2012

Two essays on information in trading

Yanhao Fang
Louisiana State University and Agricultural and Mechanical College

TWO ESSAYS ON INFORMATION IN TRADING

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agriculture and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Interdepartmental Program in Business Administration
(Finance)

by
Yanhao Fang
B.A., Fudan University, 2002
B.A., McGill University, 2005
M.A., McGill University, 2006
August 2012

ACKNOWLEDGMENTS

A journey of almost six years is about to come to an end since I started my doctoral studies at LSU. There are so many ups and downs in my personal life to look back and remember. I could not have been through those difficult times without being surrounded by my family and friends. My dad, FANG Leijie, and mom, ZHAN Mengying, have been unconditionally standing by my side and providing mental and financial support for every step I take. They give me all they have. My old friends share my pain and joy whenever I need them, listening to me with a heart and guiding me with their wise and kind words. Many of them live far away from me but we constantly talk to each other on phones and messages. It is a bless that I know I can count on them for the rest of my lifetime.

It is not easy to overcome many obstacles here and there during my time at LSU. Fortunately, my professors and colleagues are the best of the class. I would like to give special thanks to Professor Gary C. Sanger, who serves as my advisor and dissertation committee chair. He is an amazing researcher who knows the answers to every question I have for my papers. I thank Professor LI Wei, who was my advisor at one time but left LSU in my fourth year. He was willing to build my knowledge of theoretical modeling from ground zero. I thank Professor Jimmy E. Hilliard for enrolling me into the program and Professor Lin for advising me on many aspects of research. Professor HE Shan is on my dissertation committee and a mentor on my teaching assignments. She would unselfishly share her experience and best advice on teaching and being a teacher and it has always been a pleasure talking to her about research. I would like to thank Professor V. Carlos Slawson, the chair of Department of Finance. Facing budgetary shortfalls, he has managed to secure as much financial support as possible for every student in the doctoral program including senior students like me.

I always feel like that SUN Ping-wen (Steven) and ZHU Shuang (Jennifer), mean more of friends to me than of colleagues. We entered the program about the same time and are all leaving this year. We talk more about our lives than about our research. They are the best friends I made here. And I am also proud to be a member of doctoral students group at LSU. Everyone is nice and kind and ready to give me a hand if they are asked to.

Last, I reserve my deepest gratitude for someone who is very special to me at the moment. I am holding my hope high for you, if not for you only.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation comprises of two essays. The first essay, 'Index Price Discovery in the Cash Market', employs Hasbrouck's (2003) information share method to analyze the flow of information in equity markets. In particular I compare trading in Index ETFs with that of their underlying securities. Surprisingly, ETFs seem to play a significant role in the price discovery process, rather than serving as passive indexing/hedging vehicles. Using TAQ data I reconstruct the second-by-second intraday price series for the S&P 500 using its component stocks. Results show that the ETF contributes almost half of the price formation of S&P 500 in the spot market. Next, comparing trades and quotes, We determine the relative amount of informed trading (versus noise) in the ETF. When the trading of ETFs is driven more by private information, the ETF contributes more to the price discovery in the S&P 500 spot market. Thus, the evidence suggests that a portion of ETF trading is information motivated, similar to individual stocks in the index.

The second essay, 'Institutional Flow and Information', studies the dynamics of institutions' trading behavior and the information of the market using a high-frequency dataset. The daily institutional flows are estimated in a linear model which differentiates trades by their size. I find in a structural *Vector Autoregression* (*VAR*) system that the amount of trading that is private-information based has a negative contemporaneous relationship with the daily changes of institutional ownership in five of the six stocks studied. The findings prove that the trading pattern of institutional investors is related to the degree of information asymmetry in the market. In addition, there is an observable difference of institutions trading ETF or S&P 500 component stocks.

CHAPTER 1: INTRODUCTION

This dissertation is comprised of two essays. The first essay, 'Index price discovery in the cash market', probes the information dissemination process of the trading on ETF and index component stocks and its implications on the index. The second essay, 'Institutional trading and information', examines the role of institutional investors in the trading of ETF and index component stocks. A unified theme underlying the two essays is a study on old and new channels of information diffusion via different assets and market players.

Ideas of the first essay came in the wake of flash crash of the stock market in May 2010. One of the public opinion was that error in program trading of ETF brought down the price stability protection causing a market wide panic. This event intensifies the common interests in ETF as it continues to grow very fast in size and gain more popularity among investors. Research wise, it leads to a much broader topic that focuses on how the trading shaped up by current technology determines the information flow from the microstructural perspective. The objective is to reveal that information is revealed in the trading of ETF, which was considered in the past as a passive investment financial vehicle, and how it is factored into the price level of market indices, which ETF is designed to replicate.

Empirical results of the first essay show that ETF has redefined itself as an important of price discovery process of the underlying index, which give rise to the second essay focusing on institutional investors who are the major players in ETF trading. However, the discussion is extended to the general study of institutional investors' trading behavior. To what extent their trading is information based has direct implications on explaining the phenomenon found in the first essay and other observations.

Both essays deal with the set of core questions concerning high frequency trading and how it affects the information dissemination mechanism. They also share the use of same data resources, econometric models, and empirical designs. Nevertheless, each essay has its own focus and approach to the problems and supplements the other in a unified research framework.

CHAPTER 2: INDEX PRICE DISCOVERY IN THE CASH MARKET

2.1 Introduction

The rapidly growing ETF market has much advanced the way people invest. Uninformed traders are motivated to trade basket security rather than individual securities to avoid being taken advantage of by informed traders. With the advent of ETF, not only institutional investors but also individuals can easily trade indexes, which are baskets of securities. Investors flock to ETF for the same reason they used to embrace index futures: less adverse selection costs. Gorton and Pennacchi (1993) point out that a basket as a composite security has lower rate of return variance because of lower level of information asymmetries.

The advantages of ETF over futures are straightforward. Investors can easily trade the entire portfolio with no capital restriction or the complicated impacts of expiration of futures contracts. Ironically, ETF is criticized for the same reason index futures were. Both of them arguably destabilize prices by encouraging irrational speculation. That is, the proportion of noise trading would rise. Specifically, ETF was blamed for the flash crash in May 2010.

Recent trends show that ETF is no longer considered a tool merely for diversification or indexing. Hasbrouck (2003) reveals that the trading of ETF affects the short-run price dynamics in the U.S. equity index markets comprising index futures and cash markets. Although the impact is of second order compared to that of futures trading, it is no doubt that its importance is on the rising.

In this paper, We propose a method to reconstruct the intraday series of the index portfolio trading from its component stocks. The reconstructed data allows us to study the price dynamics of the entire spot market for equity index trading. If ETF is used simply for indexing, it would not contribute much to the price formation. Results show that ETF contributes on average 45% of the total information share of the underlying random walk process of S&P 500 against the synthesized index portfolio.

The second contribution of this paper is to relate the price discovery to the information contents of ETF trading. Following Hasbrouck (1991b), We decompose the price movements of ETF into private information based and public information based. If ETF is used simply for indexing, its information composition should be fairly uniform and not be related to the price formation of the index. Empirically We find the proportion of private information based trading is at the same level in ETF as in component stocks. Moreover, it is the private information based trading that drives up ETF's contribution to the price formation of S&P 500. To summarize, there is strong evidence supporting the hypothesis that ETF is an important source of information and price discovery.

The sections are arranged as followed. Section 2.2 reviews past literature. Section 2.3 details data sources and econometric methodology. In Section 2.4, We analyze the price discovery process among ETF trading and the entire index stocks trading for Standard & Poor 500. Section 2.5 focuses on the information nature of ETF trading. Robustness checks and extensions are conducted in Section 2.6. The last section concludes.

2.2 Related Work

Whether the trading of a basket of securities facilitates market efficiency has always been intriguing and interesting. Without index ETF and futures, people would devise trading strategies to trade a basket. Vijh (1994) points out that there are two possible effects of trading security basket. The positive impact would be that the nonsynchronous trading is reduced, resulting in more liquidity and accurate information. The

negative impact would be increased price pressure or excess volatility, which could disrupt the capital market. He finds that the ever popular use of trading the S&P 500 stocks as a basket generates high daily variance of the NYSE and AMEX stocks even though nonsynchroneity is decreased.

Subrahmanyam (1991) discusses the basket trading from the perspective of information revelation in his theoretical model. He proves that, conditional on sufficiently small cross-sectional variation in basket weights across securities, the trading of a basket may increase the overall informativeness of the price of the underlying portfolio. Both the portfolio price and component stock prices respond more to the systematic information and less to the security-specific information.

In recent years equity market has seen sharp increase in trading activity, characterized by higher trading volume or turnover. Chordia, Roll, and Subrahmanyam (2010) document the trend and the accompanying improvement of market efficiency. They find that intraday volatility has decreased and prices behave more closely to random walk. This trend makes the intraday data better suited for studying issues like price formation than ever before. One that benefits this paper is that it allows more accurate construction of intraday price of index portfolio. For example, if there is trading in every second for each stock in the S&P 500, we would not have to rely on delayed trade prices to compute the weighted average and the resulting index price would be fresh as up to the second. It also helps to validate the information share approach that assumes a common random walk for different price series. A unified random walk process warrants attributing information contribution between ETF and index portfolio a natural solution to studying the dynamics between these prices.

The econometric approach used in this paper is developed in Hasbrouck (1995). The approach relies on the assumption that multiple price series share a common component and therefore are cointegrated. Arbitrage prevents the differences between price series diverging without bound. The common component underlying ETF and the index portfolio is the implicit efficient price of the index. Hasbrouck defines the efficient price as the random-walk component underlying all considered price vectors. The intensity of information of the efficient price can be measured with the innovation variance. For multiple price series, the information share of each series is its proportional contribution to the innovation variance of the common efficient price.

The identification of a common random walk process is critical for the development of information share approach. A broader class is the so-called permanent-transitory decomposition. The permanent component of the decomposition is not necessarily a random walk, which means its variance would equal to the long-run variance of the underlying price. Hasbrouck (2002) compares the permanent-transitory decomposition approach proposed by Gonzalo and Granger (1995) and his own information share approach which is a special case of the former. He suggests that it is hard to infer from the permanent-transitory decomposition because it does not render a clear interpretation of efficient price.

Existing literature shows that there is a lead-lag relationship between futures and sport markets. Kawaller, Koch and Koch (1987) find the impacts of futures price movements on index lasts 20 to 45 minutes while the impacts of the other direction are arealy beyond 1 minute. Chan (1992) finds that the futures asymmetrically lead the cash index and all of its component stocks in price movements. He also shows that, when the market is driven by macroeconomic information, the futures have a larger lead. Therefore, futures market is a preferred location for investors to impound the index with market-wide news.

A stronger dependence in both directions exists in the volatility of price movements according to Chan, Chan and Karolyi (1991). They find that price innovations originating in either cash or futures markets can predict the future volatility in the other market, indicating that both markets are important sources of price discovery. Chordia, Sarkar and Subrahmanyam (2005) study the inter-market dynamics between stock and

bond market. They find a high correlation of innovation shocks of the two markets and interpret it as evidence of common factors driving both liquidity and volatility. Moreover, volatility shocks lead the changes in liquidity. Hasbrouck (2003) finds that most of the price discovery of the index comes from the index futures market through E-mini trading.

In this paper, We only consider the spot markets. Focusing solely on the spot market eliminates some of the concerns that the structural differences between spot and futures market could affect the price formation process in a way that is unrelated to the information dissemination. ETF and index component stocks are both traded on spot equity market. They are more likely to respond to the same set of information.

Measuring how the information is discovered using the approach is still a valid method even though futures market is excluded from analysis. The approach does not specify how much total information is impounded in prices. In this paper, it measures the relative contribution of the ETF and index portfolio price series to their common underlying price. The interpretation of the approach is straightforward. If a price vector is found to have a larger share, it is the first to incorporate new information in the underlying price. However, it does not imply anything beyond the order of responses.

2.3 Data Sources and Methodology

TAQ database provides the intraday trading data from all ETF and stocks in the indexes studied. The lists of index stocks are obtained from iShares website, which publishes monthly composition of their ETF tracking S&P 500 and other indexes. The composition list might not be identical to the actual list of component stocks in the index but it should be very close. Trading volume and market value of individual stocks are from CRSP database. The trading volume and market cap of ETF are only available for S&P 500 from Yahoo Finance and from YCharts, an online data provider.

The econometric model on which the information share approach is based is a vector error correction model (*VECM*), which is ideal for studying the dynamics within a joint system of prices. The general form of *VECM* is

$$\Delta p_t = \emptyset_1 \Delta p_{t-1} + \emptyset_2 \Delta p_{t-2} + \cdots + \gamma(z_{t-1} - \beta) + \varepsilon_t \quad (2.1)$$

where p_t is a vector of prices with dimension of $n \times 1$. β in the error correction term is the target of the adjustment process, i.e., the long-term value of the difference between the prices. z_{t-1} is an n - 1 column vector of the difference between the first price and the remaining prices.

$$z_t = [(p_{1t} - p_{2t})(p_{1t} - p_{3t}) \cdots (p_{1t} - p_{nt})]' \quad (2.2)$$

All prices p_t in the system share the same random-walk $p_t = m_t + s_t$ where $m_t = m_{t-1} + \omega_t$. Economically, the random walk component is the efficient price. The innovations are linear in disturbances, $\omega_t = Au_t$. If the disturbances are uncorrelated, the innovation variance can be decomposed into components explained by innovations of each price, $\sigma_\omega^2 = \sum d_i^2$, where d_i^2 is the absolute contribution of the ith price's innovations to the common efficient price. In a system of multiple prices, the relative contribution, d_i^2/σ_ω^2, is price i's information share.

If the innovations are correlated, then there is no clean decomposition of the long-run variance. However, by rotating the order of innovations in the *VECM*, we can minimize or maximize the contribution of an

innovation to the total variance, thereby setting the upper and lower bounds for the information share of each price vector in the system. The bound can be tightened by using a fine time interval.

It is difficult to interpret the coefficients estimated in *VECM*. Following Hasbrouck (2003), in addition to information shares, We use impulse response functions to describe how each price responds to the initial shock to one of the price within the system. The impulse response function is usually used to forecast the effects of a hypothetical innovation. It sets zero to all prior innovations and measure the effects of the innovation at time zero after certain periods have elapsed:

$$\varphi_s(\varepsilon_0) = E^*\{p_s|\varepsilon_0, \varepsilon_{-1} = \varepsilon_{-20} = \cdots = 0\} \quad \text{for } s \geq 0 \tag{2.3}$$

To discuss the price impacts, it is more useful to measure the cumulative impulse response function:

$$\Psi_s(\varepsilon_0) = \sum_{k=0}^{s} \varphi_s(\varepsilon_0) \tag{2.4}$$

There is little use of the actual trading of component stocks in the existing literature, possibly because the reconstruction of the intraday series of S&P 500 is a challenge. Mackinlay and Ramaswamy (1988) use the one-minute updated index quotes updated provided by Chicago Mercantile Exchange. The index quotes are actually constructed from latest trade prices of component stocks. The one-minute frequency is not small enough to avoid the sale prices in those thinly traded stocks.

To improve the intraday series, We adopt the methodology in Hasbrouck (2003) to fill the intraday price of index to the unit of seconds with last-sale price. If there is no trading at a certain second, the price would be assigned to the last trading price. We repeat this for each stock in the index and generate a value-weighted S&P 500 price. Theoretically, this synthetic S&P 500 index (*SYN*) should have had the exact same random walk as the one that is underlying the two ETF if there was trading for each component stock in the index in every second. Even though empirically it is not true, and as it would be shown in the next section, there are differences in random walk processes between ETF and the synthesized index, the information share approach can still be applied to study the price dynamics between them because one can never deviate from the other price.

2.4 Price Discovery in the S&P 500

Since its inception in 1993, the *SPDR (SPY)*, which is the first and largest ETF, has seen its market value almost one thousand times larger from 96 million to 90 billion dollars. *IVV*, the iShares ETF which also tracks S&P 500, debuted May 2000 with a market cap of 550 million dollars. As of today, it has assets of over 25 billion assets in total. Meanwhile, the market value of the entire S&P 500 increases from 313 billion in 1993 to 11 trillion dollars in 2011.

The monthly average of daily trading volume also significantly increases for both ETF and the sum of all S&P 500 component stocks between September 2006 and December 2009. The daily number of traded shares has grown from 0.86 million to 3.8 million for *IVV*, from 68 million to 161 million for *SPY*, and from 2.45 billion to 4.39 billion for all S&P 500 component stocks summed.

The initial change in one series and the response of others would appear to be contemporaneous even though they occur sequentially. Therefore, the decomposition of innovation variance cannot be uniquely determined since ETF and index portfolio prices are updated simultaneously. However, the upper and lower bounds can be set. Moreover, the finer the time interval of the price series, the tighter the bounds is. Using a

finer interval would minimize the correlation although the correlation cannot be completely eliminated. In this paper, We estimate the model with a 1-second sampling interval, with 5-minutes lags, following the same practice as in Hasbrouck (2003).

