Universitat Jaume I

Department of Economics

The Role of Internet Information in Stock Market

Doctoral Thesis

Dehua Shen

Supervisors: Andrea Teglio and Silvano Cincotti

Castellón de la Plana, April 2016

Universitat Jaume I

Department of Economics

The Role of Internet Information in Stock Market

Doctoral Thesis

Dehua Shen

Supervisors: Andrea Teglio and Silvano Cincotti

Castellón de la Plana, April 2016

Acknowledgment

Several years ago, I could never have imagined that I would have the opportunity to pursuit a doctoral degree in Spain. However, this remarkable journey turns into a reality with tremendous help from Andrea Teglio and Silvano Cincotti. They not only give me an opportunity to realize the "unrealistic" dream, but also broaden my horizons on this diversified and multi-cultural world. More important, I would like to thank Andrea for the illuminating suggestions and personal support during these days.

It is a great pleasure to be part of the department of economics for the past two years. I am very thankful to the entire department. Particularly, I would also like to thank Simone Alfarano, Thomas Lux, Miguel Ginés Vilar and Eva Camacho Cuena for their inspiring wisdom as well as the willingness to help me.

I would also like to thank the reviewers of this dissertation, Shu-Heng Chen and Tony He. Your precious comments that have helped me improve the paper substantially.

I must express my very profound gratitude to my father, Mingzheng Shen, and my mother, Lingying Wang, who tried all their best to offer me excellent education with unconditional love and support. This dissertation is as much theirs as it is mine.

Last but not least, I wish to give my heartfelt thanks to my girlfriend, Xiao Li, for always being there and sharing this journey with me.

<div align="right">Castellón de la Plana, April 15, 2016</div>

Table of Contents

List of Tables

List of Figures

The Role of Internet Information in Stock Market

Abstract

This dissertation introduces the concept of Internet Information and highlights its crucial role in stock market. At the beginning, I discuss the intrinsic relations between market efficiency and Internet Information and argue that Internet Information serves as a way of complementing the "data" for understanding the dynamics of stock market. In chapter 2, I give an overview of stock market predictions using Internet Information and put forward some research directions. Naturally, four short essays are presented to illustrate the utilization of Internet Information. In particular, Chapter 3 employs the news appeared in the Internet as the proxy for the rate of information flow and examines the nature of the information-volatility relations. Chapter 4 demonstrates that the familiarity-based investment is driven by information with a unique online information interactive channel. Chapter 5 shows that the investor sentiment constructed by investor's searching behavior can explain the excess comovement of stock returns. Chapter 6 illustrates the impact of happiness sentiment extracted from social media on stock market performance.

In general, this dissertation addresses a timely important issue presented to us in this revolutionary information environment and takes an important step in highlighting the role of Internet Information in stock market.

1. Introduction: Market Efficiency and Internet Information

The reason for this is that the "data" from which the economic calculus starts are never for the whole society "given" to a single mind which could work out the implications, and can never be so given... If we can agree that the economic problem of society is mainly one of rapid adaptation to changes in the particular circumstances of time and place, it would seem to follow that the ultimate decisions must be left to the people who are familiar with these circumstances, who know directly of the relevant changes and of the resources immediately available to meet them...We must look at the price system as such a mechanism for communicating information if we want to understand its real function.

---Hayek, F. A., The Use of Knowledge in Society, *American Economic Review*, 1945, 35(4): 519-530.

1.1. Market Efficiency

In response to the question of what is the problem when we try to construct a rational economic order, Hayek (1945) argues that it's the problem of how to make the best use of the knowledge in society, which is based on the assumption that individuals have perfect information and computational capacities. Therefore, the challenge to the best use of the knowledge is that the knowledge itself never exists in integrated form, but incompletely separated in individuals. Thus, the "data" for the economic calculus is only a part of the data given by the society. Meanwhile, Simon (1955) further argues that the "economic man" postulated by traditional economic theory is still problematic and should be taken into account of their limited computational and predictive ability, i.e., bounded rationality. These are the origins to explain why the capital market is inefficient in the sense of allocative efficiency and informational efficiency before the Efficiency Market Hypothesis (Fama, 1970) became prevailing.

Almost simultaneously after Jensen's (1978) claim that "there is no other proposition in economics which has more solid empirical evidence supporting it than the Efficient Market Hypothesis", the prospect theory (Kahneman and Tversky, 1979), the excessive volatility (Shiller, 1981), the overreaction (De Bondt and Thaler, 1985), the momentum (Jegadeesh and Titman, 1993), the sentiment (De Long *et al.*, 1990) and some psychological biases (Lakonishok *et al.*, 1994; Barberis *et al.*, 1998; Daniel *et al.*, 1998) are proposed to challenge the Efficient Market Hypothesis.

