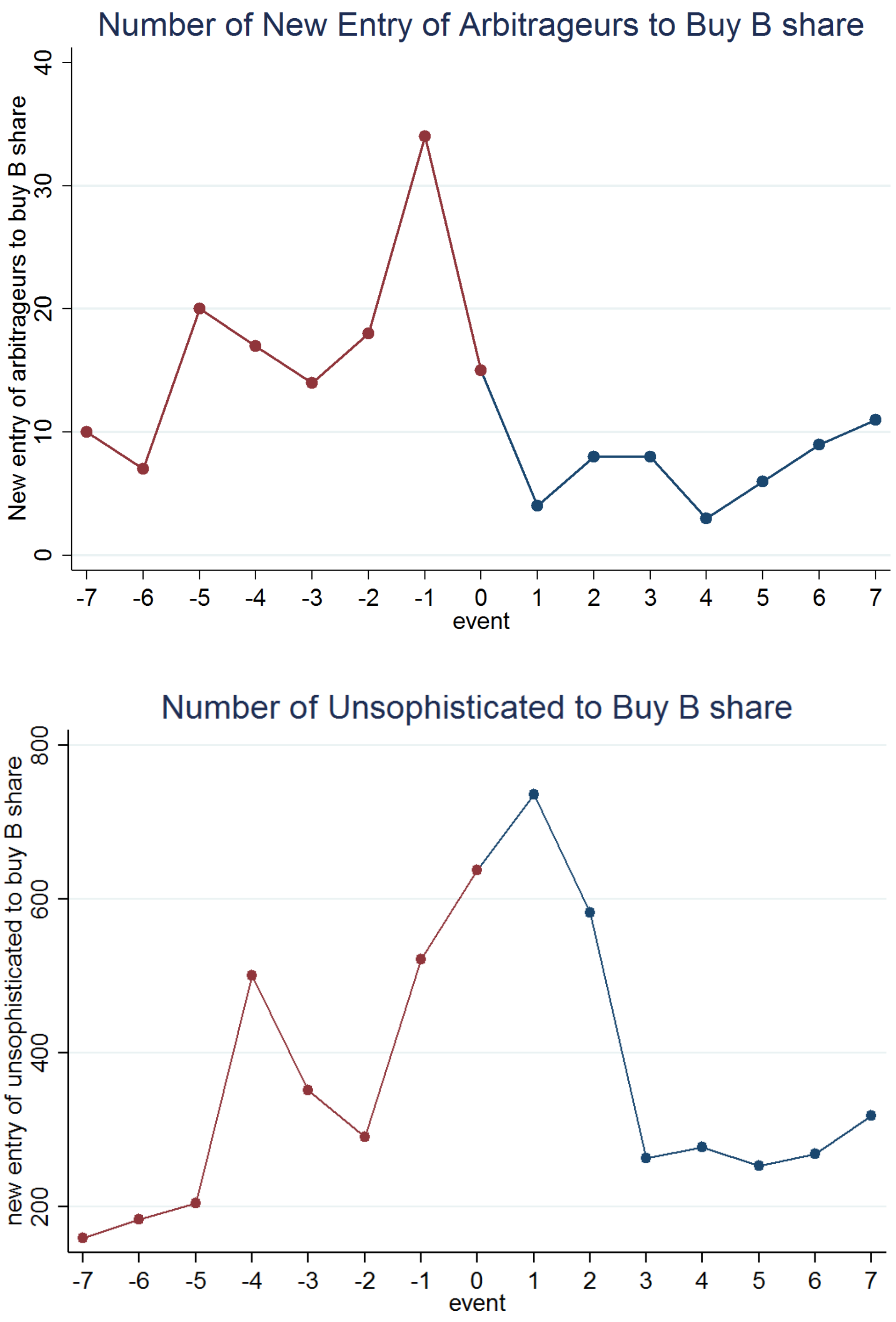


Figure 5: Arbitrageurs Riding Bubbles: New entry. This figure shows the number of new nvestors entering the market to buy B during ex post bubble periods for arbitrageurs and unsophisticated investors, respectively. The horizontal line is the event time, the vertical line is the number of new investors. The top figure is for arbitrageurs and the bottom figure is for unsophisticated investors. This is to show that the two groups of investors exhibit different trading behavior during bubble periods.

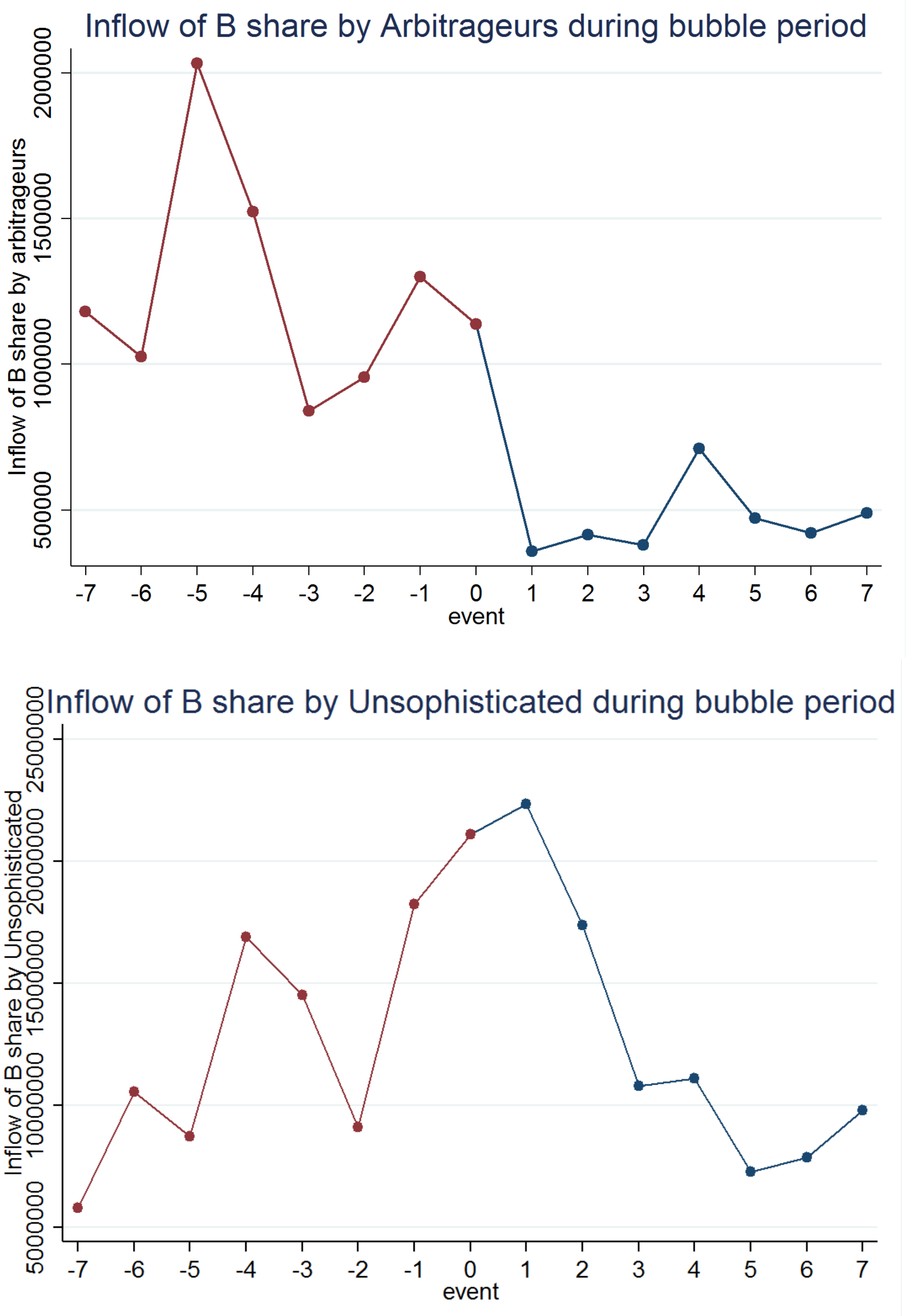


Figure 6: Arbitrageurs Riding Bubbles: Total Inflow. This figure shows the total inflows to buy B during ex post bubble periods for arbitrageurs and unsophisticated investors, respectively. The horizontal line is the event time, the vertical line is the number of new investors. The top figure is for arbitrageurs and the bottom figure is for unsophisticated investors. This is to show that the two groups of investors exhibit different trading behavior during bubble periods.