The cumulative impulse response functions are drawn in Figure 2.1. Each graph depicts the responses of all three price vectors to an initial shock in the security indicated up to 10 minutes. Each point on the graph is an average of daily estimates over all fourth quarters from 2006 to 2009. The graphs clearly show that only two ETF price impacts converge, which is not surprising considering that they have the same underlying series. The real index is calculated using an index divisor to keep the index comparable throughout times when the number of outstanding shares changes. The synthesized index is simply a weighted average of market values of all component stocks. However, all the estimation is based on one-day data and the index divisor is constant at least for any particular day. Thus, it is unnecessary to adjust the intraday price of the index. The graphs show that the synthesized index price does not converge to the two ETF, implying that the random walk processes of the synthesized index and two ETF are not 100 percent identical. But it is easy to see that the responses of the synthesized index are locked in step along with one ETF to the initial shock to the other ETF. So the two prices, one represented by the two ETF and one by the synthesized, are still cointegrated.

The graphs selected in Figure 2.1 show all three cumulative impulse response functions to the shock to SPDR in all fourth quarters. The time for *IVV* and *SPY* converging to each other is greatly reduced. In the fourth quarter of 2006, converging takes almost 4 minutes to finish. In the fourth quarter of 2009, it takes only 1 minute and 20 seconds. The reduction in converging time is more prominent when the shock is to *IVV*. In the fourth quarter of 2006, *SPY* and *IVV* still do not completely converge in the 10-minutes span. While in the fourth quarter of 2009, the process is done between 2 and 3 minutes. This improvement is helped by the more frequent trading in recent years, which eliminates arbitrage opportunities between two ETF faster. The graphs showing the impulse response functions to the shock to the synthetic index reveal the same trend that convergence between two ETF takes less time.

Price impacts to own shocks for the two ETF are greatest at the time and then quickly die out while price impacts to own shocks for the index portfolio are the smallest at the initial, jump to a much higher level and then settle at slightly lower, which is still a lot higher than the initial level. The distinction indicates that the nature of shocks to ETF and the index are very different. Shocks to the ETF are quickly reflected in all price series and the permanent effect is smaller than the initial impact. In contrast, markets at the beginning underreact to the shocks to the index portfolio. It's easy to understand the different patterns if we think that the shock to ETF is more public news related and easily absorbed by the market, which makes it 'transitory'. The shock to the index portfolio actually contains shocks to all component stocks, thus rendering the market hard time to digest all in a short time and creating a 'lasting' effect.

Figure 2.2 depicts the time series of long-run cumulative response functions. The long-run value is taken at the end of the 10-minutes iteration time window when convergence can be guaranteed. The long-run responses of the two ETF are exactly the same, both represented by the red plots. The blue series plots the responses of *SYN*. As discussed before, ETF and *SYN* still do not converge to the same level, although the distance between the two long-run levels is constant throughout the period. The three graphs represent the source of shocks respectively from *SPY*, *IVV* and *SYN*. The long-run response to the shocks have become smaller when the source of the innovation is *SPY* or *SYN* or become bigger when the source is *IVV*. To interpret, *IVV* has gained larger price impacts while *SPY* and *SYN* losing their impacts. From the sample means not shown, the innovations of *SYN* have the largest impacts, followed by *SPY* and *IVV*, although the gap is narrowing.

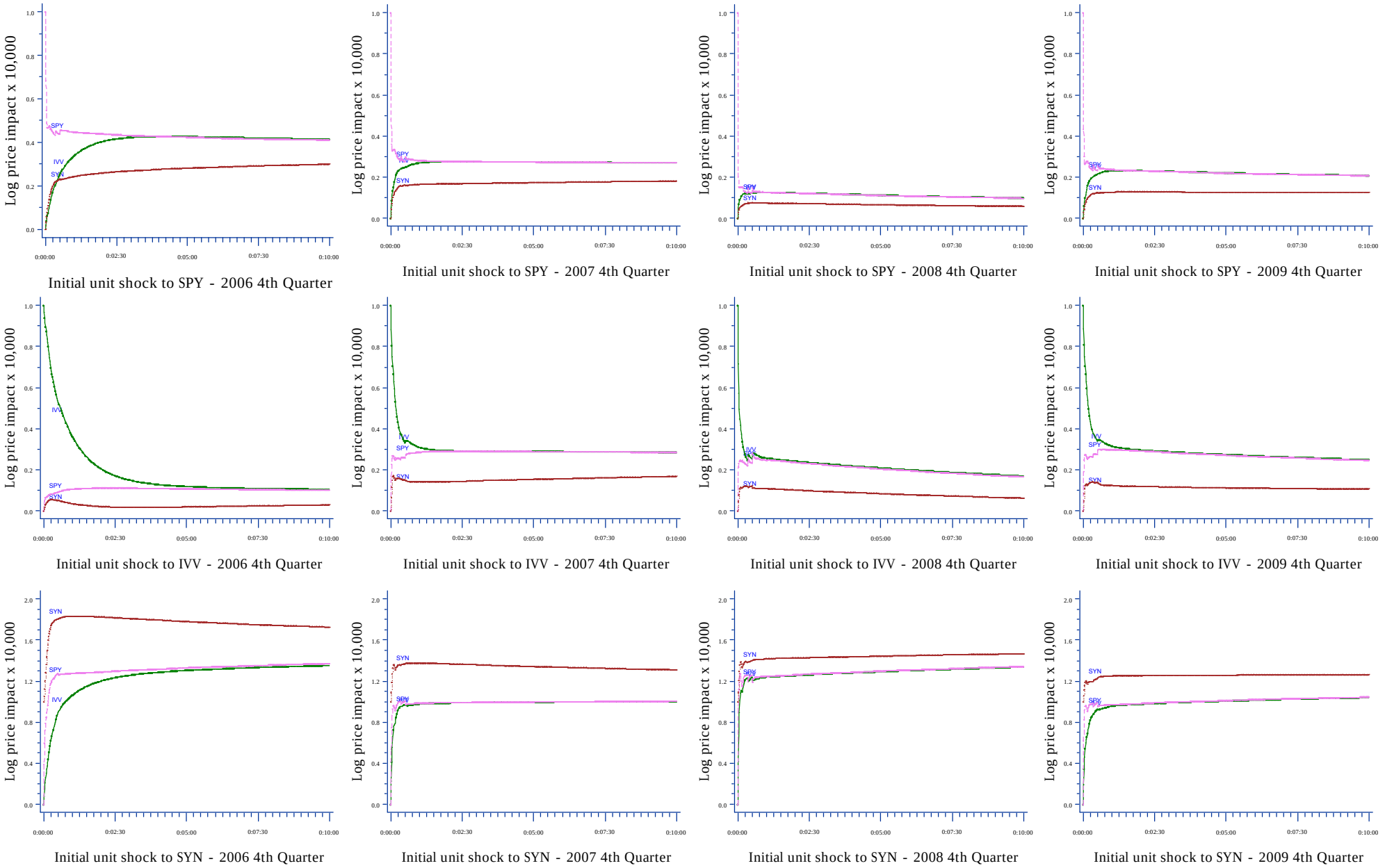

Figure 2.1
Cumulative Impulse Response Functions

The figures plot cumulative impulse response functions for prices of SPY, IVV and SYN. The three rows of figures correspond to a unit shock to SPY, IVV, and SYN repectively and show four figures of the means of fourth quarters from 2006 to 2009. A VECM system of the three price series is used to compute the estimates and is based on 1-second resolution.

Figure 2.2
Long-run Price Impacts

The figures plot the long-run price impacts on SPY and SYN from a unit shock to SPY, IVV, and SYN in the sample period from Sept 1, 2006 to Dec 31, 2009. The responses of IVV are identical to that of SPY and not shown here. The long-run level is taken at the end of the 10-minutes iteration process of the estimation of VECM consisting of the three price series.

Regardless of the needs of investors, the trading is based on the expectation of the overall index level. The top stock in the index in terms of market value is only around 5%. It is not likely that the shock to the individual component stock would have a big price impact on the ETF. Thus, the responses to the shock to ETF can be seen as the one that affects the entire index portfolio. In the case of S&P 500, it is close to the market-wide macro-related innovations. The responses to the index portfolio shocks are based on the expectations of the market as well as the individual stocks since the intraday trading is weighted across all stocks in the index. In any trade, the price impact reflects both market-wide and stock-specific innovations. It is also worth noting that in all three cumulative response functions, the *SPY* and the index portfolio have locked up each other even in small ups and downs. The *IVV* curve is smoother and takes longer time to reach its permanent level. In some sense, it indicates a more tied up relation between the trading of the index portfolio and *SPY*, which has a longer history and a much larger trading volume than *IVV*.

From above analysis on the impulse response functions, we assure that *SPY* and *IVV* have the same underlying random walk. From here, most of the statistics will be in a pair based on the underlying random walk of ETF (same for both *SPY* and *IVV*) and the synthesized index (*SYN*)[1]. In each iteration of the daily information share analysis, the contribution to the price discovery to any security is divided between the three price vectors. Then a median number is taken among all daily iterations. For each pair of median information share, We add together the information shares of *IVV* and *SPY*, since an identical random walk underlies the two price vectors. Not shown here, while the distribution of *Share_ETF_e*, based on the ETF series, is very close to a standard normal distribution, *Share_ETF_s*, based on the synthesized index series, is skewed to the left.

Panel A of Table 2.1 shows quarterly averages of these medians. The breakdown of the information shares of *SPY* and *IVV* are identical, proving again that their prices follow the same random walk. The median combined shares of *SPY* and *IVV* to the price discovery of the one underlying them vary from 40% to 59% for the studied period, with a mean of 45%, while the index portfolio has a share of 55% on average. For the random walk underlying the index portfolio, *SPY* and *IVV* combine for a share varying from 23% to 34% and averaging at 28%, while the mean of the contribution from the index portfolio is 72%. The bulk of the information in both random walks is discovered in the index portfolio trading, which is more dominant in its own price discovery. The result is not surprising in that we have seen the impulse response functions. The macro-related information is more evenly channeled in ETF and component stocks, while the stock-specific information discovery is heavily tilted toward the index portfolio. The quarterly mean of percentage of the combined ETF contributing to the price discovery of ETF and index portfolio has a correlation coefficient of 0.87, which strongly suggests that the factors that make ETF a more likely source of price formation are the same for both random walk processes.

How the information share is distributed can be correlated to trading volume and market value. It is likely that more frequent trading or larger market cap would lead to more price discovery. Panel A also shows the quarterly market cap ratio of *SPY* and *IVV* combined versus the index portfolio *R_MCAP* ((*SPY*+*IVV*)/*SYN*). Obviously ETF is growing much faster than the S&P 500 index in terms of market cap. In the fourth quarter of 2006, the index is 155 times larger than the combined ETF market cap. Three years later, the index is only 93 times larger. Trading volume is an indicator of trading interest. *R_VOL* is the ratio of average daily trading volume of the combined *SPY* and *IVV* against the index portfolio. In the fourth quarter of 2006, the daily combined trading volume of the index portfolio is 37 times that of *SPY* and *IVV* combined. In the same quarter in 2009, it is only 25 times. Despite the discrepancies shrinking fast between ETF and index portfolio in both trading volume and market cap, the information share distribution

[1] Remind that however, the price impacts brought by a shock to *SPY* and *IVV* are different, thus being treated separately.

is not shifting towards the ETF. Panel B of Table 2.1 is a correlation table of concurrent quarterly data including *Share_ETF_e*, *Share_ETF_s*, *R_VOL*, and *R_MCAP*. Two sets of correlations are reported for the two underlying random walk. Not surprisingly, the two *Share_ETF* variables are highly correlated. Both *Share_ETF* are negatively correlated to the two ratios but without significance.

Table 2.1
Price Discovery of S&P 500

This table presents the summary statistics of the price discovery of S&P 500. In Panel A, *Share_ETF* is the quarterly means of the combined share of two ETF, *SPY* and *IVV,* to the price formation of S&P 500. *Share_ETF* is estimated daily in a VAR system consisting of *SPY*, *IVV* and *SYN*, which is the intraday series of S&P 500 reconstructed from the trade prices of component stocks. Since the random walk of two ETF is slightly different from that of *SYN*, the estimates are reported separately for the three price series. For each day in the sample from September 1 2006 to December 31 2009, 600 iterations are performed using a VECM model. The daily *Share_ETF* is the mean of all medians of the estimation of information share contributed by two ETF. Only the fourth quarter results are shown. *R_VOL* and *R_MCAP* are the means of daily trading volume ratio and market capitalization ratio of two ETF combined against the entire S&P 500 component stocks combined. Panel B presents the (Pearson)correlation table of the variables. Since the estimates based on the random walk of *SPY* and *IVV* are very close, only the results based on the random walk of *SPY* (*Share_ETF_e*) and *SYN* (*Share_ETF_s*) are shown. *** denotes significance at the 99% level.

Panel A: Quarterly means

	random walk innovation	*Share_ETF*	*R_VOL*	*R_MCAP*
2006 Q4	*SPY*	51.4%	0.6%	2.6%
2007 Q4		50.6%	0.9%	5.6%
2008 Q4		34.7%	1.4%	7.3%
2009 Q4		47.6%	1.1%	3.8%
2006 Q4	*IVV*	52.4%	0.6%	2.6%
2007 Q4		50.7%	0.9%	5.6%
2008 Q4		35.1%	1.4%	7.3%
2009 Q4		48.0%	1.1%	3.8%
2006 Q4	*SYN*	29.5%	0.6%	2.6%
2007 Q4		30.4%	0.9%	5.6%
2008 Q4		25.7%	1.4%	7.3%
2009 Q4		28.8%	1.1%	3.8%

Panel B: Correlations

	Share_ETF_e	*Share_ETF_s*	*R_VOL*	*R_MCAP*
Share_ETF_e	1			
Share_ETF_s	0.873***	1		
R_VOL	-0.234	-0.217	1	
R_MCAP	-0.409	0.102	0.630**	1

Now we turn focus to the two ETF. It would be interesting to see whether *SPY* and *IVV* play the same role in terms of information discovery process. The data points when both ETF are available dates back as early as 2000. For both sets of underlying process, *SPY* dominates with above 80% of the shares. Except for the first few quarters after the inception of *IVV*, the share distribution between the two is identical, suggesting that the market quickly eliminated any difference between the underlying random walk processes. However, Panel A of Table 2.2 tells us that the dominance of *SPY* just began to disappear in 2007. Until the first quarter of 2007, *SPY* provided most of the price discovery with at least 96% of the total share. But since then, its share has drastically declined over the time. The mean share of *SPY* is 96 percent before the first

quarter of 2007 and is 61 percent after that. The sudden and large change seems puzzling. While the trading volume of *IVV* steadily remain at about 1 to 2 percent that of *SPY*, the relative market value of *IVV* to *SPY* has increased and peaked at 2006.

Table 2.2
SPY* vs. *IVV

This table presents the summary statistics of the price discovery of S&P 500 in a VAR system consisting of *SPY* and *IVV*. In Panel A, for each day in the sample from May 19 2000 to December 31 2009, 600 iterations are performed using a VECM model. The daily *Share_SPY_s (Share_SPY_i)* is the mean of all medians of the estimation of information share contributed by *SPY* based on the underlying random walk of *SPY* (*IVV*). Only the fourth quarter results are shown. *R_VOL* and *R_MCAP* are the means of daily trading volume ratio and market capitalization ratio of *IVV* against *SPY*. Panel B presents the (Pearson)correlation table of the variables. Since the estimates based on the random walk of *SPY* and *IVV* are very close (as shown in Panel A), only the results based on the random walk of *SPY* are shown.

Panel A: Quarterly means

	Share_SPY_s	*Share_SPY_i*	*R_VOL*	*R_MCAP*
all sample	85.5%	82.4%	1.7%	19.8%
2000 Q4	88.6%	75.3%	1.8%	2.6%
2001 Q4	97.6%	96.6%	1.1%	21.0%
2002 Q4	98.7%	98.7%	1.3%	13.8%
2003 Q4	98.7%	98.6%	1.4%	17.6%
2004 Q4	98.8%	98.8%	1.4%	20.8%
2005 Q4	96.1%	96.1%	1.7%	24.1%
2006 Q4	95.5%	95.5%	1.4%	28.1%
2007 Q4	61.7%	61.1%	1.4%	19.0%
2008 Q4	44.0%	44.0%	2.3%	16.5%
2009 Q4	59.1%	59.1%	2.3%	25.5%

Panel B: Correlations of daily observations

	Share_SPY	*R_VOL*	*R_MCAP*
Share_SPY	1		
R_VOL	-0.76	1	
R_MCAP	-0.01	-0.08	1

The correlation matrix between concurrent quarterly variables is shown in Panel B of Table 2.2. Only the statistics based on *SPY*'s random walk is shown since those based on *IVV* is almost identical. The daily mean of the information share of *IVV* is positively correlated to the trading volume ratio. However, it is not necessary that the growing trading volume spurs the price discovery. Intuitively, it is inevitable that *IVV* would narrow the distance between itself and *SPY* in terms of trading, since they are essentially the ETF tracking the same underlying index.

2.5 Private Information-based Trading in ETF

We have seen the importance of ETF trading from their contribution to the information share of the underlying random walk. ETF contribute more to their own random walk than to that of the synthesized index. The difference can be that the trading data of the synthesized index simply contain more private information than the trading of ETF. Therefore, knowing the nature of ETF trading would help us further understand the ever-changing role of ETF.

Let's imagine extreme cases. Assuming that ETF trading is driven by public information only, does it still contribute to the price discovery of the underlying index? What if it is entirely private information driven? If ETF is indeed transiting to a more active role in the price discovery process, can we say it is because investors trading them possess more inside information? Public or private? Once we are able to answer all these questions, we will establish a connection between price formation of index and information contents of trading as well as how ETF fits in the big picture.