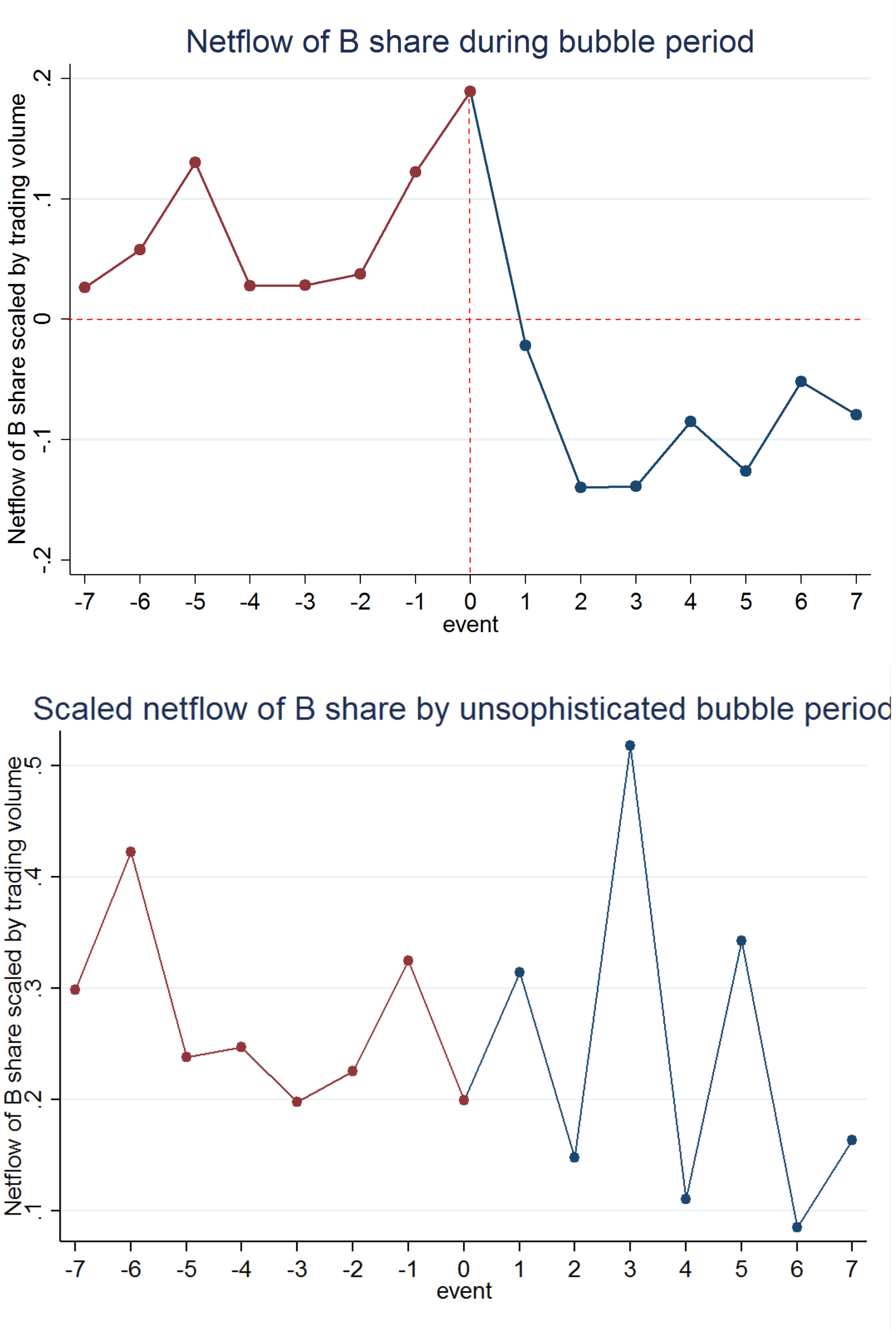


Figure 7: Arbitrageurs Riding Bubbles: Net flow Analysis. This figure shows the total net flows of B during ex-post bubble periods for arbitrageurs and unsophisticated investors, respectively. The horizontal line is the event time, the vertical line is the number of new investors. The top figure is for arbitrageurs and the bottom figure is for unsophisticated investors. This is to show that the two groups of investors exhibit different trading behavior during bubble periods.

Table 1: Descriptive Statistics of Investor Distribution

This table presents descriptive statistics of investor trading data in the two structured mutual funds. The summary statistics are based on a sample of mutual fund investors who trade in a top-30 mutual fund family in the period from September 2011 to December 2015. Panel A reports the total number of investors trading in M, A and B. Panel B presents the distribution of individual investor age in the full sample. Panel C shows the number of transactions in premium arbitrage and discount arbitrage and in trading A, B and M.

Panel A: Number of Investors

Fund Type	Investor Type	A	B	M	ABM	AB	AM	BM
Equity	Individual	981	9,012	17,748	1,297	113	265	1,650
Equity	institution	64	47	314	28	7	11	10
Bond	Individual	2,171	2,850	10,540	836	251	0	0
Bond	institution	9	14	34	10	11	0	0

Panel B: Distribution of Individual Investor Age

Age	Mean	Median	SD	Min	Max	P25	P75
Equity	44.92	43	13.14	2	99	33	51
Bond	47.48	46	14.9	8	96	35	59

Panel C: Number of Trades

Fund	Arbitrage Trades		A		B		M	
	Premium	Discount	Buy	Sell	Buy	Sell	Purchase	Redeem
Equity	1,046	1,609	4,975	4,889	38,777	33,535	32,952	20,216
Bond	494	1,567	6,443	5,428	7,623	6,929	15,356	12,323

Table 2: Descriptive Statistics of Mispricing

This table presents the summary statistics of mispricing in the structured mutual funds. Panel A reports numbers of observations, means, medians and standard deviations, along with the minimum, maximum, 25%, and 75% quintile values of the absolute value of mispricing for 150 open-end structured funds in the entire Chinese market. Panel B presents the distribution of the absolute value of mispricing for the two funds with investor trading information. Panel C reports the distribution of the absolute value of mispricing during bubble periods. Structured fund mispricing for fund i on day t is defined as $Mispricing_{it} = \frac{(P^A_{it}+P^B_{it})-NAV^M_{it}}{NAV^M_{it}}$, where $P^A_{it} + P^B_{it}$ is the synthetic market price of M in the stock market based on the prices of A and B. Mispricing of A is defined as $Mispricing^A_{it} = \frac{P^A_{it}-NAV^A_{it}}{NAV^A_{it}}$. Mispricing of B is defined as $Mispricing^B_{it} = \frac{P^B_{it}-NAV^B_{it}}{NAV^B_{it}}$. The bubble period in panel C is defined based on the structured fund mispricing. I rank the time series mispricing of each fund and define the top-15 premium peaks as bubble peaks, then define the 7 days before each bubble peak as the "bubble-formation period" and 7 days after each bubble peak as "bubble-bursting period".

Panel A: Distribution of Mispricing (whole market)

Fund Share	Type	N	Mean	SD	Min	Max	P25	P75
Equity	M	1418	0.013	0.013	0.000	0.187	0.007	0.014
	A	1418	0.067	0.035	0.000	0.217	0.044	0.085
	B	1418	0.071	0.045	0.001	0.370	0.047	0.086
Bond	M	1335	0.022	0.019	0.000	0.163	0.010	0.026
	A	1335	0.051	0.019	0.011	0.110	0.038	0.059
	B	1335	0.108	0.041	0.034	0.592	0.080	0.129

Panel B: Distribution of Mispricing (two funds)

Fund Share	Type	N	Mean	SD	Min	Max	P25	P75
Equity	M	707	0.013	0.025	0.000	0.546	0.004	0.015
	A	707	0.023	0.026	0.000	0.175	0.008	0.025
	B	707	0.033	0.053	0.000	1.131	0.008	0.049
Bond	M	914	0.011	0.018	0.000	0.206	0.003	0.012
	A	914	0.022	0.023	0.000	0.194	0.005	0.033
	B	914	0.043	0.050	0.000	0.575	0.018	0.053

Panel C: Distribution of Mispricing (bubble periods)

Fund Share	Type	N	Mean	SD	Min	Max	P25	P75
Equity	M	210	0.023	0.042	0.000	0.546	0.007	0.026
	A	210	0.031	0.035	0.000	0.175	0.011	0.027
	B	210	0.050	0.087	0.000	1.131	0.012	0.061
Bond	M	206	0.023	0.032	0.000	0.206	0.005	0.025
	A	206	0.031	0.032	0.000	0.194	0.004	0.046
	B	206	0.065	0.088	0.001	0.575	0.019	0.070

Table 3: Characters of Arbitrageurs vs. Unsophisticated Investors

This table tabulates the characteristics of the two groups of investors: arbitrageurs and unsophisticated investors for the equity fund with investor trading data. Panel A reports numbers of observations, means and standard deviations, along with the 5%, 25%, 75% and 95% quintile values of differences in age, wealth (RMB value) and trading experience for the two groups of investors. Wealth represents the average RMB value of each purchase across all trading accounts. Experience is constructed based on the number of days between the first trading date and the last trading date. Panel B presents the trading characteristics of the two groups of investors in the equity fund. Trading size is the daily average RMB value of total purchases and sales across all trading accounts for each group of investors. Trading volume is the daily average number of shares bought and sold across all trading accounts for each group of investors. Profitability measures the average realized return per transaction across all trading assets and trading accounts for each group of investors. Panel C reports the average realized return per transaction in trading other mutual funds in the same mutual fund family for the two groups of investors.

Panel A: Investor Characteristics

Investor Type	Variable	Mean	SD	P5	P25	P75	P95
Arbitrageur	Age	45.03	12.9	29	35	52	70
	Wealth (RMB)	88,353	509,094	1,380	3,576	24,857	350,000
	Experience	261.68	284.29	15	36	412	903
Unsophisticated investors	Age	43.25	12.9	25	33	51	68
	Wealth (RMB)	59,059	739,790	100	500	10,500	100,000
	Experience	161.95	221.06	7	32	195	606

Panel B: Investor Characteristics by Trading Structured Funds

Investor Type	Variable	Mean	SD	P5	P25	P75	P95
Arbitrageur	Trading size (RMB)	112,451	467,931	2,000	8,000	66,216	417,870
	Trading volume (shares)	102,498	1,253,041	1,900	6,378	49,600	290,000
	Profitability	2.39%	18.50%	-25%	-5.60%	9.00%	21.03%
Unsophisticated investors	Trading size (RMB)	69,414	1,443,271	618	2,543	23,376	113,110
	Trading volume (shares)	54,583	1,203,344	400	1,900	18,000	91,500
	Profitability	-0.43%	12.90%	-15%	-2.20%	3.90%	23.65%

Panel C: Investor Characteristics by Trading Other Funds

Investor Type	Variable	Mean	SD	P5	P25	P75	P95
Arbitrageur	Profitability	0.86%	14.22%	-11.64%	-1.62%	0.72%	17.58%
Unsophisticated investors	Profitability	-0.53%	28.03%	-40%	-3.59%	5.33%	25.90%