There are mainly two approaches to infer how much trading is private information driven. One is based on the VAR model by Hasbrouck (1991b), which is the main analyzing tool We use in the first part of this paper. The variables used in the VAR system are sequential changes in the quote midpoint and an indicator variable to define buy or sell orders. Each price movement is either associated with the most recent trade or not. Those changes that are associated with the prior trade are driven by private information. Those that are orthogonal to the prior trade are considered only based on public information[2].

The same approach is used in many papers studying microstructural effects of trading mechanism including Barclay and Hendershott (2003) and Hendershott, Jones, and Menkveld (2011).

The other approach developed to examine the amount of information revealed by trading is Probability of Informed Trading (*PIN*) developed by Easley and O'Hara (1992). Later, different variations of *PIN* are proposed to improve on the original one. Different from the sequential trade models, *PIN* is solely derived from the order arrival process. Neither the direction of a trade nor the quote midpoint changes depends on the prior order. Neither does it measure the absolute amount of information nor the proportion of private or public information.

Throughout this paper, we are interested in how the trading of ETF has affected the index itself. The sequential model is an ideal fit examining the dynamics between trades and prices as well as accommodating several price series. We would use the Hasbrouck approach to determine the information nature of trades in ETF, which is also consistent with the analysis in the first part.

The price variable considered in this section is quote midpoint p_t. The other variable in the VAR is the trade direction q_t. A trade is defined as buyer-initiated if its execution price is higher than the quote midpoint or seller-initiated if lower. There are trades executed exactly on midpoint. One solution is to look up the most recent price change. If the previous trade is a buy order, then the one following is also a buy order. We can trace all the way back if the direction of previous trade is not determined either. The signed trades are observed to be positively autocorrelated, i.e., buy orders tend to be followed by buy orders. The quote midpoints evolve according to the following:

[2] Strictly speaking, the decomposition in this section is about informed trading and non-informed trading. Vega (2006) argues that it is the arrival rate of informed or uninformed traders that matters rather than whether the information is public or private. We use these concepts interchangeably throughout this paper.

$$p_t = m_t + cq_t \tag{2.5}$$

q_t is assumed to capture the transient microstructure imperfections that drive quote midpoints deviate from the efficient price. m_t is the efficient price following a random walk process: $m_t = m_{t-1} + \omega_t$. ω_t is the innovation of the random walk, $\omega_t = u_t + \lambda v_t$. v_t is the innovation of the signed trades

$$v_t = q_t - E\{q_t | q_{t-1}, q_{t-2}, \ldots\} \tag{2.6}$$

u_t is a white noise process. Using Cholesky factorizations, we have $\sigma_\omega^2 = \sigma_u^2 + \lambda^2 \sigma_v^2$. The first component is the variance of a white noise, thus representing the public information. The second component is the variance of the signed trades' innovation, representing the private information because it is trade-related.

In a VAR system, different ordering of variables would generate different results in random-walk analysis of the innovations of VAR variables. When the variables in the system are various prices as in the first part, we can process all the different orderings and choose a statistical median or mean. In the end, it does not affect the interpretation of the results. However, it is not easy to interpret when the variables in the system are of distinctive natures, for example in this case, one describing the persistence of order flow *q* and the other price movements *p*, where the causal effects could run either way. If the trade direction is put first in the order, the structural model suggests that contemporaneous trade drives changes in price. If the order is reversed, then it is the price change that drives the direction of contemporaneous order. Moreover, we must specify the ordering first to explain the impulse response functions often used in the VAR system. Since we are more interested in how ETF trading changes the information discovery process of underlying index, only the results of the ordering where trade drives prices are reported unless mentioned else. The price variable can be trade prices or quote midpoints. Hasbrouck (2007) suggests that, in the frame of studying how trades contribute to the efficient price variance and impact the price in the long-run, the bid-ask midpoint is a better choice, because quotes are active between revisions while the most recent trade prices may still be outdated.

According to Hasbrouck (1991b), we can impose a casual ordering, which is done through Cholesky factorization to decompose the variance of random-walk innovation into trade-related components and public information. The corresponding ratio of the component variance to the total variance is their contribution to the total information. The VAR system is estimated for each day from September 2006 to December 2009 for the trading of *SPY* and *IVV*. The intraday quotes and trades of *SPY* and *IVV* are obtained from TAQ database. The quotes are within the time frame from 9:30 to 16:00 from all trading locations. There are many original unsigned trades for *SPY* and many of them occur at the same second or consecutively. We decide to discard those trades because the actual order of these trades cannot be precisely determined at the 1-second-level resolution. Fortunately, we are still left of sufficiently large data for both ETF.

As reported in Panel A of Table 2.3, the total variance of the random-walk innovation is only 0.25 bps on average for *SPY* (*VT_SPY*), which is not even one hundredth of the daily average for *IVV* (46 bps, *VT_IVV*). Figure 2.3 shows the total variance for SPY has a huge yet short-lived spike in October 2008, when the entire financial market was deep in the subprime mortgage crisis. In comparison, *IVV* has far more spikes scattered throughout the studied period and many of them are prolonged. One group of the spikes is clustered around October 2008. *SPY* is much larger in terms of trading volume and market capitalization. However, the intensity of information seems much stronger in the smaller ETF. In other words, investors prefer trading *IVV* whenever they have new information, public or private. *SPY* is more of an indexing ETF while *IVV* has more speculative trading. It is possibly because informed traders find it easy to move prices in the relatively smaller ETF.

Table 2.3
Information Content in ETF Trading

This table shows how the price discovery process of S&P 500 is linked to the trading of ETF. The estimated model is a VAR system consisting of trade directions (q) and quote midpoints (p). The value of q is 1 if the trade is buyer initiated and -1 if seller initialed. In the model, trades drive quote movements. The variance of the innovation of the quote movements (VT) can be decomposed into the trade related variance component and non-trade related. The proportion of the trade related to the total variance is PR, representing the share of private information-based trading. The remaining is public information-based. *Share_ETF*, reported in Section 3, is the contribution made by two ETF combined to the efficient price of S&P 500. The underlying random walk process is of ETF series for *Share_ETF_e* and of synthesized S&P 500 for *Share_ETF_s*. *VIX* is the S&P 500 volatility index. *R_VOL* and *R_MCAP* are the means of daily trading volume ratio and market capitalization ratio of two ETF combined against the entire S&P 500 component stocks combined. The sample period is Sept 1 2006 to Dec 31 2009. Panel A shows the means of *VT* and *PR*. Panel B is a (Pearson) correlation table of variables. Panel C presents the estimates from regressions with *Share_ETF* as the dependent variable on concurrent *VT*, *PR* and control variables in different specifications. *t-statistics* are reported under the estimates.

Panel A: Summary statistics

	VT_SPY	*VT_IVV*	*PR_SPY*	*PR_IVV*
mean	0.25bps***	46bps	7.6%***	11.9%***
t-stat	5.82	1.05	32.52	32.02

Panel B: Correlations of daily observations

	Share_ETF_e	*Share_ETF_s*	*VT_SPY*	*VT_IVV*	*PR_SPY*	*PR_IVV*	*VIX*
Share_ETF_e	1						
Share_ETF_s	0.83***	1					
VT_SPY	-0.17***	-0.13***	1				
VT_IVV	-0.09***	-0.07*	0.02	1			
PR_SPY	0.17***	0.16***	-0.14***	-0.03	1		
PR_IVV	-0.00	0.17	-0.08**	-0.04	0.70***	1	
VIX	-0.24***	-0.05	0.05	0.02	0.28***	0.49***	1

Panel C: Regressions

Dependent variable: *Share_ETF_e*

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
VT_SPY	-247***	-198***	-185***	-264***	-200**	-188***			
	-3.91	-3.2	-3.02	-4.11	-3.22	-3.06			
VT_IVV	-0.12**	-0.1**	-0.1**				-0.12**	-0.11**	-0.11**
	-2.54	-2.35	-2.38				-2.55	-2.33	-2.35
PR_SPY	0.81***	0.73***	0.74***	0.39***	0.60***	0.70***			
	6.43	5.9	6.06	4.24	6.51	7.22			
PR_IVV	-0.38***	-0.14*	-0.06				-0.007	0.23***	0.31***
	-4.86	-1.65	-0.61				-0.11	3.57	4.22
VIX		-0.003***	-0.005***		-0.004***	-0.005***		-0.004***	-0.006***
		-6.66	-5.65		-8.18	-5.7		-7.66	-6.12
R_VOL			2.33***			2.41***			2.34***
			3.78			4.05			3.68
R_MCAP			-0.39			-1.87			-0.29
			-0.07			-0.36			-0.05
r-square	8%	13%	15%	5%	13%	15%	1%	8%	9%

Table 2.3 continued

Dependent variable: *Share_ETF_s*

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
VT_SPY	-155***	-150***	-143**	-166***	-150***	-144***			
	-2.8	-2.68	-2.57	-2.97	-2.66	-2.59			
VT_IVV	-0.07*	-0.07*	-0.07*				-0.07*	-0.07*	-0.07*
	-1.77	-1.73	-1.67				-1.8	-1.73	-1.67
PR_SPY	0.58***	0.57***	0.58***	0.31***	0.36***	0.50***			
	5.23	5.12	5.23	3.84	4.34	5.74			
PR_IVV	-0.24***	-0.22***	-0.1				0.02	0.07	0.19***
	-3.58	-2.86	-1.16				0.39	1.21	2.88
VIX		-0.0004	-0.0009		-0.001**	-0.001		-0.001*	-0.001
		-0.86	-1.03		-2.33	-1.14		-1.81	-1.54
R_VOL			1.98***			2.14***			1.99***
			3.55			3.99			3.49
R_MCAP			-9.86			-12.07***			-9.78*
			-1.98			-2.59			-1.92
r-square	5%	5%	7%	3%	4%	7%	0%	1%	3%

As reported in Panel A of Table 2.3, the total variance of the random-walk innovation is only 0.25 bps on average for *SPY* (*VT_SPY*), which is not even one hundredth of the daily average for *IVV* (46 bps, *VT_IVV*). Figure 2.3 shows the total variance for SPY has a huge yet short-lived spike in October 2008, when the entire financial market was deep in the subprime mortgage crisis. In comparison, *IVV* has far more spikes scattered throughout the studied period and many of them are prolonged. One group of the spikes is clustered around October 2008. *SPY* is much larger in terms of trading volume and market capitalization. However, the intensity of information seems much stronger in the smaller ETF. In other words, investors prefer trading *IVV* whenever they have new information, public or private. *SPY* is more of an indexing ETF while *IVV* has more speculative trading. It is possibly because informed traders find it easy to move prices in the relatively smaller ETF.

Panel A of Table 2.3 also shows the mean of the daily estimates of the proportion of total variance that can be attributed to private information (*PR*). 7.54 percent of the price moves are trade related on average for *SPY* (*PR_SPY*) and 11.90 percent for *IVV* (*PR_IVV*). Multiplying *VT* with *PR*, the absolute magnitude of the private information trading in *IVV* (5.47 bps) is 290 times larger than in SPY (0.019 bps), meaning that informed trading is even more concentrated in *IVV* than in *SPY*.

Before we proceed on to multivariate analysis, there is one more variable that should be considered. *VIX*, the market volatility index, is used in the pricing of S&P 500 index options. It measures the market's expectation of the next 30 days' volatility. It is sometimes called 'fear index' since it reflects the risk appetites of investors in the short-term future. There might be a nontrivial interaction of how people perceive the market's volatility and the location of price discovery.

Correlations between the daily variables are shown in Panel B of Table 2.3 including price discovery contribution variables from Section 3. In sharp contrast of no correlation between the two total variances, the correlation coefficient between *PR_SPY* and *PR_IVV* is 0.70. This suggests that the distribution of

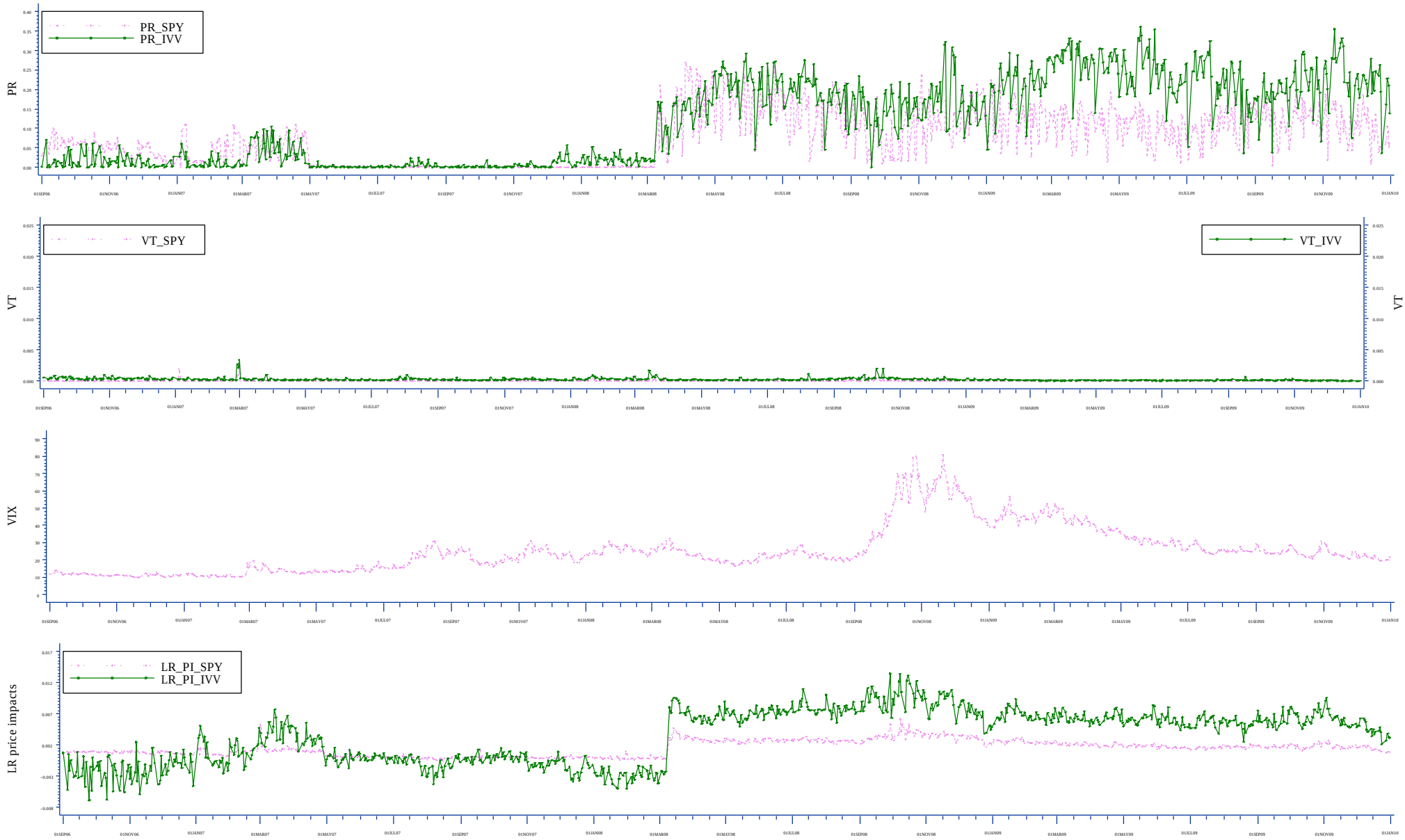

Figure 2.3
Private Infomration-based Trading

All plots are daily estimates of a VAR consisting of quotes and order flows of SPY and IVV from the sample period Sept 1 2006 to Dec 31 2009. The top figure plots the proportion of quotes movement related to trade (PR). The second figure plots the total variance of innovation of the order flow (VT). The third figure plots VIX. The bottom figure plots the long-run cumulative impulse responses on ETF quotes to a unit shock of order flow

informed traders is similar for *SPY* and *IVV* even though the information intensity is very different. The cross-correlation between *VT* and *PR* yields -0.14 and -0.04 for *SPY* and *IVV* respectively, with only the former being significant. Among all variables, *VIX* is strongly correlated with both *PR* variables at 0.27 (*SPY*) and 0.49 (*IVV*). It may be that, when the expectation of future volatility is high, investors who have the inside information are more likely to trade than those with only publicly known news, thus pushing higher the ratio of trade-related price movements. *VIX* has also significant and positive correlation with *Share_ETF_e*. The above variables are also plotted against time in Figure 2.3.

To better understand the link between information share distribution and the trading of ETF, We run several specifications of regressions. Results are shown in Panel C of Table 2.3. The basic setup is:

$$Share\_ETF_t = \alpha + \beta_{SPY} VT\_SPY_t + \beta_{IVV} VT\_IVV_t + \gamma_{SPY} PR\_SPY_t \gamma_{IVV} PR\_IVV_t + u_t \quad (2.7)$$

Share_ETF of the top half based on the random walk of ETF and the bottom on *SYN*. All estimated coefficients are significant. Both *VT* estimates are negative. Let's remember that the total variance includes both trade-related and unrelated components and represent overall information intensity in the trading. When the information surrounding the index is abundant, investors with private information might refrain from trading ETF since the trading is more likely dominated by public information. It can be said that the total information intensity weakens the role of ETF in price discovery. The coefficient on *PR_SPY* is 0.81. Quantitatively, this means that if the ratio of private information trading doubles its means from 7.5% to 15%, the combined contribution of ETF to price discovery will rise about 6%. The negative sign on *PR_IVV* is puzzling since we expect *SPY* and *IVV* behave similarly in the regressions. Considering that the two *PR* variables are highly correlated, We separate *SPY* and *IVV* variables and rerun the regression in (4) and (7). When it is the only *PR* variable in the regression, its estimate is no longer statistically significant.