Table 4: Performance of Arbitrageurs vs. Unsophisticated Investors

This table provides realized return analysis for the two groups of investors: arbitrageurs and unsophisticated investors. Arbitrageurs are defined as those investors who engage in arbitrage trading at least once when arbitrage opportunities arise at any point in their trading history with the fund. Unsophisticated investors are those investors who never conduct arbitrage trades even when they are presented with profitable arbitrage opportunities because they are not sophisticated enough to exploit those opportunities. The transactions can be separated into normal trades and arbitrage trades. Arbitrageurs conduct normal trades, such as trading A and B independently, and they also engage in arbitrage trading. However, unsophisticated investors only trade M, A and B independently in the stock market. First, I calculate the purchasing cost of each selling transaction using the FIFO (first in first out) convention based on the entire trading history of the investors and calculate realized return based on their historical purchasing cost. Then, realized returns are aggregated across all investors in each group for date t using the equal-weighted and value-weighted methods. Finally, I conduct a t-test for the time series of average realized returns. Panel A shows the equal-weighted average realized return after transaction costs, and panel B shows the value-weighted average return after transaction costs. Panel A1 presents the equal-weighted average realized return during bubble periods, and panel B1 presents the equal-weighted average realized return during bubble periods. As in table 2, I rank the time series mispricing of the fund and define the top-15 premiums as bubble peaks, then define the 7 days before each bubble peak as the "bubble-formation period" and the 7 days after each bubble peak as the "bubble-bursting period". Robust t-statistics are reported in parenthesis and based on standard errors clustered by city. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Equal-weighted realized return analysis (after transaction costs)

Group of Investors	Normal Trades			Arbitrage Trades	
	M share	A share	B share	Premium Arbitrage	Discount Arbitrage
Arbitrageur		2.96%	3.20%	3.19%	0.67%
		(16.58)***	(18.78)***	(15.87***)	(34.76***)
Unsophisticated investors	-1.07%	-0.07%	0.12%		
	(-7.28)***	(-1.08)	(2.18**)		

Panel A1: Equal-weighted realized return analysis (bubble period)

Group of Investors	Normal Trades			Arbitrage Trades	
	M share	A share	B share	Premium Arbitrage	Discount Arbitrage
Arbitrageur		-2.05%	11.51%	4.46%	0.65%
		(-8.25)***	(27.58)***	(16.55)***	(11.41)***
Unsophisticated investors	-0.95%	-2.20%	-1.05%		
	(-3.35)***	(-10.8)***	(-8.02)***		

Panel B: Value-weighted realized return analysis (after transaction costs)

Group of Investors	Normal Trades			Arbitrage Trades	
	M share	A share	B share	Premium Arbitrage	Discount Arbitrage
Arbitrageur		2.66%	0.36%	3.34%	0.72%
		(14.82)***	(2.05)***	(20.60)***	(32.87)***
Unsophisticated investors	-0.97%	-0.23%	0.11%		
	(-7.03)***	(-3.35)***	(2.06)***		

Panel B1: Value-weighted realized return analysis (bubble period)

Group of Investors	Normal Trades			Arbitrage Trades	
	M share	A share	B share	Premium Arbitrage	Discount Arbitrage
Arbitrageur		-1.73%	8.69%	3.40%	0.56%
		(-6.52)***	(21.6)***	(17.52)***	(16.81)***
Unsophisticated investors	-0.82%	-2.40%	-1.01%		
	(2.92)***	(11.39)***	(-7.51)***		

Table 5: How Do Arbitrageurs Trade: Baseline Analysis on Potential Bubbles

This table presents the baseline relationship between the participation of unsophisticated investors and arbitrageur behavior and examines the results of the following daily panel regressions with time and city fixed effects:

$$Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \alpha_0 + \alpha_1 * New\,entry^{U}_{c,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^{U}_{c,t-1} + \alpha_4 * New\,entry^{U}_{c,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^{U}_{c,t-1} * Mispricing_{t-1} + \alpha_6 * M_{c,t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in city c on day t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $New\,entry^{U}_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Other\,entry^{U}_{c,t-1}$ is the logarithm of the total number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1. $Mispricing_{t-1}$ is the positive mispricing of a structured fund for fund f on day t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles). The vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs who buy B at t-1 and mispricing for fund f on day t. I include time and city fixed effects in all specifications. Time fixed effects remove the time trend, and city fixed effects control for the time-invariant unobservable heterogeneity across different cities. Please refer to the Internet Appendix for variable definitions.

Panel A presents the impact of the participation of unsophisticated investors on arbitrage flows and panel B presents the impact of the participation of unsophisticated investors on arbitrageur riding-bubble flows. Arbitrageur is defined as potentially riding bubble if she is buying B to push up mispricing even when mispricing is positive.

The standard errors are robust to heteroskedasticity and clustered at the city level.The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Panel A: Premium arbitrage flows following unsophisticated investor participation

Variables	Log (premium arbitrage flow)					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^{U}_{c,t-1}$	1.250		1.250	0.894	0.960	0.525
	(6.82)***		(6.82)***	(5.61)***	(5.59)***	(3.63)***
$Mispricing_{t-1}$		0.107	0.056	0.060	0.171	0.119
		(0.86)	(0.71)	(0.70)	(1.39)	(1.14)
$Other\,entry^{U}_{c,t-1}$					0.121	0.077
					(1.60)	(1.22)
$New\,entry^{U}_{c,t-1} * Mispricing_{t-1}$				0.059	0.042	0.055
				(2.43)**	(2.05)**	(3.09)***
$Other\,entry^{U}_{c,t-1} * Mispricing_{t-1}$					-0.499	-0.428
					(-1.96)*	(-1.84)*
$Log(premium\,arbitrage\,flow)_{c,t-1}$						0.274
						(9.51)***
$Mispricing_{t}$						-14.325
						(-2.18)**
Constant	0.065	0.025	0.051	0.050	-0.699	-0.591
	(7.59)***	(0.69)	(2.18)**	(1.99)**	(-2.43)**	(-2.04)**
TIME FE	YES	YES	YES	YES	YES	YES
CITY FE	YES	YES	YES	YES	YES	YES
Observations	93,128	93,128	93,128	93,128	93,128	93,128
r2	0.140	0.110	0.140	0.142	0.143	0.207

Panel B: Arbitrageur riding bubble and unsophisticated investor participation

Variables	Log (num of arbis buying B)					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^{U}_{c,t-1}$	0.037		0.037	0.080	0.083	0.058
	(2.60)***		(2.59)***	(3.47)***	(3.49)***	(3.26)***
$Mispricing_{t-1}$		-0.114	-0.116	-0.116	-0.106	-0.072
		((-1.47)	(-1.46)	(-1.46)	(-1.46)	(-1.39)
$Other\,entry^{U}_{c,t-1}$					0.043	0.021
					(1.69)*	(1.30)
$New\,entry^{U}_{c,t-1} * Mispricing_{t-1}$				-0.007	-0.008	-0.005
				(-3.67)***	(-3.87)***	(-3.77)***
$Other\,entry^{U}_{c,t-1} * Mispricing_{t-1}$					-0.015	-0.007
					(-1.21)	(-0.56)
$Log(num\,of\,arbis\,buying\,B)_{c,t-1}$						0.276
						(5.67)***
$Mispricing_{t}$						-0.591
						(-1.65)*
Constant	-0.002	0.026	0.027	0.027	-0.018	-0.005
	(-3.30)***	(1.39)	(1.40)	(1.40)	(-0.61)	(-0.25)
TIME FE	YES	YES	YES	YES	YES	YES
CITY FE	YES	YES	YES	YES	YES	YES
Observations	93,128	93,128	93,128	93,128	93,128	93,128
r2	0.097	0.091	0.099	0.107	0.108	0.175

Table 6: Two-Stage Regression Analysis

This table provides results of a two-stage least-square specification used to estimate the impact of the participation of unsophisticated investors on arbitrageur behavior in the period from 2011 to 2015. In the first stage, I predict the estimated number of new entries of unsophisticated investors using the following panel regression:

$$Newentry_{c,t} = \beta_0 + \beta_1 Posreturn\,num^U_{c,t-1} + \beta_2 Posreturn\,mean^U_{c,t-1} + \beta_3 Negreturn\,num^U_{c,t} + \beta_4 Negreturn\,mean^U_{c,t-1} + \beta_5 CEFD_{t-1} + \epsilon_{c,t}$$

where subscripts c and t refer to city and week, respectively. $Newentry^U_{c,t}$ is the number of new unsophisticated investors entering the stock market to buy B in city c and week t-1; New unsophisticated investor is defined as the first time such an investor enters the market to trade; $Posreturn\,num^U_{c,t-1}$ is the number of unsophisticated investors who experience positive returns since their purchase in city c and week t-1; $Posreturn\,mean^U_{c,t-1}$ is the average positive returns of those investors in city c and week t-1; $Negreturn\,num^U_{c,t}$ is the number of unsophisticated investors who experience negative returns since their purchase in city c and week t-1; and $Negreturn\,mean^U_{c,t-1}$ is the average negative returns of those investors in city c and week t-1. $CEFD_{t-1}$ is the closed-end fund discount used as a proxy to control for investor sentiment in the market.

In the second stage, I regress the arbitrage flows and riding-bubble flows on the predicted number of new entries of unsophisticated investors in the following regression:

$$Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \alpha_0 + \alpha_1 * Predicted\,new\,entry^U_{c,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^U_{c,t-1} + \alpha_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^U_{c,t-1} * Mispricing_{t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in city c on day t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $Predicted\,new\,entry^U_{c,t-1}$ is the logarithm of the estimated new entries of unsophisticated investors, which is the fitted part of $\beta_1 Posreturn\,num^U_{c,t-1}$ in the first-stage regression; and $Other\,entry^U_{c,t-1}$ is the logarithm of the total number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1. $Mispricing_{t-1}$ is the structured fund mispricing for fund f on day t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles).

Panel A presents the regression results of the first stage and panel B presents the results of the impact of unsophisticated participation on arbitrage flows and riding-bubbles flows. Arbitrageur is defined as potentially riding bubble if she or he is buying B to push up mispricing even when mispricing is positive.