It is necessary to control some important trending factors in the regressions. Although theoretically we do not know how exactly ETF's contribution to price formation is linked to market factors, it can be affected by the market trend based on the simple fact that the ETF of S&P 500 index in many ways is tied to the entire market. *VIX* is added in regression (2):

$$Share\_ETF_t = \alpha + \beta_{SPY} VT\_SPY_t + \beta_{IVV} VT\_IVV_t + \gamma_{SPY} PR\_SPY_t \gamma_{IVV} PR\_IVV_t + \zeta VIX_t + u_t \quad (2.8)$$

The coefficient is significantly negative but the magnitude is extremely small (-0.003). Interestingly, the addition of *VIX* to the mix raises the R-square from 0.08 to 0.13. Monthly volume ratio (*R_VOL*) and quarterly market cap ratio (*R_MCAP*) are added in (3) as additional control variables. It should be noted that both ratios are two ETF combined against the entire S&P 500. As expected, the volume ratio has a positive coefficient. But the addition of these two ratios only raises the R-square by 0.02.

In (8) and (9), the *PR_IVV* becomes significantly positive with the magnitude less than half of *PR_SPY*. Estimates on *SPY* are consistent for all regression specifications. The results clearly show an economically significant impact of trading of ETF on the price formation process. In the end, it is the private information that drives price discovery. In this sense, *SPY* and *IVV* do not behave anything like people are using them merely for indexing. Even though the mean of *PR_SPY* is only 7.5%, the change of *PR_SPY* has almost one-to-one change in *Share_ETF_e*. To summarize, after controlling for market volatility, trading volume, and market capitalization, how much informed trading of ETF crucially determine their contribution to the price formation of S&P 500 index.

Share_ETF_e is the dependent variable in the alternative specifications. Most of the results are qualitatively the same but the magnitudes of the estimated coefficients are visibly smaller, which is not surprising because

the underlying random walk for *SYN* has a larger information set than that for ETF. Moreover, *VIX* has lost its explanatory power in four of the six regressions it is in and it does not raise the R-square as much as in the set of regressions on *Share_ETF_s*.

We can see that there is a strong link between informed trading in ETF and its contribution to the price discovery of S&P 500 index. Intuitively, when we are talking about where the price is 'discovered', the source of information is mostly likely private. From this perspective, ETF behaves like a 'normal' stock. If ETF is used in most occasions for indexing, we would expect to see a low level of *PR* compared to other stocks. Moreover, the information contribution of ETF should not be driven by the level of informed trading. Therefore it is necessary to check how ETF compare to individual stocks in the index in the information contents revealed by trading.

2.6 Robustness Checks and Extensions

2.6.1 Benchmark Analysis

In the above analysis, we see that *SPY* and *IVV* behave anything but as simply passive investment financial vehicles. To have a better understanding, We construct a benchmark group consisting of component stocks from S&P 500 index and apply the same VAR system to study whether they share the same trading characteristics with ETF. To be considered for selection, the stock must stay in the S&P 500 throughout the 40-months period. 383 stocks, qualified for this criterion, are then ranked based on their average weights in the index. The stocks ranked at the 1st (*XOM*), 101st (*AFL*), 201st (*CEG*), and 301st (*SVU*) are included in the benchmark group to enable us compare cross-section of market capitalization. Using the TAQ data, We perform the same price-trade analysis by Hasbrouck (1991b).

The summary statistics are listed in Panel A of Table 2.4. *XOM* by far has the largest market value, averaging 417 million dollars, followed by *SPY* (72 million). The rest of the three selected range from 5 to 24 million. As of trading volume, *SPY* averages 45 million shares a day which is 7.7 times larger than the second place *XOM*, which in turn is 6.8 times larger than the third place *AFL*. *IVV* ranks right in the middle of the four selected stocks in both market cap and volume. Both ETF have a much larger monthly turnover rate than the individual stocks. *PR* of the benchmark ranges from 7.19% to 12.68%. Putting ETF in the ranking, *SPY* is in fifth and *IVV* is in second. The comparison of *VT* reveals the biggest gap between ETF and component stocks. *VT_SPY* is the smallest but still comparable to benchmark. However, *VT_IVV* seems just too big. Not reported here, all *PR* variables are strongly cross-correlated to each other. This is because most trading is public information related for ETF as well as component stocks.

Figure 2.4 plots the time series of *VT* and *PR* for each benchmark stock. The regime shifts occur around the same time as ETF. We will leave the discussion in the next chapter. We rerun the regressions in specification (6) Panel C of Table 2.3 replacing *SPY* and *IVV* with the four selected stocks. Let's remember that the dependent variable puts ETF 'against' the index portfolio in which all the four benchmark stocks are in it. We don't have a theory of how the trading of component stocks is linked to the price formation of S&P 500. Nonetheless, the benchmark regressions may provide more insight of the relationship between trading and information. Results are shown in Panel B of Table 2.4 along with estimates of *SPY* and *IVV* from Panel C of Table 2.3. All the coefficients on *PR* are positive but only those of *XOM* and *AFL* are significant. Within the benchmark group, larger estimates are associated with larger market cap. For 1% increase of private information trading in *XOM*, the ETF contributes 0.28% more to the price discovery. The magnitude is smaller than *IVV*, which has a fraction of the market value of *XOM*. The R-square

Table 2.4
Benchmark Analysis

*This table shows whether ETF is different from the component stocks of S&P 500 in their trading linked to price discovery. Component stock considered for selection must be present in the index throughout the sample Sept 1 2006 to Dec 31 2009. Qualified stocks are then ranked by average monthly market capitalizaiton. Stocks ranked 1st, 101st, 201st, and 301st are then selected. The estimated model is a VAR system consisting of trade directions (q) and quote midpoints (p). The variance of the innovation of the quote movements (VT) can be decomposed into the trade related variance component and non-trade related. The proportion of the trade related to the total variance is PR, representing the share of private information-based trading. The remaining is public information-based. Share_ETF, reported in Section 3, is the contribution made by two ETF combined to the efficient price of S&P 500. VIX is the S&P 500 volatility index. R_VOL and R_MCAP are the means of daily trading volume ratio and market capitalization ratio of two ETF combined against the entire S&P 500 component stocks combined. Panel A shows the summary stats of the selected stocks. Panel B presents the estimates from regressions with Share_ETF as the dependent variable on concurrent VT, PR and control variables for each selected stock. t-statistics are reported under the estimates. *, **, *** denotes significance at 90%, 95%, and 99% level respectively. The stats of SPY and IVV from previous previous*

Panel A: Comparison of summary statistics

	XOM	*AFL*	*CEG*	*SVU*	**enchmark grou**	*SPY*	*IVV*
value rank	1st	101st	201st	301st			
value weight	3.95%	0.21%	0.10%	0.05%	1.08%		
market cap (in billions)	417.09	23.19	11.07	5.83	114.30	71.65	17.91
trading volume (in millions)	5.94	0.88	0.45	0.60	1.97	45.80	0.79
turnover	1.13	1.88	2.42	2.82	2.07	73.21	5.19
VT (in bps)	0.315	0.921	1.808	0.344	0.847	0.248	46
PR	11.98%	12.68%	7.19%	10.61%	10.61%	7.55%	11.91%
corr_VT_PR	-0.17	-0.11	-0.32	-0.13	-0.18	-0.14	-0.04

Panel B: Regressions by stocks

	VT	*PR*	*VIX*	*R_VOL*	*R_MCAP*	r-square
Dependent variable: *Share_ETF_e*						
XOM	23.7	0.28***	-0.006***	2.05***	3.13	
	0.29	3.66	-5.9	3.26	0.56	8%
AFL	-51.1**	0.23***	-0.005***	2.30***	1.88	
	-2.05	3.46	3.62	3.62	0.33	9%
CEG	-31.8	0.19	-0.006***	1.98***	7.10	
	-1.52	1.56	-5.94	3.03	1.34	7%
SVU	-94.8**	0.12	-0.006***	1.76***	7.03	
	-2.35	1.36	-5.76	2.78	1.29	8%
SPY	-188.9***	0.70***	-0.005***	2.41***	-1.87	
	-3.06	7.22	-5.7	4.05	-0.36	15%
IVV	-0.11**	0.31***	-0.006***	2.34***	-0.29	
	-2.35	4.22	-6.12	3.68	-0.05	9%
Dependent variable: *Share_ETF_s*						
XOM	132.6*	0.23***	-0.002*	1.87***	-8.41*	
	1.82	3.28	-1.78	3.32	-1.69	3%
AFL	-20.7	0.16***	-0.001	1.99***	-9.18*	
	-0.93	2.7	-1.15	3.51	-1.82	2%
CEG	-9.4	0.10	-0.001	1.62***	-4.75	
	-0.51	0.88	-1.4	2.78	-1.01	1%
SVU	-53.7	0.10	-0.001	1.7***	-5.92	
	-1.49	1.34	-1.38	3.01	-1.22	2%
SPY	-144.7***	0.50***	-0.001	2.14***	-12.07***	
	-2.59	5.74	-1.14	3.99	-2.59	7%
IVV	-0.11**	0.31***	-0.006***	2.34***	-0.29	
	-2.35	4.22	-6.12	3.68	-0.05	3%

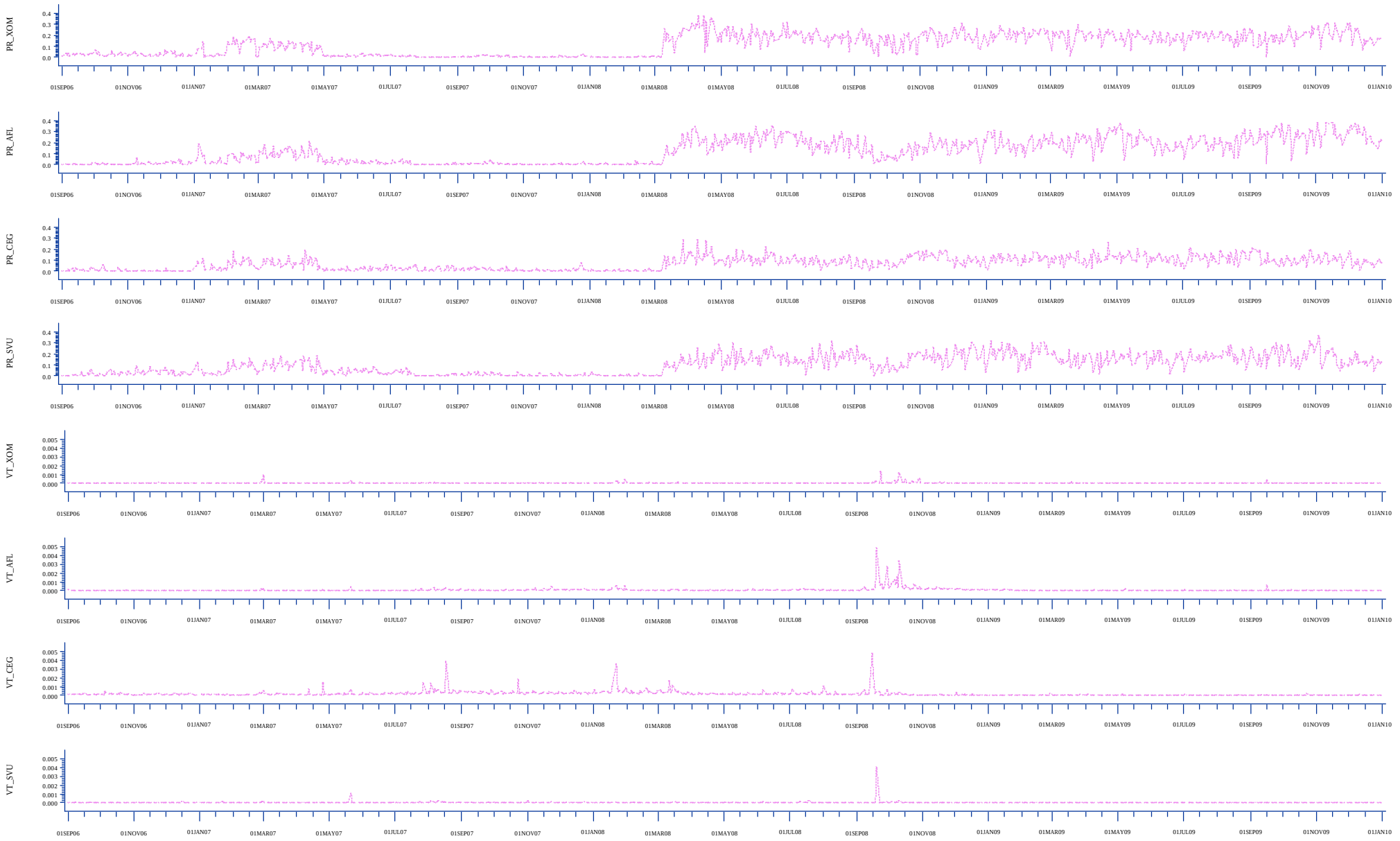

Figure 2.3
Private Infomration-based Trading

All plots are daily estimates of a VAR consisting of quotes and order flows of SPY and IVV from the sample period Sept 1 2006 to Dec 31 2009. The top figure plots the proportion of quotes movement related to trade (PR). The second figure plots the total variance of innovation of the order flow (VT). The third figure plots VIX. The bottom figure plots the long-run cumulative impulse responses on ETF quotes to a unit shock of order flow

statistics are also correspondingly smaller. The fact that *PR* of large-cap stocks also has the explanatory power can be explained by the commonality of information composition of the entire market. Private information is revealed through the trading of all stocks including ETF. At the same time, ETF makes a larger contribution to price formation on days when there is more informed trading. On the other side, we can view component stocks part of the ETF. So it's not surprising that the regressions on individual stocks yield smaller estimates with the same signs.

Figure 2.3 also depicts the time series of the long-run cumulative responses of quote midpoints to a unit shock of signed trades. The estimation of response functions is based on the iterations of 60 lags of each quote. Each point in the graph is the daily average of the level of the cumulative response functions at the 60th lag. In general, the same series of all six stocks move in the same direction over the time with the presence of the regime shift around March 2008. Particularly, the quote midpoints of two ETF have almost identical responses to the order flow. As for the long-run level of cumulative impulse responses, *SPY* and *XOM* show a sudden increase at the regime shift and the other four display a steady increase over the time.

The long-run responses of prices to signed trades for both ETF strikingly resemble that of the ratio of trade-related price movements, with two structural breaks at the same time. It implies that, if trading based on private information is of larger portion for the day, it also has a larger long-run impact on the prices. In fact, the correlation between long-run cumulative responses of price to a shock to signed trade is 0.67 with significance for both ETF.

The small distinctions between ETF and benchmark stocks reveal the fact that ETF is becoming a more active investment tool just like any other individual stock as opposed of remaining as a passive indexing tool. In the most, it can be said that ETF is just a much bigger 'stock' on which investors trade with their private information.

2.6.2 Regime Shifts

There is one concern. We see the apparent two structural breaks in Figure 2.3. *PR_SPY* on March 7 2008 is 0.086 percent but the next day it jumps to 16.96 percent. Similarly, it increases from 0.0164 percent to 16.89 percent for *IVV* between the two days. Another break occurs at the end of April 2007. Table 10a gives the mean *PR* for each period when we divide the time around the two breaks. The means for the three periods are 4.9%, 0.8% and 12% respectively for *SPY*. For *IVV*, it's 2.3%, 0.8% and 21%. According to the TAQ documents on WRDS, there is no known major change made to the database. The regime shifts also appear in Figure 2.4 for the benchmark group. Thus, the driving factor must be system wide, affecting both ETF and component stocks. In this section, We provide some preliminary analysis.

Mean statistics of all periods are presented in Panel A of Table 2.5. Period 1 has 163 daily observations from September 1 2006 to April 27 2007. Period 2 has 216 observations from April 30 2007 to March 7 2008. Period 3 has 459 observations from March 10 2008 to December 31 2009. *VT_IVV* changes dramatically between the breaks. It jumps from 2.3 bps in the second period to 82 bps in the third. Not reported here, the correlation between the *VT_SPY* and *VT_IVV* is significantly high in Period 1 and 2 at 0.67 and 0.54. The correlation between the two *PR* is negative in Period 2. The cross-correlations between *VT* and *PR* are all negative except for *IVV* in Period 2. Most of the abnormal statistics are caused by the extremely low *PR* in Period 2.

Table 2.5
Regime Shifts

This table shows the results divided by two regime shifts in April 2007 and March 2008. The estimated model is a VAR system consisting of trade directions (q) and quote midpoints (p). In the model, trades drive quote movements. The variance of the innovation of the quote movements (VT) can be decomposed into the trade related variance component and non-trade related. The proportion of the trade related to the total variance is PR, representing the share of private information-based trading. The remaining is public information-based. Share_ETF, reported in Section 3, is the contribution made by two ETF combined to the efficient price of S&P 500. The underlying random walk process is of ETF series for Share_ETF_e and of synthesized S&P 500 for Share_ETF_s. VIX is the S&P 500 volatility index. R_VOL and R_MCAP are the means of daily trading volume ratio and market capitalization ratio of two ETF combined against the entire S&P 500 component stocks combined. Panel A shows the means of VT and PR. Panel B presents the estimates from regressions with Share_ETF as the dependent variable on concurrent VT, PR and control variables in different specifications. t-statistics are reported under the estimates. *, **, *** denotes significance at 90%, 95%, and 99% level respectively.

Panel A: Summary statistics by periods

	Period 1: 9/01/06-04/27/07	Period 2: 4/30/07-3/07/08	Period 3: 03/10/08-12/31/09	all sample
Share_ETF_e	45.91%	45.47%	44.47%	44.97%
Share_ETF_s	27.24%	29.19%	28.34%	28.37%
VT_SPY (bps)	0.44	0.21	0.2	0.25
VT_IVV (bps)	3.56	7.45	81.7	46
PR_SPY	4.95%	0.07%	12.00%	7.56%
PR_IVV	2.27%	0.77%	20.60%	11.92%
VIX	12	21	33	26
R_VOL	2.97%	5.58%	5.12%	4.82%
R_MCAP	0.63%	0.79%	1.05%	0.90%

Regressions are also redone by periods in Panel B of Table 2.5 with comparison to the overall results from Panel B. In general, the results of Period 3 are the closest to the overall. For the underlying process of ETF, the coefficient on *PR_IVV* is significantly positive in Period 3 while it is significantly negative in the second, which leads to further suspicion that Period 2 might be an anomaly. Most interestingly, the R-square are all larger except for one in regressions by periods. One possible explanation is that, some fundamental differences between the three periods reduce the explanatory power of regressions when all observations are pooled together. R-square is also generally the largest in Period 3, which could mean that the trading of ETF has larger impacts on the price discovery process of S&P 500 index! The results for dependent variable *Share_ETF_s* are similar[1].

[1] Unfortunately, we still don't have a clear explanation of what causing the regime shifts. All we know is that it affects the entire market, including individual stocks and ETF. But it is definitely interesting for further studies because it might be related to some fundamental changes in the microstructure world.

Table 2.5 continued

Panel B: Regressions by periods

Dependent variable: *Share_ETF_e*												
	all sample	Period 1: 9/01/06-04/27/07	Period 2: 4/30/07-3/07/08	Period 3: 03/10/08-12/31/09	all sample	Period 1: 9/01/06-04/27/07	Period 2: 4/30/07-3/07/08	Period 3: 03/10/08-12/31/09	all sample	Period 1: 9/01/06-04/27/07	Period 2: 4/30/07-3/07/08	Period 3: 03/10/08-12/31/09
VT_SPY	-185.8***	-109.8	-25.73	-57.26	-188.9***	-136.97*	-521.8**	-137.3				
	-3.02	-1.39	-0.1	-0.35	-3.06	-1.82	-2.22	-0.85				
VT_IVV	-0.1**	-73.89	-360.7***	-0.09**					-0.11**	-105.9*	-366.9***	-0.11**
	-2.38	-1.27	-3.77	-2.08					-2.35	-1.91	-4.42	-2.42
PR_SPY	0.74***	0.68	8.05	0.90***	0.70***	0.74	5.55	0.96***				
	6.06	1.2	0.66	6.06	7.22	1.31	0.42	6.43				
PR_IVV	-0.06	0.58	-4.92***	0.27**					0.31***	0.56	-4.88***	0.43***
	-0.61	0.79	-4.22	2.17					4.22	0.76	-4.22	3.43
VIX	-0.005***	-0.008	-0.002	-0.005***	-0.005***	-0.017	-0.012***	-0.004***	-0.006***	-0.011	-0.002	-0.006
	-5.65	-0.53	-0.47	-4.78	-5.7	-1.3	-3.13	-4.27	-6.12	-0.77	-0.54	-5.37
R_VOL	2.33***	-9.65**	-0.37	3.85***	2.41***	-6.93*	0.54	3.19***	2.34***	-8.21**	-0.40	5.00***
	3.78	-2.42	-0.2	4.19	4.05	-1.92	0.27	3.59	3.68	-2.07	-0.22	5.39
R_MCAP	-0.39	74.04	21.12*	-5.68	-1.87	22.24	26.05**	-6.76	-0.29	52.98	18.87	-12.3*
	-0.07	0.74	1.76	-0.81	-0.36	0.23	2.09	-0.96	-0.05	0.54	1.65	-1.71
r-square	15%	19%	25%	26%	15%	17%	12%	24%	9%	16%	24%	18%
Dependent variable: *Share_ETF_s*												
VT_SPY	-143.4**	-164.6**	120.3	4.59	-144.7***	-155.9**	-206.5	-39.76				
	-2.57	-2.47	0.48	0.03	-2.59	-2.48	-0.92	-0.27				
VT_IVV	-0.07*	11.77	-232.6**	-0.05					-0.07*	-30.1	-211.3**	-0.07
	-1.67	0.24	-2.46	-1.29					-1.67	-0.64	-2.57	-1.64
PR_SPY	0.58***	0.21	5.87	0.80***	0.50***	0.26	3.66	-0.83***				
	5.23	0.44	0.48	5.8	5.74	0.56	0.29	6.07				
PR_IVV	-0.1	0.64	-4.09***	0.15					0.19***	0.52	-4.03***	0.28**
	-1.16	1.04	-3.54	1.29					2.88	0.83	-3.51	2.43
VIX	-0.0009	0.004	0.000	-0.001	-0.001	0.006	-0.007*	-0.001	-0.001	0.000	0.000	-0.002
	-1.03	0.34	0.06	-0.98	-1.14	0.56	-1.96	-0.65	-1.54	0.01	0.02	-1.65
R_VOL	1.98***	-5.40	0.12	2.39***	2.14***	-4.35	0.92	2.02**	1.99***	-3.52	0.05	3.44***
	3.55	-1.61	0.07	2.82	3.99	-1.44	0.49	2.48	3.49	-1.13	0.03	4.03
R_MCAP	-9.86	23.9	-4.04	-5.94	-12.07***	19.74	-1.89	-6.53	-9.78*	25.0	-5.44	-11.87*
	-1.98	0.28	-0.34	-0.92	-2.59	0.25	-0.16	-1.01	-1.92	0.3	-0.48	-1.79
r-square	7%	8%	14%	14%	7%	7%	6%	13%	3%	2%	14%	6%

2.7 Concluding Remarks

Using a reconstructed intraday series of S&P 500 index, this paper presents a unique microstructural perspective of ETF on its ever changing role. *SPY* and *IVV*, the two ETF tracking S&P 500, make up nearly half of the price discovery in the spot markets comprised of the trading of all component stocks and ETF. The trading of *SPY* and *IVV* is as much private information driven as individual stocks in the index are. The two major findings outline the growing importance of ETF in the information dissemination process in the equity market.

From the empirics, we can see that the transformation has far from being complete. Therefore, the future research should continue observing the relationship between ETF and its underlying index and update on the existing literature.

The findings in my paper also reveal a mixing role of ETF. They are basket securities but in the meantime the information contents in their trading are not much different from individual stocks. A theoretical model examining the interaction between ETF and index component stocks is indeed needed for a better understanding of ETF.

CHAPTER 3: INSTITUTIONAL FLOW AND INFORMATION

3.1 Introduction

The discovery of large amount of information flow induced by the trading of ETF is seen as an increasingly popular active investment strategy. From the perspective of promoting information dissemination and market efficiency, institutional investors are important players who can exert significant impacts through their large holdings and huge volumes of trading. They observe the flow of information on the market in a very different manner than individual investors and trade accordingly. Chordia, Roll, and Subrahmanyam (2011) observe an increase in market efficiency for those stocks widely held by institutions. They attribute the decrease in intraday volatility and price conforming more closely to random walk to a more effective trading by institutions on their private information. Hendershott, Jones, and Menkveld (2011) demonstrate that automated algorithmic trading greatly facilitate the trading for institutions. Considering a higher than average ownership by institutions, the trading of ETF is believed to strengthen the information flow to the ETFs' underlying indexes.

However, the hazard of a potential manipulation by the institutional investors exists. The problem is particularly acute in ETF trading because of their large impacts on the prices and thus the overall market indices. In a new study of the tech bubble at the end of the 20th century, Griffin et al. (2011) provides evidence on that institution' trading behavior could be very disruptive and causing harm to the market. They show that institutions speculatively purchased overpriced tech stocks until a coordinated broad sell-off without a justified expectation of future price appreciation consistent with fundamentals. Individual investors, on the other hand, are victimized by continuously buying while institutions are selling.

Of all the stocks we could consider to investigate the issue, ETF stands out as particularly interesting securities as far as we are concerned with institutional ownership. In Chapter 1 we find that ETF is transforming from a pure passive investment vehicle incepted with indexing as the main use to a popular choice of incorporating private information. Institutions have reasons to love ETF. They are perhaps the most liquid securities with huge trading volumes. They are ideal for portfolio indexing or hedging purposes. And most recently, they become an increasingly important source of information origination. Studying the institutional flow of ETF would also provide us a better understanding of the role of ETF in the market.

To answer the question whether institutions conduct information-based or speculative trading and its implications on the market, we examine the relationship between institutional trading flow and the information nature of trading. We hypothesize that: 1) a strong link between the trading flow and the proportion of private information-based trading is an indication of institutional investors' incorporating private information in their trading; 2) a weak or nonexistent link would mean a likely 'noise trader' role for institutions.

Using the same set of data, we also test whether institutions behave differently when trading ETF. Chapter 1 shows that the proportion of private information-based trading in ETF is as much as in a selected group of S&P 500 component stocks. This implies that the overall information quality in ETF trading is not fundamentally different in other stocks. However, institutional investors have more motivations to trade ETF than individual investors and some of them are not driven by information. We hypothesize that the link between institutional flow and information contents is stronger among non-ETF stocks. Observing the alternative that they behave similarly in terms of relationship would be further proof that ETF is utilized by institutional investors to profit from private information. In general, we want to contribute to not only the high frequency data literature on institutional trading but also analyze the evolving ETF.

The paper is organized as follows. In the follow section we briefly review the literature related to institutional trading and describe TAQ and 13-F, the two databases we use, in details. In section 3.3, we compute and discuss the trading imbalance which is used in constructing daily flow. In section 3.4, the daily institutional flows are estimated using the in-sample regression method. We then related it to the informational nature of trading in a set of structural and reduced form regressions in section 3.5. Additional analysis on the construction of institutional flow is performed in section 3.6. And the last section concludes.

3.2 Literature Review and Data Description

Most of the previous work using the 13-F filings to the SEC focus on the relationship of contemporaneous or lag returns and the ownership changes. In the influential work by Nofsinger and Sias (1999), they identify herding and feedback trading as a common practice among institutional investors. Excessive buying or selling may cause the prices deviating from their fundamental level and destabilize the market. The momentum-chasing behavior of institutions is documented using intraday frequency data in addition to the often used quarterly data. Griffin, Harris, and Topaloglu (2003) study the Nasdaq 100 and find that institutions chase short-term price movements while individual investors are in the opposite trading positions. In the other direction, the trading of institutional investors has little impact on the future returns economically. Neither are they able to find evidence of price reversal which could result from institutions' trend chasing behavior.

However, the quarterly data is inadequate to examine any kinds of microstructural behavior of institutions. Recently researchers start using high frequency data to infer the trading behavior of institutional investors. Puckett and Yan (2011), relying on a large proprietary database, find institutions earn significant abnormal returns intra-quarterly, suggesting that they are able to capture short-lived profitable trading opportunities. It supports the argument that managers do possess superior stock-picking abilities to some extent. Without proprietary data, other researchers attempt to construct institutional trades from the Trade and Quote (TAQ) database. Lee and Radhakrishna (2000) adopt a cutoff criterion that defines a trade initiated by institutions if the trade size is above a certain threshold. Campbell, Ramadorai, and Schwarts (2009) propose a finer method to infer institutional trading flows. They argue that trades of different sizes have disproportional institutional trading and one can therefore obtain a more accurate description of the flow. They divide trades into twenty bins based on their dollar volumes with preset bounds. The same quarter trading imbalance of classified trades of each bin is aggregated and used as explanatory variable in a regression with the quarterly changes from 13-F as the regressand. The obtained estimates are then used to compute the unobservable daily institutional flow.

We use Hasbrouck (1991b) to analyze the information contents of trading. He assumes that prices follow a random walk process and can be deviated by informed trading. In a Vector Autoregression Error Correction (VECM) model consisting of quote midpoints and trade direction indicators, the total variance of the innovation term is then divided into the proportion that is contributed by the quotes or by the trades. The latter is then defined as the private information trading.

Our work is the first to associate institutional ownership changes with the information dissemination mechanism with high frequency data that is better suited to analyze information related microstructure topics. We study the dynamics of institutional flow and information in a structural Vector Autoregression (VAR) model. We also add more evidence of the transforming nature of ETF by discussing the role of institutional investors.

This paper uses two databases, TAQ (Trade and Quote) and 13-F Institutional Holdings from Thomson-Reuters (formerly CDA/Spectrum). The intraday trades and quotes that are needed to compute daily imbalance come from TAQ. Following the widely used algorithm proposed in Lee and Ready (1991), we define a trade as buyer-initiated (seller-initiated) when its execution price is higher (lower) than the quote midpoint. Usually trades that are executed exactly on midpoint are classified using a tick test. Since the number of daily trades increase dramatically in recent years, there is a huge quantity of trades of which the trade direction cannot be identified without tick test. But in order to precisely compute daily imbalance, we avoid any ambiguity by considering those trades as unclassifiable.

Following Campbell et al. (2009), for each stock studied, we first exclude the trades that are labeled as 'bundled or 'split' in TAQ and those of which the timestamp is before 10:00am because trades size is an important variable to construct daily institutional ownership change. In the first half hour of trading, the price is usually determined by opening auction in which the recording of trade size is not reliable just as in cases when trades are bundled or split. These observations along with those of which the direction of trades need to be determined by tick test are together defined as unclassifiable.

TAQ is also used to decompose daily trades into private or public information driven. Following Hasbrouck (1991), we study quote prices and trade directions (buyer or seller initiated) in a Vector Autoregression (*VAR*) system to determine how much of the total variance of innovation (*VT*) to the underlying price can be attributed to trade related, which is then defined as the proportion of private information (*PR*). The values of *VT* and *PR* are recalculated since the sample period is different from in Chapter 1.

13-F provides us the quarterly changes in institutional ownership. There are many kinds of inconsistencies in 13-F. We have done a lot of cleaning. 1) The current quarter ownership is not equal to lag ownership plus change. We assume that the ownership number not the change is correct; 2) multiple entries for the same manager-stock-quarter observation. We retain only the first record in the original order; 3) there are missing stock-quarter observations for the same manager. We drop those unreported quarters from the sample. There are 3,992 occasions in which the gap is one quarter and 1,033 two quarters, compared to 108,976 occasions in which there is no missing observation. We could fill out the data using the average of the two months preceding and following the missing quarter. However, since we are interested in the aggregate institutional ownership changes, there is no point of assuming any change for a single institution.

3.3 Imbalance and Holdings

The sample we choose is from 2002 to 2009. *IVV* debuted in 2000. The first few quarters of trading data from IVV is very volatile so we want to pick up a starting point where things are stabilized. The studies period covers the financial crisis and may provide us an opportunity to examine any potential structural breaks the data presents.

Our first step is to identify the volume by institutional investors from all daily trading. Griffin, Harris, and Topaloglu (2003) rely on the proprietary data on Nasdaq 100 which classifies the originator of both sides of all trades as an individual, an institution or a market maker. Our method is based on Campbell et al. (2009). Their proposed specification allows separate coefficients for each dollar-volume bin:

$$\Delta H_{it} = \propto + \rho \Delta H_{it-1} + \phi H_{it-1} + \beta_{b_1} B_{1it} + \cdots + \beta_{b_n} B_{nit} + \beta_U U_{it} + \varepsilon_{it} \quad (3.1)$$

where the dependent variable ΔH_{it} is the quarterly changes in institutional ownership[1] of the stock from 13-F. Regressors include lag quarterly change ΔH_{it-1}, lag quarterly ownership level H_{it-1}. The remaining regressors are aggregated unclassified total trade volume (U_{it}), and aggregated trade imbalances[2] for all bins ($B_{1it}\ to\ B_{nit}$) from all trading days in a quarter.

We make a few modifications. First, their bins are in the units of dollar volumes. Considering the cross-sectional price level can vary in a wide range, we argue that it would be more accurate to use trading volume normalized by the number of outstanding shares. For example, a $10,000 transaction for a relatively small stock is more likely to be initiated by an institution than a $10,000 transaction for a large stock. Second, their sample used for estimation pools all stocks together and then the estimates are used to compute daily flow of each stock. To improve, we estimate the regression for each stock and obtain separate sets of estimates to be used to compute the daily flow for that particular stock. This also addresses the concern that the relationship between total flow and institutional flow may be different for stocks. Third, Campbell et al. eventually estimate a transformation function that smooth out the trade flows between bins, because there is lack of large trades for small stocks. All the stocks we consider are large-caps from S&P 500 and therefore we stick to the linear model (3.1).

For each quarter, we sort each classified buy/sell into a certain number of bins based on normalized trading volumes and then compute the daily imbalance for each bin. It is however difficult to determine the number of bins and the length of each bin. With results not being shown here, we experimented with many settings with the number of bins up to ten. However, the addition of more explanatory variables results in too many imprecisely estimated parameters and does not make additional contribution to the explanatory power of the model. We settle for simply using three bins, with the bottom 20% as the 'small' trades, top 20% as the 'large' trades, and those in between the intermediate ones. Normally, we consider the combined intermediate and large trades are initiated by institutions. In later section, we would also utilize the rule used in Lee and Radhakrishna (2000) and set the cutoff point at the smallest 20% in terms of trade size and other settings.