The standard errors are robust to heteroscedasticity and clustered at the city level. The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Panel A: 1st stage social contagion predicting new entry of unsophisticated investors

Variables	New entry of unsophisticated investors					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$Posreturn\, num^U_{c,t-1}$	0.191	0.199			0.192	0.101
	(6.19)***	(5.80)***			(6.30)***	(5.83)***
$Posreturn\, mean^U_{c,t-1}$		-0.061				0.248
		(-0.61)				(1.42)
$Negreturn\, num^U_{c,t-1}$			-0.022	0.030	0.004	0.030
			(-1.31)	(1.68)*	(0.36)	(1.51)
$Negreturn\, mean^U_{c,t-1}$				-0.340		-1.977
				(-0.71)		(-1.88)*
$CEFD_{t-1}$	-0.442	-0.460	-0.577	-0.707	-0.438	-2.033
	(-3.32)***	(-3.28)***	(-7.08)***	(-2.13)**	(-3.32)***	(-1.60)
Constant	-2.370	-2.921	5.102	5.599	-2.514	14.421
	(-1.88)*	(-1.94)*	(6.59)***	(1.90)*	(-2.22)**	(1.31)
TIME FE	YES	YES	YES	YES	YES	YES
CITY FE	YES	YES	YES	YES	YES	YES
Observations	17,301	12,050	17,301	10,060	17,301	4,809
r2	0.420	0.430	0.271	0.381	0.420	0.512

Panel B: 2nd stage arbitrageur behavior following predicted unsophisticated investor participation

Variables	Log(premium arbitrage flow)			Log(num of arbis buying B)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$Predicted\, new\, entry^U_{c,t-1}$	0.053	0.056	0.037	0.002	0.003	0.003
	((4.09)***	(3.96)***	(4.44)***	(6.62)***	(4.75)***	(4.99)***
$Mispricing_{t-1}$	0.106	-0.060	-0.060	-0.112	-0.114	-0.112
	(0.87)	(-0.37)	(-0.48)	(-1.67)*	(-1.56)	(-1.66)*
$Other\, entry^U_{c,t-1}$		0.532	0.538		0.006	0.010
		(1.98)**	(1.22)		(0.25)	(0.38)
$Predicted\, new\, entry^U_{c,t-1} * Mispricing_{t-1}$	0.009	0.006	0.005	-0.005	-0.004	-0.004
	(2.99)***	(2.98)***	(3.06)***	(-2.01)**	(-2.03)**	(-2.04)**
$Other\, entry^U_{c,t-1} * Mispricing_{t-1}$		-0.545	-0.470		-0.001	-0.004
		(-2.88)***	(-2.95)***		(-0.63)	(-0.32)
Constant	0.203	0.088	-0.498	0.020	0.028	0.017
	(2.01)**	(-1.73)*	(-1.91)*	(0.82)	(1.51)	(0.63)
CONTROLS	NO	NO	YES	NO	NO	YES
TIME FE	YES	YES	YES	YES	YES	YES
CITY FE	YES	YES	YES	YES	YES	YES
Observations	93,128	93,128	93,128	93,128	93,128	93,128
r2	0.114	0.116	0.192	0.091	0.091	0.092

Table 7: Why Do Arbitrageurs Ride Bubble?

This table provides realized return analysis of arbitrageurs riding bubble and arbitrageurs not riding bubble during ex post bubble periods. Arbitrageurs are labeled as riding bubbles if they ever buy B during the bubble period when the structured fund mispricing already exceeds the transaction costs, meaning that arbitrageurs are supposed to buy M, partition it into A and B and sell them in the market. Arbitrageurs are labeled as not riding bubbles if they never buy B during the bubble period when the market premium already exceeds the transaction costs, meaning that profitable arbitrage opportunities exist. As in table 2, I rank the time series mispricing of each fund and define the top-15 premium peaks as bubble peaks, then define the 7 days before each bubble peak as the "bubble-formation period" and 7 days after each bubble peak as "bubble-bursting period".

First, I calculate the purchasing cost of each selling transaction using the FIFO (first in first out) convention based on the entire trading history of the investors and calculate realized return based on their historical purchasing cost. Then, realized returns are aggregated across all investors in each group for date t using the equal-weighted and value-weighted methods. Finally, I conduct a t-test for the time series of average realized returns. Panel A shows the equal-weighted average realized return for total realized return, arbitrage trading return and riding bubble return for arbitrageurs after transaction costs and panel B shows the value-weighted average return for total realized return, arbitrage trading return and riding bubble return for arbitrageurs after transaction costs. Robust t-statistics are reported in parenthesis and based on standard errors clustered by city. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Equal-weighted realized return analysis (after transaction cost)

Group of Investors	Realized Return(bubble period)	Arbitrage Trades	Riding Bubble
Arbitrageur riding on bubble	7.39%	5.21%	11.35%
	(14.38)***	(18.24)***	(15.86)***
Arbitrageur not riding on bubble	3.14%	3.58%	
	(6.63)***	(9.08)***	

Panel B: Value-weighted realized return analysis (after transaction cost)

Group of Investors	Realized Return(bubble period)	Arbitrage Trades	Riding Bubble
Arbitrageur riding on bubble	6.39%	5.52%	8.26%
	(30.84)***	(18.52)***	(20.73)***
Arbitrageur not riding on bubble	2.54%	3.67%	
	(54.12)***	(9.84)***	

Table 8: Who Are Riding Bubble among Arbitrageurs?

This table presents the group of arbitrageur who are riding bubbles and examines the results of the following daily panel regressions with time and city fixed effects:

$$Riding\,bubble_{c,t} = \lambda_0 + \lambda_1 * New\,entry^{U}_{c,t-1} + \lambda_2 * Mispricing_{t-1} + \lambda_3 * Non\,inventory_{c,t-1} + \lambda_4 * New\,entry^{U}_{c,t-1} * Mispricing_{t-1} + \lambda_5 * Non\,inventory_{c,t-1} * Mispricing_{t-1} + \lambda_6 * New\,entry^{U}_{c,t-1} * Mispricing_{t-1} * Non\,inventory_{c,t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and day, respectively. $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B at time t; $New\,entry^{U}_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c at time t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Non\,inventory_{c,t-1}$ is the logarithm of the number of arbitrageurs who do not have inventory of B in their account in city c at time t-1. $Mispricing_{t-1}$ is the structured fund mispricing for fund f at t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles). Please refer to the Internet Appendix for variable definitions.

The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Panel A: Who ride bubbles among arbitrageurs?

Variables	log (num of arbis buying B)					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^{U}_{c,t-1}$	0.037	0.037	0.075	0.044	0.020	0.022
	(4.02)***	(4.03)***	(4.31)***	(3.53)***	(1.70)*	(1.82)*
$Mispricing_{t-1}$		-0.116	-0.116	-0.117	-0.117	-0.108
		(-1.57)	(-1.55)	(-1.55)	(-1.55)	(-1.68)*
$Non\,inventory_{c,t-1}$			0.021	0.001	-0.000	-0.000
			(2.38)**	(0.16)	(-0.02)	(-0.02)
$New\,entry^{U}_{c,t-1} * Mispricing_{t-1}$			-0.007	-0.007	-0.003	-0.003
			(-4.09)***	(-4.31)***	(-2.59)***	(-2.67)***
$New\,entry^{U}_{c,t-1} * Non\,inventory_{c,t-1}$				0.050	0.086	0.086
				(3.33)***	(3.30)***	(3.30)***
$New\,entry^{U}_{c,t-1} * Mispricing_{t-1} * Non\,inventory_{c,t-1}$					-0.006	-0.006
					(-2.15)**	(-2.15)**
Other entry						0.024
						(0.98)
Constant	-0.002	0.027	0.027	0.027	0.027	-0.001
	(-2.47)**	(1.47)	(1.46)	(1.46)	(1.45)	(-0.04)
TIME FE	YES	YES	YES	YES	YES	YES
CITY FE	YES	YES	YES	YES	YES	YES
Observations	93,128	93,128	93,128	93,128	93,128	93,128
r2	0.097	0.098	0.109	0.119	0.124	0.124

Panel B: Riding bubble arbitrageur characters

Investor Type	Variable	Mean	SD	Median	P5	P95
Riding	Age	46.90	12.61	45.00	29.00	70.00
	Experience	392.47	261.39	376.00	28.00	876.00
Not Riding	Age	44.53	13.21	41.00	28.00	70.00
	Experience	181.50	222.31	106.00	14.00	770.00

Panel C: Flow Distribution

Investor Type	Variable	Mean	SD	Median	P5	P95
Riding	arbitrage flow 0.33	0.29	0.28	0.00	0.93	
	riding bubble flow	0.46	0.35	0.45	0.00	0.97
Not Riding	arbitrage flow	0.21	0.27	0.09	0.00	0.83

Table 9: How Do Arbitrageurs Trade: Migrant Arbitrageur

This table presents the relationship between migrant arbitrageurs behavior and the unsophisticated investors participation in their city of residence vs. their hometown and examines the results of the following daily panel regressions with time and city fixed effects:

$$Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \gamma_0 + \gamma_1 * New\,entry^U_{c,t-1} + \gamma_2 * Mispricing_{t-1} + \gamma_3 * Hometown\,entry^U_{c,t-1} + \gamma_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \gamma_5 * Hometown\,entry^U_{c,t-1} * Mispricing_{t-1} + \gamma_6 * M_{c,t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and day, respectively. Migrant arbitrageurs are defined as those arbitrageurs whose hometown province and province of residence are different. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage for migrant arbitrageurs in city c on day t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of migrant arbitrageurs who buy B on day t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Hometown\,entry^U_{c,t-1}$ is the logarithm of the average number of new unsophisticated investors entering the stock market to buy B for the hometown cities of all migrant arbitrageurs in city c on day t-1. $Mispricing_{t-1}$ is the structured fund mispricing for fund f at t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles), and the vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs who buy B at t-1 and the mispricing for fund f on day t.