Panel A of Table 3.1 shows the mean of quarterly imbalance (*imb*) of each stock-bin aggregated from daily imbalance. Imbalances are positive for all stock-bins and increase monotonically with the larger bins for non-ETF, while the largest imbalance appears in the intermediate bin for ETF. To take into account the fact that the average trade size of a higher quintile is always larger, we standardize each quarter's imbalance by dividing the quarterly average size of a trade within the bin (*sdd_imb*). All of the six stocks have the largest standardized imbalance in the intermediate bin except for *XOM*. Larger imbalance possibly indicates that market participants' opinions are more aligned. So for the small and large bins where the concentration of individual and institutional investors is high, there is a higher degree of heterogeneity of the trade direction than for the intermediate bins. It is also in line with the findings of Chakravarty (2001) and Barclay and Warner (1993) that institutional investors prefer medium-sized trades to minimize price impacts. In Panel B, we examine the correlations of imbalances across stocks for each bin. It is easy to spot that correlation of intermediate bins is largest. It is more interesting to note that negative almost all the negative correlations are related to the smallest bin of *CEG* and the largest bin of *SPY*. Panel C shows the correlations of imbalances across bins for each stock. It is not surprising that the intermediate and large bins have the largest correlations where the composition of traders is similar (populated with institutional investors).

[1] All ownership variables, without specifically indicated, are measured in percentages of the outstanding number of shares.
[2] Imbalances are computed as buys minus sells.

Table 3.1
Imbalance

This table presents the statistics for trading imbalance during the sample period of 2002 to 2009. All trades are categorized based on their sizes relative to the number of outstanding shares in a particular quarter. The 'small' bin contains bottom 20 percent of trades sizes; the 'large' bin contains top 20 percent; and the 'intermediate' bin has all 60 percent in the middle. Imbalance (imb) is defined as buying volume minus selling in percentage numbers relative to the number of outstanding shares. In Panel A, sdd_imb is the imbalance in each bin standardized by the average size of a trade in that bin. The last two columns show the standardized trade sized of buy and sell orders classified using Lee and Ready (1991). Panel B presents Pearson Correlation of imb across stocks for each bin and Panel C across bins for each stock.

Panel A: Mean and standardized imbalance

	bin level	imb	sdd_imb	sdd_buy	sdd_sell
AFL	small	0.02	973	87,182	86,209
	intermediate	0.18	2,088	210,163	208,075
	large	0.90	1,329	73,529	72,200
CEG	small	0.04	659	52,059	51,401
	intermediate	0.22	1,281	105,677	104,396
	large	1.27	1,001	40,304	39,303
IVV	small	0.03	-234	41,922	42,156
	intermediate	0.63	1,679	123,592	121,912
	large	0.55	-463	39,747	40,210
SPY	small	0.04	1,440	1,489,571	1,488,131
	intermediate	0.32	4,896	4,061,787	4,056,891
	large	0.08	992	1,328,719	1,327,727
SVU	small	0.22	3,428	43,789	40,361
	intermediate	1.23	7,624	136,325	128,702
	large	4.45	4,304	43,759	39,455
XOM	small	0.01	7,809	339,574	331,765
	intermediate	0.05	-11,095	1,187,187	1,198,282
	large	0.30	-12,175	369,437	381,612

Panel B: Correlations across stocks

`	*AFL*	*CEG*	*IVV*	*SPY*	*SVU*	
CEG	-0.39					small
	0.37					intermediate
	0.56					large
IVV	0.57	-0.27				small
	0.50	0.38				intermediate
	0.15	0.15				large
SPY	0.16	-0.54	0.20			small
	0.27	-0.11	0.27			intermediate
	-0.20	-0.36	-0.07			large
SVU	0.18	-0.16	-0.02	0.33		small
	0.57	0.62	0.44	0.31		intermediate
	0.35	0.39	0.29	-0.17		large
XOM	0.65	-0.35	0.40	0.00	0.49	small
	0.80	0.42	0.48	0.38	0.74	intermediate
	0.75	0.45	0.26	-0.21	0.45	large

Table 3.1 continued

Panel C: Correlations across bins

	small-intermediate	small-large	intermediate-large
AFL	0.31	0.41	0.8
CEG	-0.03	0.1	0.62
IVV	0.44	0.31	0.5
SPY	0.61	-0.11	0.2
SVU	0.28	0.2	0.82
XOM	0.49	0.58	0.87

Figure 3.1 shows the evolution of imbalance and the almost identical standardized imbalance for each stock-bin. For all stock-bins except for large bin of *SPY*, the volatility of imbalance or standardized imbalance substantially increases in recent year with a downward trend in level. The resembling patterns prove that there are some common time factors that affect the order imbalance for all stocks. The larger swing is mostly likely the result of dramatic increase in trading activities which is documented in Chordia, Roll, and Subrahmanyam (2011). The average imbalance moving from positive to negative implies that the trading pressure shifts from buyers to sellers and seemingly coincides with the returns of the overall market in recent years.

Table 3.2
Institutional Ownership

This table presents institutional holdings (H) related statistics. The quarterly H is calculated from 13-F database. The daily imbalance (imb) and standardized imbalance (sdd_imb) are computed from TAQ database and then aggregated over a quarter. Sample period is 2002 to 2009. Panel A shows the mean and standard deviations of H. Panel B shows the Pearson correlations of the two measures of imbalance. Panel C shows the correlations between H and imb/sdd_imb.

Panel A: Institutional holdings from 13-F (H)

	AFL	*CEG*	*IVV*	*SPY*	*SVU*	*XOM*
mean (%)	58.3	65.4	41.5	60	84.9	49.6
s.t.d. (%)	4.3	4.6	7.5	17.3	11.1	2.4

Panel B: Correlations of measures of imbalance (imb, sdd_imb)

AFL	*CEG*	*IVV*	*SPY*	*SVU*	*XOM*
0.52	0.46	0.59	0.59	0.33	0.64

Panel C: Correlations of quarterly holdings and imbalance

	AFL	*CEG*	*IVV*	*SPY*	*SVU*	*XOM*
imbalance	-0.47	-0.2	0.13	0.22	0.09	-0.16
standardized imbalance	-0.42	-0.19	0.14	-0.02	0.08	-0.18

In Panel A of Table 3.2, computed from 13-F data, the average holdings by institutions (*H*) during the studied period vary from 41% to 85% with *IVV* the lowest. Figure 3.2 shows that except for *IVV*, the institutional ownership has increased over the years. Panel B presents the correlations of the two measures of trading imbalance. The correlation coefficient between standardized imbalance and unstandardized ranges from 0.33 to 0.63, indicating that we should be careful of interpreting using the raw imbalance. Nevertheless, there is not much difference between the two when being linked to holdings. The correlation between *imb* and *H* in Panel C is positive for the two ETF along with the smallest benchmark stock *SVU*. For *sdd_imb*, the correlation for *SPY* becomes slightly negative. All other three have negative correlations.

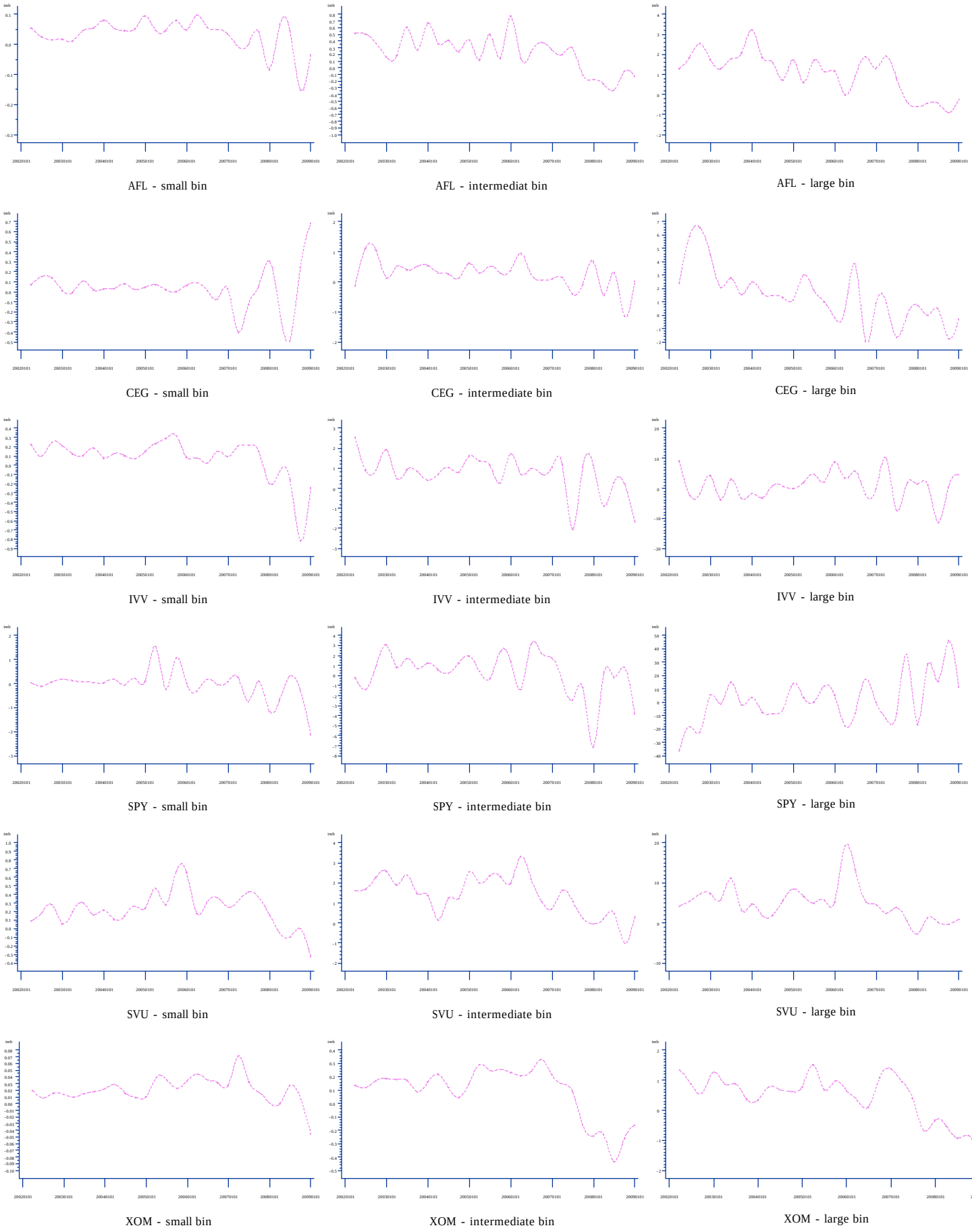

Figure 3.1
Imbalance 2002 to 2009

The figures plot the trading imbalance of each stock-bin combination. Trades are sorted by sizes relative to outstanding shares in each stock-quarter. Top and bottom 20 percent are defined as small and large bins and the rest as intermediate. The studied period is from 2002 to 2009

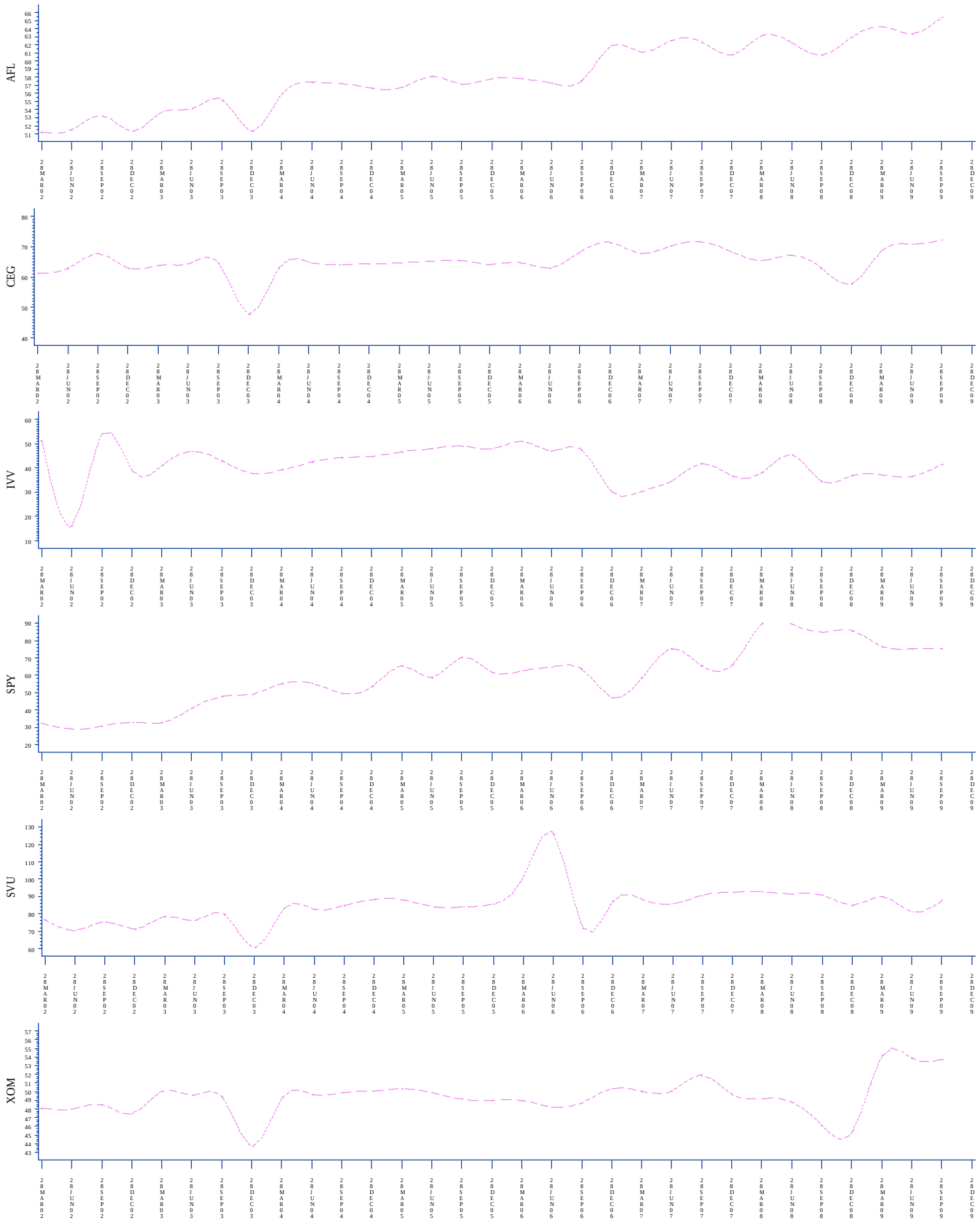

Figure 3.2
Institutional Ownership

The figures plot the institutional holdings in percentage (H) from 13-F during the sample period is from 2002 to 2009. Studied stocks are two S&P 500 ETF, SPY and IVV and four index component stocks, AFL, CEG, SVU and XOM.

Large order imbalance possibly signals homogeneity of market opinion or lower. It is possible that more institutional investors choose to hold *ETF* as indexing or hedging tools based on some macro-level information. On the other side, more institutions can imply a more divided opinion on individual stocks.

3.4 Daily Institutional Flow

In the estimated regression model (3.1), the changes in institutional holdings (ΔH) is explained by the right-side variables including quarterly aggregated imbalance by bins, unclassified total volume computed from TAQ, the lagged total holdings by institutions (*lag*H) and the lagged changes in holdings (*lag*ΔH). Here we assume that the coefficients of buy orders and sell orders are identical within the same bin. The regression can be seen as aggregating imbalances of all bins over all days in a quarter to see how they combine to explain the quarterly institutional flow. The estimated coefficients are then used to compute the fitted values at the daily level. In Table 3.3 we present the results of estimation based on different bin settings, *large-intermediate-small*, *large-small* along with those based on no division at all and on deciles.

None of the coefficients on bin imbalances is precisely estimated. However, the *F*-stats show that jointly the coefficients on the independent variables are not zero except for *SPY*. Multicollinearity tests based on *variance inflation factor* or Tolerance not shown here exclude the possibility of existence of high degree of collinearity among explanatory variables. It is shown in a setting where there are ten equally divided bins the coefficients are then jointly significant for *SPY*, which implies that the institutional trading in *SPY* is more layered in terms of trading size than any other stocks. All *lagH* are significantly negatively correlated to *H,* which indicates the long-term mean-reversion of holdings by institutions. *lag*ΔH has mostly insignificant negative estimated coefficients. Unexpectedly the adjusted R^2 is smaller when we assume different coefficients on bins than identical coefficients on all bins except only for *SPY* and *XOM*. From here, we will construct the daily institutional flow.

Means and medians of the daily and quarterly flow are shown in Panel A of Table 3.4. The largest and smallest estimated flows usually appear in the setting with no separate bins or with 10-bins. The mean daily institutional flow estimations are positive but we cannot reject the null hypothesis that they are zero. The mean is suspiciously larger for *SPY* compared to other stocks, possibly due to more trading interests in *SPY* from institutional investors. Panel B presents the Pearson correlations between different measures of daily institutional flow by stocks. In the case of *CEG* and *IVV*, all correlations are larger than 0.96. For the other stocks, 3-bins settings have on average the highest correlations with other measures except for *XOM*. It seems that the ideal number of bins is different for each stock. However, it is likely to be safe using the 3-bins setting as its estimates are highly correlated with those of other settings.