The standard errors are robust to heteroskedasticity and clustered at the city level. The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Variables	Log (premium arbitrage flow)			Log (num of arbis buying B)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^U_{c,t-1}$	0.894	0.483	0.344	0.037	0.030	0.029
	(5.61)***	(3.60)***	(2.83)**	(2.59)***	(2.11)**	(2.02)**
$Mispricing_{t-1}$	0.060	0.027	0.011	0.034	-0.006	-0.130
	(0.70)	(0.42)	(0.19)	(1.27)	(-0.25)	(-1.34)
$Hometown\,entry^U_{t-1}$		0.041	-0.549		0.046	0.041
		(0.17)	(-3.15)***		(1.73)*	(1.66)*
$New\,entry^U_{c,t-1} * Mispricing_{t-1}$	0.059	0.062	0.032	-0.001	-0.006	-0.006
	(2.43)**	(3.16)***	(2.19)*	(-1.65)*	(-2.62)**	(-2.56)***
$Hometown\,entry^U_{t-1} * Mispricing_{t-1}$		0.055	0.001		-0.002	-0.002
		(1.64)	(0.03)		(-0.97)	(-1.01)
Constant	0.050	0.040	0.016	-0.002	-0.002	0.017
	(1.99)**	(2.03)**	(0.90)	(-3.30)***	(-2.82)***	(1.26)
CONTROLS	NO	NO	YES	NO	NO	YES
TIME FE	YES	YES	YES	YES	YES	YES
CITY FE	YES	YES	YES	YES	YES	YES
Observations	93,128	93,128	93,128	93,128	93,128	93,128
r2	0.106	0.295	0.338	0.097	0.176	0.178

Table 10: The Impact of Social Contagion and Feedback Trading on Mispricing

This panel regression tests the impact of new entry of unsophisticated investors due to social contagion and feedback trading on mispricing of the equity structured funds:

$$Mispricing_{f,t} = \phi_0 + \phi_1 New\,entry^U_{f,t-1} + \phi_2 Repeat\,entry^U_{f,t-1} + \phi_3 B\,volume_{f,t-1} + \phi_4 A\,volume_{f,t-1} + \phi_5 A\,volatility_{f,t-1} + \phi_6 B\,volatility_{f,t-1} + \epsilon_{f,t}$$

where $Mispricing_{f,t}$ is the structured fund mispricing calculated as the difference between the synthetic trading price of A, B and the NAV of M for fund f at time t. $New\,entry^U_{f,t-1}$ is the total number of new unsophisticated investors aggregated at fund level, which is due to social contagion effects among unsophisticated investors. $Repeat\,entry^U_{f,t-1}$ is the total number of repeated entry of unsophisticated investors aggregated at fund level, which is due to the positive feedback trading for each unsophisticated investor herself (for evidence of positive feedback trading, please refer to the appendix). Control variables for the regression include trading volume of A, B for fund f at time t and standard deviation of past 5 days return of B. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Dependent Variable is the mispricing of the structured fund						
Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^U_{f,t-1}$	0.005	0.005			0.001	0.004
	(3.82)***	(4.57)***			(2.26)**	(3.22)***
$Repeat\,entry^U_{f,t-1}$			0.008	0.006	0.007	0.003
			(2.05)**	(4.93)***	(1.06)	(1.78)*
$B\,volume_{f,t-1}$		-0.001		-0.001		-0.001
		(-2.06)**		(-2.51)**		(-2.35)**
$A\,volume_{f,t-1}$		-0.003		-0.002		-0.003
		(-3.27)***		(-2.51)**		(-4.32)***
$A\,volatility_{f,t-1}$		-0.000		-0.000		-0.000
		(-0.74)		(-0.32)		(-0.27)
$B\,volatility_{f,t-1}$		0.003		0.004		0.003
		(3.31)***		(3.99)***		(3.24)***
Constant	-0.005	0.032	-0.033	0.002	-0.031	0.029
	(-4.00)***	(3.64)***	(-2.01)**	(0.31)	(-1.24)	(3.13)***
Observations	706	574	706	574	706	574
r2	0.047	0.129	0.072	0.104	0.072	0.134

Table 11: The Impact of Social Contagion and Feedback Trading on Arbitrageur Behavior

This table presents the relationship between the participation of unsophisticated investors and arbitrageur behavior and examines the results of the following daily panel regressions with time and city fixed effects:

$$Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \delta_0 + \delta_1 * New\,entry^U_{c,t-1} + \delta_2 * Mispricing_{t-1} + \delta_3 * Repeat\,entry^U_{c,t-1} + \delta_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \delta_5 * Repeat\,entry^U_{c,t-1} * Mispricing_{t-1} + \delta_6 * M_{c,t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares partitioned from M for the premium arbitrage in city c on day t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1, which is mostly due to social contagion effect among unsophisticated investors. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Repeat\,entry^U_{c,t-1}$ is the logarithm of the number of existing unsophisticated investors entering the stock market to buy B after the first trade in B in city c on day t, which is mostly due to positive feedback trading effect. $Mispricing_{t-1}$ is the structured fund mispricing for fund f at t-1 which is calculated as the difference between synthetic trading price in the market and the NAV of M; here it represents the days when mispricing is positive (when there is ex ante potential bubbles). The vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs riding bubbles at t-1 and mispricing for fund f on day t. Please refer to Internet Appendix for variable definitions.

The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Variables	Log (premium arbitrage flow)			Log (num of arbis buying B)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^U_{c,t-1}$	0.894	0.724	0.404	0.080	0.050	0.044
	(4.70)***	(3.94)***	(2.40)**	(4.43)***	(2.93)***	(2.81)***
$Mispricing_{t-1}$	0.060	0.069	0.030	-0.116	-0.113	-0.078
	(0.25)	(0.27)	(0.16)	(-1.57)	(-1.60)	(-1.73)*
$Repeat\,entry^U_{c,t-1}$		0.227	0.085		0.049	0.023
		(2.43)**	(0.98)		(5.93)***	(3.49)***
$New\,entry^U_{c,t-1} * Mispricing_{t-1}$	0.059	0.034	0.047	-0.007	-0.005	-0.004
	(2.48)**	(2.63)***	(2.12)**	(-4.11)***	(-2.89)***	(-2.37)**
$Repeat\,entry^U_{c,t-1} * Mispricing_{t-1}$		0.070	0.057		-0.003	-0.002
		(1.39)	(2.57)**		(-1.94)*	(1.07)
$Log(DV)_{c,t-1}$			0.267			0.263
			(12.37)***			(6.27)***
$Mispricing_t$			0.213			-0.137
			(0.53)			(-1.79)*
Constant	0.088	0.088	0.115	0.027	0.027	0.018
	(1.59)	(1.59)	(1.66)*	(1.48)	(1.52)	(1.58)
CONTROLS	NO	NO	YES	NO	NO	YES
TIME FE	93,128	93,128	93,128	93,128	93,128	93,128
CITY FE	YES	YES	YES	YES	YES	YES
Observations	YES	YES	YES	YES	YES	YES
r2	0.142	0.150	0.210	0.107	0.122	0.178

Table 12: Robustness Check (new definitions of arbitrageur)

This table presents the robustness check for the baseline regression in table 5 using new definitions of arbitrageurs and examines the results of the following daily panel regressions with time and city fixed effects:

$$Arbitrage\,flow(new)_{c,t}/Riding\,bubble(new)_{c,t} = \alpha_0 + \alpha_1 * New\,entry^U_{c,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^U_{c,t-1} + \alpha_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_6 * M_{c,t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\,flow(new)_{c,t}$is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in city c on day t under new definitions; $Riding\,bubble(new)_{c,t}$is the logarithm of the total number of arbitrageurs (under new definitions) who buy B on day t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Other\,entry^U_{c,t-1}$ is logarithm of the total number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1. $Mispricing_{t-1}$ is the positive structured fund mispricing for fund f at t-1 which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here it represents the days when mispricing is positive (when there is ex ante potential bubbles). The vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs who buy B at t-1 and mispricing for fund f at time t. Please refer to Internet Appendix for variable definitions.

In panel A, arbitrageurs are defined as those investors who conduct arbitrage trading at least twice. In panel B, arbitrageurs are defined as those investors who conduct arbitrage trading in their first year, and I examine their tradng behavior in the rest of the sample periods.

The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Panel A: Arbitrageurs who conduct at least twice arbitrage trading

Variables	Log (premium arbitrage flow)			Log (num of arbis buying B)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^{U}_{c,t-1}$	0.624	0.649	0.408	0.039	0.040	0.028
	(5.75)***	(5.89)***	(3.91)***	(2.89)***	(2.90)***	(2.08)**
$Mispricing_{t-1}$	0.102	0.277	0.224	-0.120	-0.113	-0.083
	(1.66)*	(3.11)***	(2.87)***	(-1.47)	(-1.62)	(-1.64)
$Other\,entry^{U}_{c,t-1}$		0.084	0.047		0.019	0.012
		(0.91)	(0.56)		(0.76)	(0.58)
$New\,entry^{U}_{c,t-1} * Mispricing_{t-1}$	0.063	0.052	0.048	-0.004	-0.004	-0.003
	(3.56)***	(3.00)***	(2.96)***	(-3.24)***	(-3.11)***	(-2.27)**
$Other\,entry^{U}_{c,t-1} * Mispricing_{t-1}$		-0.322	-0.258		0.001	-0.000
		(-3.27)***	(-2.84)***		(0.17)	(-0.06)
$Log(DV)_{c,t-1}$			0.282			0.279
			(13.37)***			(4.90)***
$Mispricing_{t}$			-0.084			-0.382
			(-0.66)			(-0.81)
Constant	0.094	-0.307	-0.226	0.029	0.006	0.007
	(1.69)*	(-1.34)	(-1.28)	(1.43)	(0.28)	(0.34)
CONTROLS	NO	NO	YES	NO	NO	YES
TIME FE	93,128	93,128	93,128	93,128	93,128	93,128
CITY FE	YES	YES	YES	YES	YES	YES
Observations	YES	YES	YES	YES	YES	YES
r2	0.089	0.09	0.163	0.066	0.066	0.138