Figure 3.3 plots the aggregate quarterly institutional flow over the time. For all stocks, there is a steady increase in the flow and a peak appearing sometime around the first quarter of 2009. It is interesting that these evolutions mirror those of imbalances in Figure 3.1. In other words, the imbalances are falling, while institutional investors more actively increase their holdings. Reconciliation could be that institutional investors hold more divided opinions than individual investors.

Table 3.3
Flow Estimation

This table presents the estimates of the regression (3.1), where the dependent variable is the quarterly changes from 13-F. Regressors include lag quarterly change (lag ΔH), lag quarterly ownership level (lagH), daily unclassified total trade volume (U), and daily trade imbalances for all bins. The number of bins in different regression settings are three, two, one and ten respectively. In 3-bins setting, 'small' contains the bottom 20 percent, 'large' the top 20 percent, and 'intermediate' the middle. In 2-bins setting, 'small' the bottom 20 percent and 'large' the rest. Deciles are represented in the 10-bins setting. 'One universal bin' does not differentiate trade sizes. Significance level are signified in shades. *, **, *** denotes significance at 90%, 95%, and 99% level respectively.

	AFL	*CEG*	*IVV*	*SPY*	*SVU*	*XOM*
3-bins by trade size						
adj. R-square	0.35	0.42	0.48	0.03	0.44	0.19
lagH	-0.17	-0.8***	-1.07***	-0.27**	-0.42*	-0.59**
lagΔH	-0.38**	0.12	-0.03	-0.16	-0.29	-0.08
small	4.94	-6.39	5.64	2.21	2.37	-25.07
intermediate	0.48	0.51	2.2	-0.7	-0.47	-4.64
large	-0.34	-0.61	-0.01	0.13	-0.71	1.87
unclassed	0.06	0.02	0.02	0.01	-0.01	0.16
2-bins by trade size						
adj. R-square	0.34	0.41	0.49	0.04	0.44	0.18
lagH	-0.16	-0.78***	-1.04***	-0.29**	-0.42**	-0.63**
lagΔH	-0.4**	0.09	-0.01	-0.13	-0.29	-0.01
small	4.54	-6.22	4.12	1.12	2.58	-23.97
large	-0.19	-0.46	0.18	0.08	-0.69	0.72
unclassified	0.06	0	-0.04	0.02	-0.02	0.18
one universal bin						
adj. R-square	0.31	0.39	0.5	0.08	0.43	0.13
lagH	-0.18	-0.71***	-1.08***	-0.27**	-0.41*	-0.52**
lagΔH	-0.39**	0.06	-0.01	-0.13	-0.3*	0.03
imbalance	-0.16	-0.66	0.24	0.09	-0.53	0
unclassified	0.04	-0.08	-0.08	0.13	0	0.12
deciles by trade size						
adj. R-square	0.23	0.11	0.37	0.74	0.57	0.72
lagH	-0.25*	-0.94**	-1.18***	-0.28**	-0.42	-0.73**
lagΔH	-0.27	0.26	0.02	0.06	-0.28	-0.12
b_1 (smallest)	-2.85	-0.67	-3.53	20.32***	17.32	21.44
b_2	11.01	-7.6	9.21	5.22*	1.08	0.73
b_3	19.67	-12.3	18.91	8.68	1.13	4.61
b_4	-1.44	15.59	-4.38	-0.66	2.75	-13.84
b_5	-29.63**	-5.97	3.38	1.42	-12.16	-19.55
b_6	5.33	2.69	7.15	-8.69**	0.55	40.58**
b_7	-1.68	0.11	-9.3	-3.95	-0.47	-18.89
b_8	7.61	4.83	-4.2	4.28**	4.38	0.53
b_9	1.67	2.22	5.56	-2.38	-5.84	-16.68
bin_10 (largest)	-1.15	-1.16	-0.12	0.19*	-0.19	1.6
unclassified	0.13	0.1	-0.07	0	0.04	0.01*

Table 3.4
Institutional Flow

Panel A of this table presents the mean and median statistics for the estimated institutional flows from regressions (3.1), where the dependent variable is the quarterly changes from 13-F. Regressors include lag quarterly change (lagΔH), lag quarterly ownership level (lagH), daily unclassified total trade volume (U), and daily trade imbalances for all bins. The fitted values are calculated daily excluding quarterly variables (lagH, lagΔH).The quarterly flow is aggregated using the daily fitted values. In 3-bins setting, 'small' contains the bottom 20 percent, 'large' the top 20 percent, and 'intermediate' the middle. In 2-bins setting, 'small' the bottom 20 percent and 'large' the rest. Deciles are represented in the 10-bins setting. '1-bin' does not differentiate trade sizes. Panel B presents the correlations of different measures of daily flows.

Panel A: Estimated flows

	AFL	*CEG*	*IVV*	*SPY*	*SVU*	*XOM*
	daily mean					
3-bins	0.03	0.18	0.34	2.61	0.3	0.03
2-bins	0.04	0.2	0.27	1.17	0.26	0.02
1-bin	0.05	0.1	0.12	1.18	0.28	0.05
10-bins	0.03	0.13	0.31	5.08	0.31	0.01
	daily median					
3-bins	0.02	0.1	0.2	1.75	0.2	0.02
2-bins	0.02	0.11	0.15	0.74	0.16	0.02
1-bin	0.03	0.06	0.07	0.74	0.18	0.04
10-bins	0.02	0.07	0.18	3.49	0.21	0.01
	quarterly mean					
3-bins	2.3	11.18	21.66	164.42	18.95	1.63
2-bins	2.38	12.32	16.91	73.9	16.16	1.35
1-bin	3.36	6.32	7.45	74.35	17.93	3
10-bins	1.74	7.92	19.6	320.16	19.5	0.79
	quarterly median					
3-bins	2.03	9.37	14.48	122.52	17.27	1.46
2-bins	2.07	9.97	10.91	57.69	13.61	1.25
1-bin	3.03	5.13	5.7	57.37	15.43	2.92
10-bins	1.51	6.89	13.17	234.34	18.48	0.79

Panel B: Correlations of estimated daily flows

	AFL	*CEG*	*IVV*	*SPY*	*SVU*	*XOM*
3-bins/2-bins	0.997	0.998	0.992	0.961	0.984	0.938
3-bins/1-bin	0.979	0.996	0.991	0.974	0.986	0.823
3-bins/10-bins	0.937	0.996	0.999	0.99	0.968	0.904
2-bins/1-bin	0.98	0.999	0.969	0.999	0.999	0.967
2-bins/10-bins	0.932	0.99	0.994	0.914	0.915	0.95
1-bin/10-bins	0.973	0.988	0.989	0.934	0.92	0.929

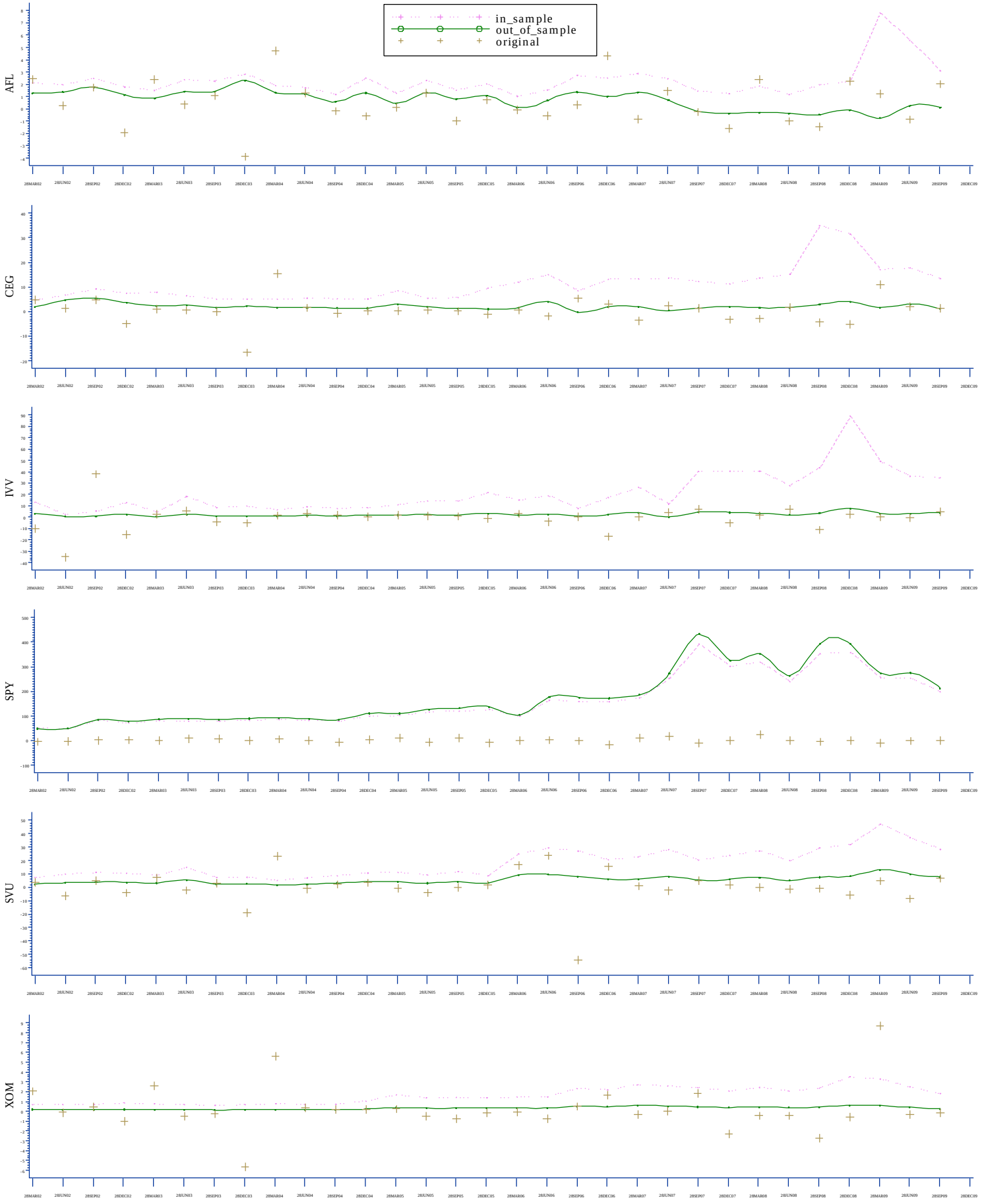

Figure 3.3
Quarterly Institutional Flow

The figures plot the estimated quarterly institutional flows in-sample and out-of-sample methods along with the quarterly ownership changes computed from 13-F in the sample period 2002 to 2009. Out-of-sample estimation is conducted on a rolling basis. The initial estimation period is 2002 to 2003. The estimates are then used to fit the regression for the 1st quarter of 2004. The daily flows are then aggregated to quarterly level and used to replace the original variable from the 13-F database. The newly constructed flow is then used to expand the sample and the procedures are repeated until we reach the last quarter of 2009.

3.5 Institutional Trading and Information

The motives to hold equity or change holdings can be different for institutional and individual investors, which in turn may be reflected in the information contents of trading. Institutions in many ways are believed to possess more private information than individual investors and their trading behavior is more informative. We hypothesize that, in the days when the magnitude of institutions changing their positions is larger, the trading is more likely to be private information driven.

It is also a good opportunity to examine how close the motive on trading ETF is to that on trading common stocks. If it is observed that the nature of the relationship between institutional trading flow and informational trading is the same for ETF and benchmark stocks, we may conclude that investors trade ETF behind similar motives.

We measure the trading intensity (*flow*) of institutions by taking the absolute value of the daily trading flow (ΔH). The trading flow is not symmetric as we have seen that the mean is positive and there is an increasing trend in recent years. The new information that arrives in the market and motivates institutional trading is possibly not the same when institutions increase or decrease one percent of their holdings. Nevertheless, most of the daily changes in institutional holdings are positive. Conversely, the trading pattern of institutional investors can possibly predict future composition of information of the market, as it is possibly driving the price level. To study the information-trading dynamics, we introduce two variables from the first chapter. *PR* is the proportion of private information in a trading day, measured by the contribution to the total variance (VT) of trade-related price movement in a Vector Error Correction Model (VECM) consisting of trades and prices.

$$flow_t = \alpha + \beta_0 PR_t + \sum_{i=1}^{5} \beta_i PR_{t-i} + \sum_{i=1}^{5} \gamma_i flow_{t-i} + \varepsilon_{t,flow} \quad (3.2)$$

$$PR_t = \alpha + \sum_{i=1}^{5} \beta_i PR_{t-i} + \sum_{i=1}^{5} \gamma_i flow_{t-i} + \varepsilon_{t,PR} \quad (3.3)$$

In the structural *Vector Autoregression* (VAR) system shown above that consists of the magnitude of institutional flow (*flow*) and *PR*, we treat both variables as endogenous and each is regressed on up to five lags of both. *PR* is allowed to have a contemporaneous impact on *flow*. The structural regressions are performed on the two ETF and four component stocks of S&P 500. We also estimate the reduced form regression for the same VAR.

Panel A of Table 3.5 shows the estimation results. For equation (3.2), all coefficients of the contemporaneous *PR* are negative and precisely estimated except for *IVV* of which the coefficient is not significant. Therefore, for those days institutional investors in overall decide to not change their positions in these stocks, the proportion of private-information based trading is actually higher. Indeed we observe a strong correlation between institutional flow and information contents. But this is the opposite of what we hypothesize about the sign. It is usually assumed that institutional investors possess more private information and incorporate the information in the trading. If we hold on to this assumption, possible reconciliation to the paradox are that the private information held by different institutions is different or opinions on the same information among institutions remain widely differentiated so that the imbalance is reduced as institutions trade against each other. The *PR* coefficient of *SPY* (-1.34) is much larger than that of other stocks in terms of magnitude. All of the coefficients on the lagged absolute flows are statistically positive except for one, showing some degree of autocorrelation of overall institutional flow magnitude. The 5th lag of *flow* comes with coefficient of much larger magnitude than other lags. Since the regression is conducted in daily frequency, there is likely to be weekday effects on the institutional flow. The first two lags of *PR* are mostly negatively correlated to *flow* while the signs on larger *PR* lags are mostly positive. The

Table 3.5
Flow and Information

Panel A contains the estimates of a structural Vector Autoregression of daily institutional trading flow (in absolute value) and PR, the proportion of private-information based trading in a day with five lags. PR is allowed to have a contemporaneous impact on flow. Reduced form estimates are only partially presented for regressions with flow as dependent variable. Estimates with statistical significance are represented in different shades. Panel B presents the estimates from a VAR consisting of flow, PR, and VT, the daily total variance of the random walk of the underlying price. The structural VAR allows contemporaneous effects of VT on flow. Estimated coefficients of flow, PR and their lag values are omitted. *, **, *** denotes significance at 90%, 95%, and 99% level respectively.

Panel A: Two-variable VAR

dependent variable	*flow*						*PR*					
	AFL	CEG	IVV	SPY	SVU	XOM	AFL	CEG	IVV	SPY	SVU	XOM
PR	-0.11***	-0.14***	0.07	-1.34***	-0.33***	-0.07**						
PR_lag1	0.00	-0.01	0.13	-1.28	0.06	-0.06	0.26***	0.25***	0.22***	0.25***	0.2***	0.33***
(reduced form)	-0.03	-0.05	0.15	-1.62	-0.01	-0.08**						
PR_lag2	0.00	0.05	0.14	0.15	0.03	-0.02	0.17***	0.18***	0.22***	0.23***	0.25***	0.25***
(reduced form)	-0.02	0.03	0.16	-0.16	-0.05	-0.04						
PR_lag3	0.03	-0.06	0.16	1.13	-0.15	0.03	0.18***	0.14***	0.15***	0.19***	0.12***	0.13***
(reduced form)	0.01	-0.07	0.17	0.87	-0.19	0.02						
PR_lag4	0.03	-0.04	0.04	0.25	0.18	-0.03	0.18***	0.17***	0.24***	0.1***	0.13***	0.08***
(reduced form)	0.01	-0.06	0.06	0.12	0.13	-0.03						
PR_lag5	0.07**	0.05	-0.16	0.22	0.07	0.12	0.12***	0.15***	0.11***	0.16***	0.18***	0.14***
(reduced form)	0.06	0.03	-0.15	0.01	0.02	0.11						
flow_lag1	0.11***	0.23***	0.04*	-0.01	0.08***	0.01	0.01	-0.02*	0.00	0.00	0.00	0.00
flow_lag2	0.07***	0.13***	0.01	0.04**	0.09***	0.00	0.02	0.00	-0.004**	0.00	0.00	0.06***
flow_lag3	0.07***	-0.01	0.00	0.03*	0.06***	0.01	0.02	0.01	0.01**	0.00	0.00	0.03
flow_lag4	0.12***	-0.00255	0.11***	0.12***	0.11***	0.07***	-0.01	0.00	0.01	0.00	0.00	-0.03
flow_lag5	0.33***	0.31***	0.19***	0.62***	0.47***	0.41***	0.02	-0.01	0.00	0.00	0.00	0.03
intercept	0.01***	0.05***	0.08***	0.34***	0.07***	0.03***	0.01***	0.02***	0.00	0.01***	0.01***	0.00

Sample Mean

PR						*VT (in bps)*					
AFL	CEG	IVV	SPY	SVU	XOM	AFL	CEG	IVV	SPY	SVU	XOM
14.15%	12.97%	7.42%	12.40%	12.91%	11.54%	10	18	382	11	8	3

Table 3.5 continued

Panel B: 3-variable VAR

		VT	*VT_lag1*	*VT_lag2*	*VT_lag3*	*VT_lag4*	*VT_lag5*
flow	AFL	0.22**	-0.17*	0.01	0.03	-0.10	0.03
	CEG	-0.21	0.33***	0.77***	0.78***	-0.14	-0.92***
	IVV	0.01	0.03	-0.03	-0.03	-0.10	0.12*
	SPY	1.02	0.44	-0.69	-2.53*	-2.09	-1.61
	SVU	0.94*	-1.05*	-0.88	-0.68	-0.77	0.51
	XOM	0.44**	-0.07	0.34	0.02	-0.19	0.06
PR	AFL		-0.06	0.01	0.04	-0.03	0.16**
	CEG		-0.04	0.05	-0.11	0.05	-0.02
	IVV		0.03	0.00	-0.02*	0.02*	-0.04***
	SPY		0.05	0.03	0.05	-0.04	-0.02
	SVU		0.02	0.04	-0.02	0.13	-0.06
	XOM		0.24	0.28*	0.35**	-0.54***	0.45***
VT	AFL		0.19***	0.16***	0.11***	0.14***	0.1***
	CEG		0.34***	0.05***	0.08***	0.05***	0.16***
	IVV		0.3***	0.17***	0.17***	0.11***	0.2***
	SPY		0.08***	0.05**	0.02	0.02	0.02
	SVU		0.16***	0.13***	0.13***	0.13***	0.12***
	XOM		0.24***	0.08***	0.07***	0.21***	0.04***

reversal of the signs on lagged *PR* implies that it is possible that institutions refrain from trading against abundant information on the market up to two days and start reposition themselves in the trading of those stocks hereafter.