Panel B: Ex-ante identified arbitrageurs

Variables	Log (premium arbitrage flow)			Log (num of arbis buying B)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^{U}_{c,t-1}$	0.265	0.282	0.076	0.026	0.034	0.026
	(3.06)***	(3.43)***	(1.07)	(2.62)***	(3.02)***	(3.02)***
$Mispricing_{t-1}$	0.004	0.006	-0.002	0.004	0.257	0.158
	(0.82)	(0.33)	(-0.15)	(0.11)	(1.49)	(1.21)
$Other\,entry^{U}_{c,t-1}$		2.071	0.837		0.196	0.146
		(1.97)*	-1.01		(1.65)*	(1.21)
$New\,entry^{U}_{c,t-1} * Mispricing_{t-1}$	0.001	0.001	0.001	-0.002	-0.004	-0.003
	(3.80)***	(4.53)***	(5.09)***	(-2.12)**	(-3.35)***	(-3.40)***
$Other\,entry^{U}_{c,t-1} * Mispricing_{t-1}$		0.003	0.003		-0.061	-0.035
		(0.49)	(0.69)		(-2.32)**	(-1.62)
$Log(DV)_{c,t-1}$			0.284			0.227
			(8.55)***			(5.64)***
$Mispricing_{t}$			0.018			0.417
			(1.27)			(2.66)***
Constant	00.002	-7.131	-4.157	-0.005	-0.658	-0.775
	(0.01)	(-2.99)***	(-1.87)*	(-0.27)	(-2.13)**	(-2.33)**
CONTROLS	NO	NO	YES	NO	NO	YES
TIME FE	38,901	38,901	38,901	38,901	38,901	38,901
CITY FE	YES	YES	YES	YES	YES	YES
Observations	YES	YES	YES	YES	YES	YES
r2	0.175	0.176	0.242	0.137	0.140	0.183

Table 13: Robustness Check (new definition of location)

This table provides the robustness check of the baseline regression in table 5 and examines the results of the following daily panel regressions with time and province fixed effects:

$$Arbitrage\,flow_{g,t}/Riding\,bubble_{g,t} = \alpha_0 + \alpha_1 * New\,entry^{U}_{g,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^{U}_{g,t-1} + \alpha_4 * New\,entry^{U}_{g,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^{U}_{g,t-1} * Mispricing_{t-1} + \alpha_6 * M_{g,t-1} + \epsilon_{g,t}$$

where the subscripts g and t refer to province and day, respectively. Province information if derived from the zip codes. $Arbitrage\,flow_{g,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in province g on day t; $Riding\,bubble_{g,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $New\,entry^{U}_{g,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B for province g on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Other\,entry^{U}_{g,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in provinces other than province g on day t-1. $Mispricing_{t-1}$ is the positive structured fund mispricing of M for fund f at t-1 which is calculated as the difference between synthetic trading price in the market and the NAV of M; here it represents the days when mispricing is positive (when there is exante potential bubbles). The vector $M_{g,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs buying B at t-1 for province g and mispricing for fund f at time t. Please refer to Internet Appendix for variable definitions.

The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Variables	Log (premium arbitrage flow)			Log (num of arbis buying B)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^{U}_{g,t-1}$	0.822	0.823	0.541	0.119	0.119	0.084
	(5.31)***	(5.29)***	(3.63)***	(6.22)***	(6.20)***	(5.19)***
$Mispricing_{t-1}$	0.066	0.012	-0.052	0.002	-0.004	-0.005
	(1.38)	(0.19)	(-0.88)	(1.10)	(-0.82)	(-1.23)
$Other\,entry^{U}_{g,t-1}$		-0.055	-0.043		-0.010	-0.007
		(-0.65)	(-0.55)		(-1.52)	(-1.18)
$New\,entry^{U}_{g,t-1} * Mispricing_{t-1}$	0.102	0.103	0.098	-0.011	-0.011	-0.009
	(5.26)***	(5.18)***	(5.42)***	(-4.62)***	(-4.54)***	(-5.42)***
$Other\,entry^{U}_{g,t-1} * Mispricing_{t-1}$		-0.022	-0.003		0.002	0.002
		(-1.23)	(-0.16)		(1.51)	(2.01)**
$Log(DV)_{g,t-1}$			0.234			0.268
			(9.69)***			(7.15)***
$Mispricing_{t}$			1.708			0.193
			(0.55)			(0.73)
Constant	0.362	0.443	0.385	-0.046	-0.032	-0.033
	(0.74)	(0.88)	(0.71)	(-2.61)***	(-1.58)	(-1.78)*
CONTROLS	NO	NO	YES	NO	NO	YES
TIME FE	10,680	10,680	10,680	10,680	10,680	10,680
CITY FE	YES	YES	YES	YES	YES	YES
Observations	YES	YES	YES	YES	YES	YES
r2	0.336	0.336	0.373	0.200	0.200	0.255

Table 14: Robustness Check (weekly regression)

This table provides the robustness check of the baseline regression in table 5 and examines the results of the following daily panel regressions with week and city fixed effects:

$$Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \alpha_0 + \alpha_1 * New\,entry^U_{c,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^U_{c,t-1} + \alpha_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_6 * M_{c,t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and week, respectively. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares “partitioned” from M for premium arbitrage in city c and week t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B in week t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c and week t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Other\,entry^U_{c,t-1}$ is the logarithm of number of new unsophisticated investors entering the stock market to buy B in week t-1; $Mispricing_{t-1}$ is the positive structured fund mispricing for fund f in week t-1 which is calculated as the difference between synthetic trading price in the market and the NAV of M; and vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow and the logarithm of the number of arbitrageurs who buy B at t-1 in city c and the average mispricing for fund f in week t-1. Please refer to Internet Appendix for variable definitions.

The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Variables	Log (premium arbitrage flow)			Log (num of arbis buying B)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$New\,entry^U_{c,t-1}$	0.323	0.328	0.203	0.182	0.181	0.176
	(2.75)***	(2.79)***	(1.72)*	(8.30)***	(8.28)***	(8.02)***
$Mispricing_{t-1}$	0.012	0.233	0.195	0.001	0.007	0.005
	(0.42)	(3.04)***	(2.59)***	(0.67)	(1.09)	(0.84)
$Other\,entry^U_{c,t-1}$		-0.013	-0.037		0.007	0.006
		(-0.37)	(-1.10)		(1.90)*	(1.66)*
$New\,entry^U_{c,t-1} * Mispricing_{t-1}$	0.177	0.181	0.155	-0.016	-0.016	-0.017
	(6.40)***	(6.57)***	(5.82)***	(-3.29)***	(-3.25)***	(-3.52)***
$Other\,entry^U_{c,t-1} * Mispricing_{t-1}$		-0.041	-0.035		-0.001	-0.001
		(-3.66)***	(-3.18)***		(-1.18)	(-0.96)
$Log(DV)_{c,t-1}$			0.168			0.007
			(7.19)***			(2.28)**
$Mispricing_t$			0.006			-0.115
			(0.78)			(-1.27)
Constant	0.914	0.952	0.890	-0.013	-0.026	-0.007
	(1.34)	(1.39)	(1.49)	(-0.16)	(-0.32)	(-0.11)
CONTROLS	NO	NO	YES	NO	NO	YES
TIME FE	116,98	16,981	16,981	16,981	16,981	16,981
CITY FE	YES	YES	YES	YES	YES	YES
Observations	YES	YES	YES	YES	YES	YES
r2	0.251	0.251	0.272	0.375	0.375	0.376

Table 15: Robustness Check (subsample of mispricng)

This table presents the baseline relationship between the participation of unsophisticated investors and arbitrageur behavior and panel A examines the regression model as in table 5 with different subsamples of structured fund mispricing, panel B examines the following regression with negative bubbles:

$$Arbitrage\,flow_{c,t}/Riding\,bubble_{c,t} = \alpha_0 + \alpha_1 * New\,entry^U_{c,t-1} + \alpha_2 * Mispricing_{t-1} + \alpha_3 * Other\,entry^U_{c,t-1} + \alpha_4 * New\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_5 * Other\,entry^U_{c,t-1} * Mispricing_{t-1} + \alpha_6 * M_{c,t-1} + \epsilon_{c,t}$$

where the subscripts c and t refer to city and day, respectively. $Arbitrage\,flow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in city c on day t; $Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B on day t; $New\,entry^U_{c,t-1}$ is the logarithm of the number of new unsophisticated investors entering the stock market to buy B in city c on day t-1. New unsophisticated investor is defined as the first time such an investor enters the market to trade. $Other\,entry^U_{c,t-1}$ is the logarithm of the total number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1. $Mispricing_{t-1}$ is the positive mispricing of a structured fund for fund f on day t-1, which is calculated as the difference between the synthetic trading price in the market and the NAV of M; here, it represents the days when mispricing is positive (when there are ex ante potential bubbles). The vector $M_{c,t-1}$ stacks a list of control variables, including the logarithm of the premium arbitrage flow or the logarithm of the number of arbitrageurs who buy B at t-1 and mispricing for fund f on day t. I include time and city fixed effects in all specifications. Time fixed effects remove the time trend, and city fixed effects control for the time-invariant unobservable heterogeneity across different cities. Please refer to the Internet Appendix for variable definitions.

The superscripts ***, **, and * refer to the 1%, 5%, and 10% levels of statistical significance, respectively. The sample period is from 2011 to 2015.