Results from the reduced form regressions are only partially shown because most of the estimated coefficients are identical. Even those that are different when *flow* is the dependent variable yield similar coefficients to the structural form estimations.

For equation (3.3), the current daily *PR* is regressed on lag values of *PR* and *flow* up to five lags. There is a strong autocorrelation of *PR*, suggesting that private information intensity tends to extend for a couple of days. The coefficients of *flow* on *PR* are mixed in signs with few significant estimates. Institutional investors are known to frequently trade against noise traders. If we believe that institutional investors engage in stealth trading to move stock price without revealing themselves, then it's not hard to see why their trading does not give away too much on the information side either.

$$flow_t = \alpha + \beta_0 PR_t + \gamma_0 VT_t + \sum_{i=1}^{5} \beta_i PR_{t-i} + \sum_{i=1}^{5} \gamma_i flow_{t-i} + \sum_{i=1}^{5} \gamma_i VT_{t-i} + \varepsilon_{t,flow} \quad (3.4)$$

$$PR_t = \alpha + \sum_{i=1}^{5} \beta_i PR_{t-i} + \sum_{i=1}^{5} \gamma_i flow_{t-i} + \sum_{i=1}^{5} \gamma_i VT_{t-i} + \varepsilon_{t,PR} \quad (3.5)$$

$$VT_t = \alpha + \sum_{i=1}^{5} \beta_i PR_{t-i} + \sum_{i=1}^{5} \gamma_i flow_{t-i} + \sum_{i=1}^{5} \gamma_i VT_{t-i} + \varepsilon_{t,VT} \quad (3.6)$$

We also explore the addition of *VT* in the endogenous variable list. As *VT* measures the total information intensity, it may be related to the perception of institutional investors of the entire market and their

corresponding trading behavior. Panel B of Table 5 presents selected estimates from the 3-variables-system regressions. We omit those coefficients on lag *flow* and *PR*, which are very close to those in the 2-variables system. Of five of the six stocks (*CEG* is the exception) and three with statistical significance, *VT* has a positive correlation with the contemporaneous daily flow. The opposite signs of *VT* and *PR* on *flow* indicate that institutions respond to public and private information very differently. The coefficients are mostly positive on recent lags of *VT* and negative on larger lags. In the regressions with *VT* as dependent variables, the autocorrelation coefficients are decreasing with larger lags but keeping the same signs.

The empirical results above reveal another important observation. Of the six stocks, *PR* ranges from 7.45% to 14.15%, with the lowest value in *IVV*. *VT* of *IVV* is 0.1311 and the second largest is 0.018[1]. The total information intensity of the daily trading in *IVV* is about almost ten times larger than the rest. The relative amount of private information in the trading (*PR*) of *IVV* is 7.5%, about at least 35% lower than the rest. Moreover, we find a strong relationship between institutional trading flows and *PR* in all but *IVV*. The standout of *IVV* against other stocks implies that *SPY* might resemble other stocks in terms of informational trading more than *IVV*, which is also tracking S&P 500.

Interestingly, *IVV* also has the lowest institutional holdings with 41%. It can be argued that the relatively small holdings by institution investors who possess more private information makes the trading of *IVV* fundamentally different from other stocks. A small *PR* means more speculative trading. For ETF, it also means that the trading is less information based and more likely for indexing or hedging purposes. Therefore, even though *IVV* and *SPY* are both ETF tracking the same index, the composition of market participants and the trading motives behind are quite different.

3.6 Additional Analysis

The estimation of daily flows of institution investors is critical for the analysis of its relationship with the nature of information in the trading. We first perform an out-of-sample estimation on a rolling basis with the regression method. The initial estimation period is 2002 to 2003. The estimates are then used to fit ΔH, the change in daily institutional ownership for the 1st quarter of 2004. We aggregate the daily flows to quarterly level and replace with these fitted values the original variable from the 13-F database. We then use the newly constructed ΔH to expand the sample and rerun the estimation. The procedures are repeated until we reach the last quarter of 2009.

Table 3.6 shows the summary statistics with in-sample and out-of-sample estimation in comparison. In terms of flow magnitude, the mean and standard deviation of daily flow are much smaller for the out-of-sample estimation except for *SPY*. However, for *AFL* and *CEG*, the out-of-sample values are only marginally correlated to the in-sample values, with correlation coefficient at 0.73 and 0.50 (larger than 0.91 for the other four). The plots of the daily flows in Figure 3.3 confirm both observations. We also present the correlation table of the aggregated quarterly daily flows and the original institutional ownership changes from 13-F. The correlation coefficients of in-sample and out-of-sample flows are now negative and even lower for *AFL* and *CEG* respectively (-0.13 and 0.23) and still strong for the rest. However, the raw ΔH data from 13-F is not correlated to the constructed quarterly flows at all. It should be noted that the quarterly variables (*lag* ΔH and *lag* H) that are used in the regression to estimate daily flows are actually left out when fitting the values. In the end, our purpose is to capture the daily dynamics and the accuracy of quarterly aggregated data is not of much concern. Nevertheless, it shows that we need to distinguish

[1] The summary statistics of *VT* and *PR* are slightly different from Chapter 1 because the sample period is different.

between different stocks when choosing the methodology for the linear estimation of institutional flows. Not shown here, the estimates from the regressions using the out-of-sample method in comparison with the in-sample method are qualitatively identical, suggesting that the dynamics of daily flows are captured in the same way for both estimation methods.

Table 3.6
Out-of-sample Estimation of Flow

This table compares the statisticis of the institutional flows estimated using in-sample and out-of-sampe methods. Out-of-sample estimation is conducted on a rolling basis. The initial estimation period is 2002 to 2003. The estimates are then used to fit ΔH, the change in daily institutional ownership for the 1st quarter of 2004. The daily flows are then aggregated to quarterly level and used to replace the original variable from the 13-F database. The newly constructed ΔH is then used to expand the sample and the procedures are repeated until we reach the last quarter of 2009. The first table shows means and standard deviations. The second table shows the two-way correlations.

	AFL	*CEG*	*IVV*	*SPY*	*SVU*	*XOM*
Means and standard deviations of quarterly flows						
in-sample	2.3	11.18	21.66	164.42	18.95	1.63
(s.t.d.)	1.3	7.04	18.05	99.83	11.15	0.85
out-of-sample	0.71	2.1	2.33	178.7	5.62	0.36
(s.t.d.)	0.76	1.21	1.5	110.31	2.83	0.15
Correlations of quarterly flows						
in-sample/out-of-sample	-0.129	0.226	0.901	0.999	0.977	0.968
in-sample/13-F	0.056	-0.158	0.054	-0.013	-0.057	0.163
out-of-sample/13-F	0.025	-0.162	0	-0.009	-0.024	0.168

We also conduct robustness tests to address the concern that the number of bins to separate trading size may also affect the estimation of institutional flows. The increase in the number of bins would leave many of the bins blank while no separation of trading size is expected to yield imprecise estimates of daily flows. The estimates from using the fitted values with the setting of ten bins or without separation of trade sizes are again qualitatively the same as the 3-bins setting. It seems that the dynamics is not sensitive to how we construct the daily institutional flows.

3.7 Concluding Remarks

Institutional investors have long been considered as 'informed' traders who use their information advantage to chase returns. Accordingly the past literature has been focusing on the relationship between return and trading in the microstructural framework. Our paper is the first to analyze the actual information quality of the institutional trading by examining a structural *VAR* model which is estimated from high frequency trading data. This is probably a more fundamental question than the return issue because it directly asks whether institutions make moves based on private information. We use 3-bins setting to classify trades based on sizes and infer daily flows transacted by institutional investors. We find that the daily ownership changes by institutions in five of the six studied stocks are negatively correlated to contemporaneous *PR*,

the proportion of private information-driven trading, and its recent two lags. There are two possible explanations: 1) in overall institutional investors are not synchronous with a high level of private information-based trading. There are divided opinions on the information among institutions, resulting in buy volumes cancelling sell volumes. Therefore, even though *PR* is high, the imbalance is low; 2) institutional investors prefer to camouflage themselves in stealth trading. They would rather trade against noise traders than other institutions. Moreover, with the presence of noise traders, it is difficult for outsiders to capture the information revealed by institutional investors' trading. Therefore, we observe large institutional flows when the amount of information is abundant on the market and observe small flows then the information is relatively scarce.

Even though we are short of a complete theory, the strong negative correlation between institutional flow and informational contents in the trading is sufficient to prove that the behavior of institutional investors is motivated by information in some way or another. In the meanwhile, we also show that institutions might have motives to trade ETF different from index stocks, especially for *IVV*.

Directions of future studies would rely on a more accurate high-frequency trading datasets of institutions. More importantly, we would need a theory to explain the intraday or short-term behavior of institutional investors related to the information reflected on the market. After all, empirically it is easy to show the interactions of trading and information but hard to prove the causality and therefore the motives behind institutional trading.

CHAPTER 4: CONCLUSIONS

This dissertation provides empirical evidence showing the interaction of information flows and the trading of various financial securities by different market participants. The first essay finds that ETF contributes significantly to the price formation of its underlying index and their contribution is positively correlated to how much their trading is driven by private information. The second essay proves that institutional investors are not likely noise traders. The direction of institutional daily volumes is negatively correlated to the proportion of private information in the overall trading. we also show that institutions might have motives to trade ETF different from index stocks.

Emphasis is on the comparisons between ETF and index component stocks. Similarities are revealed in the findings of the first essay while differences in those of the second essay. Nevertheless, we believe that ETF is transforming into similar roles played by index component stocks, in the sense that they are traded as part of active portfolio management.

A complete theoretical model would help us better understand the empirics regarding ETF's role in the price discovery and information distribution. The theory should parsimoniously describe the investors' choice trading ETF or component stocks based on the private information they possess. It also should distinguish the trading motivations between institutional and individual investors. Another direction of future studies would be the utilization of newly introduced ultra-high-frequency data. The new database would undoubtedly give researchers tools to analyze the evolving trading mechanism adopted by investors equipped with necessary technology.

REFERENCES

Ates, Aysegul, and George H. K. Wang, 2005, Information transmission in electronic versus open-outcry trading systems: an analysis of U.S. equity index futures markets, *Journal of Futures Markets* 25, 679-715.

Barclay, Michael J., and Terrence Hendershott, 2003, Price discovery and after trading hours, *Review of Financial Studies* 16, 1041-1073.

Barclay, Michael J., and Jerold B. Warner, 1993, Stealth trading and volatility: Which trades move prices?, *Journal of Financial Economics* 34, 281-305.

Campbell, John Y., Tarun Ramadorai, and Allie Schwartz, 2009, Caught on tape: Institutional trading, stock returns, and earnings announcements, *Journal of Financial Economics* 92, 66-91.

Chakravarty, Sugato, 2001, Stealth-trading: Which traders' trades move stock prices?, *Journal of Financial Economics* 61, 289-307.

Chan, Kalok, 1992, A further analysis of the lead-lag relation between the cash market and stock index futures market, *Review of Financial Studies* 5, 123-152.

Chan, Kalok, K.C. Chan, and G. Andrew Karolyi, 1991, Intraday volatility in the stock index and stock index futures markets, *Review of Financial Studies* 4, 634-684.

Chiyachantana, Chiraphol N., Pankaj K. Jain, Christine Jiang, and Robert. A. Wood, 2004, International evidence on institutional trading behavior and price impact, *Journal of Finance* 59, 869-898.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2011, Recent trends in trading activity and market quality, *Journal of Financial Economics* 101, 243-263.

Chordia, Tarun, Asani Sarkar, and Avanidhar Subrahmanyam, 2005, *An empirical analysis of stock and bond market liquidity*, Review of Financial Studies 18, 85-129.

Easley, David, and Maureen O'Hara, 1992, Time and the process of security price adjustment, *Journal of Finance* 47, 576-605.

Gonzalo, Jesus, and Clive Granger, 1995, Estimation of common long-memory components in cointegrated systems, *Journal of Business and Economic Statistics* 13, 27-35.

Gorton, Gary B., and George G. Pennacchi, 1993, Security baskets and index-linked securities, *Journal of Business* 66, 1-27.

Griffin, John M., Jeffrey H. Harris, Tao Shu, and Selim Topaloglu, 2011, Who drove and burst the tech bubble?, *Journal of Finance* 66, 1251-1290.

Griffin, John M., Jeffrey H. Harris, and Selim Topaloglu, 2003, The dynamics of institutional and individual trading, *Journal of Finance* 58, 2285-2320.

Hasbrouck, Joel, 1991a, Measuring the information content of stock trades, *Journal of Finance* 46, 179-207.

Hasbrouck, Joel, 1991b, The summary informativeness of stock trades: An econometric analysis, *Review of Financial Studies* 4, 571-595.

Hasbrouck, Joel, 1995, One security, many markets: determining the contributions to price discovery, *Journal of Finance* 50, 1175-1199.

Hasbrouck, Joel, 2002, Stalking the "efficient price" in market microstructure specifications: an overview, *Journal of Financial Markets* 5, 329-339.

Hasbrouck, Joel, 2003, Intraday price formation in U.S. equity index markets, *Journal of Finance* 58, 2375-2400.

Hasbrouck, Joel, 2007, *Empirical market microstructure: the institutions, economics, and econometrics of securities trading* (Oxford University Press: New York).

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1-33.

Kawaller, Ira G., Paul D. Koch, and Timothy W. Koch, 1987, The temporal price relationship between S&P 500 futures and the S&P 500 index, *Journal of Finance* 42, 1309-1329.

Keswani, Aneel, and David Stolin, 2008, Which money is smart? Mutual fund buys and sells of individual and institutional investors, *Journal of Finance* 63, 85-118.

Kumar, Alok, 2009, Who gambles in the stock market?, *Journal of Finance* 64, 1889-1933.

Kumar, Alok, and Charles M.C. Lee, 2006, Retail investor sentiment and return comovements, *Journal of Finance* 61, 2451-2486.

Kyle, Albert, 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315-1335.

Mackinlay, A. Craig, and Krishna Ramaswamy, 1988, Index-futures arbitrage and the behavior of stock index futures prices, *Review of Financial Studies* 1, 137-158.

Nofsinger, John R., and Richard W. Sias, 1999, Herding and feedback trading by institutional and individual investors, *Journal of Finance* 54, 2263-2295.

Puckett, Andy, and Xuemin Yan, 2011, The interim trading skills of institutional investors, *Journal of Finance* 66, 601-633.

Sarkar, Asani, and Robert A. Schwartz, 2009, Market sideness: insights into motives for trade initiation, *Journal of Finance* 64, 375-423.

Subrahmanyam, Avanidhar, 1991, A theory of Trading in stock index futures, *Review of Financial Studies* 4, 17-51.

Vega, Clara, 2006, Stock price reaction to public and private information, *Journal of Financial Economics* 82, 103-133.

Vijh, Anand M., 1994, S&P 500 trading strategies and stock betas, *Review of Financial Studies* 7, 215-251.

VITA

Yanhao Fang was born in Shanghai, China. He studied for a Bachelor of Arts degree majoring in English literature at Fudan University. After his graduation in 2002, he went on to obtain his second bachelor degree in economics at McGill University in Montreal, Quebec, Canada in 2005. He spent one more year in the Master program of economics at McGill University and graduated in 2006. He was expecting to become a Doctor of Philosophy in Business Administration with a concentration in finance at Louisiana State University in August 2012. Working as a teaching assistant at LSU, he taught corporate finance for business major and non-business majors and investments at the undergraduate level. His research interests include empirical market microstructure, asset pricing and corporate finance.