Variables	Log (premium arbitrage flow)		Log (num of arbis buying B)	
	All sample	Above 0.5%	All sample	Above 0.5%
$New\,entry^U_{c,t-1}$	0.455	0.076	0.053	0.031
	(4.28)***	(0.35)	(4.79)***	(2.71)***
$Mispricing_{t-1}$	0.238	0.582	-0.029	-0.056
	(2.69)***	(1.18)	(-1.77)*	(-0.92)
$Other\,entry^U_{c,t-1}$	0.077	0.768	0.035	0.085
	(0.82)	(1.62)	(2.21)**	(2.31)**
$New\,entry^U_{c,t-1} * Mispricing_{t-1}$	0.075	0.094	-0.005	-0.003
	(4.36)***	(3.66)***	(-4.30)***	(-2.37)**
$Other\,entry^U_{c,t-1} * Mispricing_{t-1}$	-0.286	-0.365	0.003	-0.021
	(-2.62)***	(-1.53)	(0.33)	(-1.43)
$Log(DV)_{c,t-1}$	0.273	0.242	0.237	0.175
	(14.57)***	(10.41)***	(11.03)***	(3.85)***
$Mispricing_t$	-0.084	-2.863	0.021	-0.372
	(-0.58)	(-1.33)	(1.46)	(-2.20)**
Constant	-0.325	1.360	-0.091	0.264
	(-1.68)*	(1.18)	(-2.36)**	(2.40)**
CITY FE	YES	YES	YES	YES
Observations	225,155	62,443	225,155	62,443
r2	0.174	0.209	0.176	0.123

Appendix A. Variable Definition

Appendix A.1. Mispricing

$Mispricing_{it}$ is the structured fund mispricing which is defined as
$Mispricing_{it} = \frac{(P^A_{it}+P^B_{it})-NAV^M_{it}}{NAV^M_{it}}$, where $P^A_{it} + P^B_{it}$ is the synthetic market price of M on the stock market based on the prices of A and B. The structured fund mispricing is consisted of two components: mispricing of A and mispricing of B.
$Mispricing^A_{it}$ is the mispricing of A which is defined as $Mispricing^A_{it} = \frac{P^A_{it}-NAV^A_{it}}{NAV^A_{it}}$, where $NAV^A_t = 1 + \frac{R*t}{365}$.
$Mispricing^B_{it}$ is the mispricing of B which is defined as $Mispricing^B_{it} = \frac{P^B_{it}-NAV^B_{it}}{NAV^B_{it}}$, where $NAV^B_t = NAV^M_t - NAV^A_t$.

Appendix A.2. Arbitrageur Trading Variable

$Arbitrageflow_{c,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in city c on day t.
$Arbitrageflow_{g,t}$ is the logarithm of the total number of shares "partitioned" from M for premium arbitrage in province g on day t.
$Riding\,bubble_{c,t}$ is the logarithm of the total number of arbitrageurs who buy B in city c on day t.
$Riding\,bubble_{g,t}$ is the logarithm of the total number of arbitrageurs who buy B in province g on day t.

Appendix A.3. Unsophisticated Investor Variable

$Newentry^U_{c,t-1}$ is the logarithm of number of new unsophisticated investors entering the stock market to buy B for city c and time t-1.
$Newentry^U_{g,t-1}$ is the logarithm of number of new unsophisticated investors entering the stock market to buy B for province g and time t-1.
$Other\,entry^U_{c,t}$ is the logarithm of number of new unsophisticated investors entering the stock market to buy B in cities other than city c on day t-1.
$Other\,entry^U_{g,t}$ is the logarithm of number of new unsophisticated investors entering the stock market to buy B in provinces other than province g on day t-1.
$Predicted\,new\,entry^U_{c,t}$ is the logarithm of the estimated new entry of unsophisticated investors, which is the fitted part of $\beta_1 Posreturnnum_(c, t-1)$ in the first stage regression.
$Hometown\,entry^U_{c,t}$ is the logarithm of the average number of new unsophisticated investors entering the stock market to buy B for the hometown cities of all arbitrageurs in city c on day t-1.
$Non\,inventory_{c,t}$ is the logarithm of the number of arbitrageurs who do not have inventory of B in their accounts in city c on day t-1.
$Repeat\,entry^U_{c,t}$ is the logarithm of the number of existing unsophisticated investors entering the stock market to buy B after the first trade in B in city c on day t.

$New\, entry^{U}_{f,t}$ is the total number of new entries of unsophisticated investors aggregated at fund level, which is due to social contagion effect among unsophisticated investors.
$Repeat\, entry^{U}_{f,t}$ is the total number of repeated entries of unsophisticated investors aggregated at fund level, which is due to the positive feedback trading effect for unsophisticated investor herself.
$Posreturn\, num^{U}_{c,t-1}$ is the number of investors who experience positive returns since their purchase last week in city c.
$Negreturn\, num^{U}_{c,t-1}$ is the number of investors who experience negative returns since their purchase last week in city c.
$Posreturn\, mean^{U}_{c,t-1}$ is the average positive returns by the investors last week in city c.
$Negreturnmean^{U}_{c,t-1}$ is the average negative returns by the investors last week in city c.

Appendix A.4. Control Variables

$M\, volatility$ is the std.Dev of past 5 days' net asset value of M.
$A\, volatility$ is the std.Dev of past 5 days' return of A.
$B\, volatility$ is the std.Dev of past 5 days' return of B.
$A\, volume$ is the logarithm of trading volume of A at t-1.
$B\, volume$ is the logarithm of trading volume of B at t-1.
$Arbitragedeals$ is the number of arbitrage deals conducted by investors.
$Arbitrageflows$ is the total number of shares that are "partitioned" or "synthesized" aggregated from all arbitrage deals on each trading day.
$Returnmax$ is max (A daily return, B daily return).
$Returnmin$ is min (A daily return, B daily return).
$Largereturnmax$ is max (Return max -5%, 0).
$Volatilitymax$ is max (A Std.Dev of past 5 days' return, B Std.Dev of past 5 days' return).
$Turnovermax$ is max (A turnover, B turnover).
$Turnovermin$ is min (A turnover, B turnover).
$Closingprice_A$ is the closing price of A for each fund at day t-1.
$Closingprice_B$ is the closing price of B for each fund at day t-1.
$Returnlag1_{i,f,t}$ is the return on the most recent transaction cycle of investor i in B of fund i before date t.
$Returnlag2_{i,f,t}$ is the average return of the transaction cycles of investor i in B of fund i prior to the first most recent transaction cycle before date t.

Appendix B. Empirical Evidence on Arbitrage Activities and Mispricing

Table B.16: VAR Model of Arbitrage Deals and Mispricing

This table test the following Vector Autoregressive model:

$$\begin{bmatrix} Mispricing_{1,t} \\ ArbitrageDeals_{2,t} or ArbitrageFlows_{2,t} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} + \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} * \begin{bmatrix} Mispricing_{1,t-1} \\ ArbitrageDeals_{2,t-1} or ArbitrageFlows_{2,t} \end{bmatrix} + \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix} * \begin{bmatrix} Mispricing_{1,t-2} \\ ArbitrageDeals_{2,t-2} or ArbitrageFlows_{2,t} \end{bmatrix} + \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix} * \begin{bmatrix} Mispricing_{1,t-3} \\ ArbitrageDeals_{2,t-3} or ArbitrageFlows_{2,t} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

with daily mispricing and the number of arbitrage deals or arbitrage flows as dependent variables and past 3 days' mispricing and deals as independent variables. Mispricing is the absolute value of the structured fund mispricing. Arbitrage flows are scaled by total number of shares in the stock market for fund A and B. (t-statistics are in parentheses). I use ***, **, and * to denote significance at the 1%, 5%, and 10% level, respectively.

Panel A: Mispricing and Arbitrage Deals				
	Equity Fund		Bond Fund	
Variables	M mispricing	Deals	M mispricing	Deals
$L.mispricing$	0.899	95.330	0.605	7.776
	(16.82)***	(4.15)***	(11.91)***	(0.55)
$L2.mispricing$	0.084	-28.020	0.117	-19.866
	(3.89)***	(-3.01)***	(2.34)**	(-1.44)
$L3.mispricing$	0.020	19.440	0.132	6.700
	(0.92)	(2.10)**	(3.64)***	(0.67)
$L.deals$	-0.000	0.679	0.000	0.503
	(-4.03)***	(14.13)***	(0.99)	(10.47)***
$L2.deals$	0.000	-0.092	-0.000	0.098
	(1.89)*	(-1.59)	(-1.50)	(2.14)**
$L3.deals$	-0.000	0.063	-0.000	0.172
	(-0.42)	(1.30)	(-1.61)	(4.04)***
Constant	0.001	-0.333	0.002	0.304
	(0.96)	(-0.97)	(3.80)***	(1.74)*
Observations	264	264	405	405
$chi2$	462.3	462.3	1068	1068
Panel B: Mispricing and Arbitrage Flows				
Variables	M mispricing	Flows	M mispricing	Flows
$L.mispricing$	0.944	0.439	1.064	0.056
	(17.94)***	(2.20)**	(19.64)***	(1.17)
$L2.mispricing$	0.055	0.129	-0.120	0.024
	(3.20)***	(1.98)**	(-1.73)*	(0.40)
$L3.mispricing$	0.042	0.089	0.081	0.009
	(2.41)**	(1.33)	(1.61)	(0.19)
$L.flow$	-0.056	0.475	0.130	0.185
	(-5.01)***	(11.27)***	(1.17)	(1.89)*
$L2.flow$	0.014	0.019	0.001	0.315
	(1.14)	(0.40)	(0.01)	(3.23)***
$L3.flow$	-0.007	-0.001	0.044	0.029
	(-0.67)	(-0.03)	(0.49)	(0.37)
Constant	0.000	-0.002	-0.000	-0.000
	(0.49)	(-0.73)	(-0.06)	(-0.66)
Observations	264	264	356	356
chi2	506.5	506.5	1379	1379

Appendix C. Empirical Evidence on Limits to Arbitrage: Market Risk

Table C.17: Market Volatility Risk on Probability of Arbitrage

This table reports the results of the maximum likelihood estimates in the logit regression:

$$
\begin{aligned}
&Logit(P)_{f,i,t} = \alpha + \beta_1 * (Mispricing)_{f,t-1} + \beta_2 * (Largereturnmax)_{f,t-1} + \beta_3 * (Largereturnmax)_{f,t-2} \\
&+ \beta_4 * (Largereturnmax)_{f,t-3} + \beta_5 * (Volatilitymax)_{f,t-1} + \beta_6 * Largereturnmax_{f,t-1} * Mispricing_{f,t-1} + \\
&\beta_7 * Largereturnmax_{f,t-2} * Mispricing_{f,t-1} + \beta_8 * Largereturnmax_{f,t-3} * Mispricing_{f,t-1} \\
&+ \beta_9 * Volatilitymax * Mispricing_{f,t-1} + Controls
\end{aligned}
$$

where f, i, t refers to fund-investor-time account, $Mispricing_{f,t-1}$ is the structured fund mispricingin fund f at time t-1, $Largereturnmax_{f,t-1}$ is the max(Return max- 5%, 0), $Volatilitymax_{f,t}$ is the max(standard deviation of past 5 days' return, B standard deviation of past 5 days' return). Control variables include: $(Returnmax)_{f,t-1,t-2,t-3}$ is the max(A daily return, B daily return) for day t-1, t-2 and t-3 respectively. $(Returnmin)_{f,t-1,t-2,t-3}$ is the min(A daily return, B daily return) for day t-1,t-2,t-3 respectively. $(Turnovermax)_{f,t-1,t-2,t-3}$ is the max(A turnover, B turnover) for day t-1,t-2,t-3. $(Turnovermin)_{f,t-1,t-2,t-3}$ is the min(A turnover, B turnover) for day t-1,t-2,t-3 respectively. I use ***, **, and * to denote significance at the 1%, 5%, and 10% level, respectively.

Panel A: Dependent Variable =1 if an investor conducts premium arbitrage at day t

	Individual Investor			Institutional Investor		
Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$Mispricing_{t-1}$	48.783	37.272	38.349	56.169	278.847	278.881
	(19.11)***	(6.76)***	(7.15)***	(1.67)*	(2.07)**	(2.07)**
$Largereturnmax_{t-1}$	0.262	-0.065	-0.028	0.074	1.785	1.785
	(5.05)***	(-0.58)	(-0.26)	-0.23	-1.51	-1.51
$Largereturnmax_{t-2}$	0.083	0.183	0.205	0.31	0.764	0.764
	(1.16)	(1.59)	(1.76)*	-0.69	-0.82	-0.82
$Largereturnmax_{t-3}$	0.338	-0.162	-0.129	0.205	1.592	1.592
	(6.42)***	(-1.29)	(-1.01)	-0.41	-1.55	-1.55
$Volatilitymax$	0.103	0.081	0.051	0.252	1.129	1.128
	(3.07)***	(1.17)	(0.69)	-0.69	-1.1	-1.09
$Largereturnmax_{t-1} * Mispricing_{t-1}$	-1.301	-0.283	-0.117	5.139	-19.39	-19.391
	(-2.31)**	(-0.33)	(-0.13)	-0.81	(-2.10)**	(-2.10)**
$Largereturnmax_{t-2} * Mispricing_{t-1}$	-2.299	-3.863	-4.138	-8.358	-13.708	-13.71
	(-3.31)***	(-4.26)***	(-4.63)***	(-1.46)	(-1.43)	(-1.43)
$Largereturnmax_{t-3} * Mispricing_{t-1}$	-5.405	-2.286	-2.494	-5.595	-16.6	-16.603
	(-9.43)***	(-2.74)***	(-2.95)***	(-0.77)	(-2.71)***	(-2.71)***
$Volatilitymax * Mispricing_{t-1}$	-3.741	-2.636	-2.896	-5.409	-17.93	-17.932
	(-7.25)***	(-3.19)***	(-3.51)***	(-0.56)	(-2.04)**	(-2.04)**
Intercept	-8.726	-9.061	-7.870	-7.563	-107.042	-82.764
	(-15.09)***	(-7.90)***	(-6.12)***	(-5.70)***	(-1.30)	(-1.55)
Controls	NO	YES	YES	NO	YES	YES
Fund Fixed Effect	NO	NO	YES	NO	NO	YES
Time Fixed Effects	YES	YES	YES	YES	YES	YES
Observations	655,424	379,753	379,753	1,716	1,632	1,632
R2	0.191	0.176	0.177	0.131	0.18	0.18

Panel B: Dependent Variable =1 if an investor conducts discount arbitrage at day t

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$Mispricing_{t-1}$	-35.741	-51.128	-53.763 -	68.638	-297.027	-402.541
	(-4.02)***	(-4.29)***	(-4.48)***	(-0.80)	(-1.94)*	(-2.58)***
$Volatilitymax$	0.028	0.038	0.049	0.502	0.843	1.104
	(0.82)	(0.71)	(0.92)	(4.23)***	(1.99)**	(2.62)***
$Volatilitymax * Mispricing_{t-1}$	-0.263	4.165	4.740	36.737	124.914	152.111
	(-0.10)	(1.14)	(1.29)	(1.28)	(1.62)	(1.95)*
Intercept	-8.701	-8.521	-8.070	-8.556	-9.848	-5.110
	(-14.97)***	(-13.42)***	(-11.28)***	(-13.40)***	(-4.48)***	(-1.30)
Controls	NO	YES	YES	NO	YES	YES
Fund Fixed Effect	NO	NO	YES	NO	NO	YES
Time Fixed Effects	YES	YES	YES	NO	NO	NO
Observations	1,286,835	746,515	746,515	15,098	10,802	10,802
R2	0.0486	0.0468	0.0469	0.0126	0.102	0.177

Appendix D. Empirical Evidence on Positive Feedback Trading

Table D.18: Positive Feedback Trading among Unsophisticated Investors

The proportional hazards model specifies that $\lambda_{i,f,t}(\tau)$ is the hazard function of starting a new transaction cycle for existing investors i in fund f on day t, τ trading days after the end of the investor's last transaction cycle, takes the form

$$\lambda_{i,f,t}(\tau) = \lambda(\tau) * e^{x_{i,f,t}*\beta}$$

where $\lambda(\tau)$ is the baseline hazard rate and $x_{i,f,t}$ is a vector of covariates that proportionally shift the baseline hazard. For investors who have previously completed one transaction cycles $x_{i,f,t} * \beta$ is given by

$$x_{i,f,t} * \beta = a1 * Returnlag1_{i,f,t} + b1 * (Returnlag1_{i,f,t} > 0) + c1 * Returnlag2_{i,f,t} + d1 * (Returnlag2_{i,f,t} > 0) + controls + u1_{i,f,t}$$

where $Returnlag1_{i,f,t}$ is the return on the most recent transaction cycle of investor i in B of fund i before date t. The dummy variable $Returnlag1_{i,f,t} > 0$ takes value one if $Returnlag1_{i,f,t} > 0$ and zero otherwise. I also control for fund and time fixed effects.

This regression explain the re-entries of unsophisticated investors who have previously traded structured fund using the investors' previous transaction cycle returns for three groups of investors. The three groups are those who have previously completed one, two, and three or more transactions in B. The unit of observation is an investor-fund-date, and for each investor, the left-hand side variable takes value one if investor i begins a new transaction cycle on date t, and zero otherwise. The main explanatory variables are $Returnlag1_{i,f,t}$ and the dummy variable $I(Returnlag1_{i,f,t} > 0)$. $Returnlag2_{i,f,t}$ is the average return of the transaction cycles of investor i in fund f prior to the first most recent transaction cycle before date t. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Dependent variable is equal to one if investor begins a new transaction cycle				
	One-cycle investor		Two-cycle investor	
Variables	Coefficient	P-value	Coefficient	P-value
$Returnlag1_{i,f,t}$	0.120	0.120	-0.178	0.637
$Returnlag2_{i,f,t}$			-0.485	0.226
$I(Returnlag1_{i,f,t} > 0)$	0.391	$< .0001$	0.201	$< .0001$
$I(Returnlag2_{i,f,t} > 0)$			0.120	0.0051
$FundReturn_{f,t-1}$	3.779	$< .0001$	0.815	0.083
$FundReturn_{f,t-2}$			-0.899	0.019
$Turnover_{f,t-1}$	0.292	$< .0001$	0.154	$< .0001$
$Turnover_{f,t-2}$			-0.005	0.863
Observations	1402226		405494	
TIME FE	Yes		Yes	

Appendix E. Risk-free Asset A

Table E.19: Risk-free Asset A: an example

In the structured fund,investors can invest in a base asset, a fund labeled M, which is almost identical to a standard open-end mutual fund, except that investors can choose to convert M into two structured assets at a pre-determined conversion ratio – a fixed-income asset (called asset A) and a levered-equity asset (called asset B). Both A and B are traded on the stock exchange and can be converted back to M by investors. Their net asset values (NAVs) are calculated from the NAV of M and announced by the fund family on a daily basis, but their trading prices often deviate from their NAVs, creating the classical scenario of mispricing in which trading price of an asset deviates from its fundamental value. The policy of the structured funds ensures that A is almost a risk-free security. Here is a detailed example from the website of China Securities Regulatory Commission (CSRC).

A structured fund with conversion ratio 1:1 for A and B (one A plus one B can be converted to two M). Jennifer is an investor with 10,000 shares of M, A, B in account. The net asset values of M, A, B are 0.661, 1.076, 0.246 separately. When the net asset value of B is lower than 0.25 (which is the lower bound), the mutual fund family would activate a policy that ensures the safety of A share in the following chart.

The mechanism works like this: on a date t, net asset value of B is lower than 0.25 which is the lower bound. Mutual fund family would activate the policy to ensure the safety of A asset. Net asset value of M, A, B would be set back to the originally value of 1 and the number of shares would change accordingly in each investor's account (note that the total assets remain the same for M, A, B in Jennifer's account). She gets 8,300 number of M shares that she can redeem from the mutual fund to get cash. This policy ensures both the interests and principle of A investors.

	Before		After	
Asset	NAV	Num of shares	NAV	Num of shares
M	0.661	10,000	1.000	6610 shares of M
A	1.076	10,000	1.000	2460 shares of A+8300 shares of M
B	0.246	10,000	1.000	2460 shares of B