The crucial issue in stock market is the way in which information is generated, diffused and utilized. A large amount of studies investigate the role of mass media (e.g., newspaper, advertising and television) in stock market. Cutler *et al.* (1989) investigate the relations between the news from *New York Times* and the extreme market-wide returns and conclude that news can only explain a small proportion of the return variations. Mitchell and Mulherin (1994) find the *Dow Jones* announcements are closely related to market activity. Klibanoff *et al.* (1998) find the news appeared on the front page of *New York Times* can affect the reaction of closed-end country funds. Liang (1999) shows that the analysts' recommendations in the "Dartboard" column of the *Wall Street Journal* have short-term impact on the stock returns. Huberman and Regev (2001) show that the stale news on "a potential development of new cancer-curing drugs" published by *New York Times* can lead the price to an extreme high level. Chan (2003) employs the *Dow Jones* Interactive Publications Library of past newspapers, periodicals and newswires and finds strong drift after bad news and reversal after extreme price fluctuation unaccompanied by public news. Meschke (2004) investigates the impact of CEO's interview on CNBC on the stock prices and trading volume and concludes the interview can generate transitory buying pressure. Grullon *et al.* (2004) find that firm's advertising expenditures is related to the breadth of ownership measured by the number of investors and the liquidity. Fang and Peress (2009) investigate the relations between the news on *New York Times*, *USA Today*, *Wall Street Journal and Washington Post* and expected stock returns and conclude that the breadth of information diffusion affects the stock prices. Tetlock (2007) find that the unusually high or low sentiment extracted from *Wall Street Journal* column can predict high trading volume. Griffin *et al.* (2011) provide global evidence that news published on *Wall Street Journal*,

Financial Times, *Dow Jones*, and *Reuters* can move the stock prices. Engelberg and Parsons (2011) confine their study to the pre-Internet period of 1991 to 1996 and find that the local newspaper coverage can increase the trading volume of local investors. Solomon *et al.* (2014) find the news on the *Wall Street Journal*, *New York Times*, *Washington Post*, and *USA Today* can affect investor's allocation behavior in mutual fund. Other similar studies include Liu *et al.* (1990), Pound and Zeckhauser (1990), Barber and Loeffler (1993), Tetlock *et al.* (2008), Tetlock (2010) and Tetlock (2011), Gurun and Butler (2012), Solomon (2012), Ahern and Sosyura (2014) and Dang *et al.* (2015).

1.2. Internet Information

Over the past twenty years, the Internet has emerged itself as the mainstream platform for generating, processing and diffusing information from public to investors. In stock market, the Internet has changed the landscape for financial advice, brokerage and decision-making of investors (Barber and Odean, 2001). Each day the Internet creates millions of firm-specific and market-wide information in stock markets worldwide. In a broad concept of "big data", Einav and Levin (2014) claim that "There was five exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days, and the pace is increasing". The efficiency of the stock market is in large attribute to the amount of information and its diffusion process. In that sense, the Internet is undoubtedly playing a crucial role in stock market.

In this dissertation, I define the concept of "Internet Information" as both the firm-specific and the market-wide news, rumors and information that appeared in the Internet (e.g., stock message board, search engine, microblogging, spam emails) as well as the newly emerged online information interaction channels (e.g., the online investor education and online investor relationship management), which can potential cause the formations and updates of investors' expectation, influence their decision-making and eventually have a material impact on the stock prices. This definition is mainly from the perspective of complementing the "data" for

understanding the dynamics of stock market. As shown in Figure 1.1, Internet Information is an additional information source that can be employed for analytics and computing. Other sources of information include the information on tradition communication media (e.g., newspaper, advertising and television) and stock market data (e.g., stock prices, trading volume and short interest). Only when we reach fully integrated the Internet Information, traditional information and stock market data into the analytics and computing, can we have a panoramic understanding of the market. It is crucial to recognize that this analytics and computing is an interdisciplinary field which draws on financial economics, behavioral economics, artificial intelligence and computer science. Figure 1.1 also illustrates that the definition of Internet Information is distinct from the concept of "big data". Since the (ultra)high frequent stock market data (e.g., tick data, trades & quotes and millisecond timestamp data) is always viewed as an important part of big data in finance.
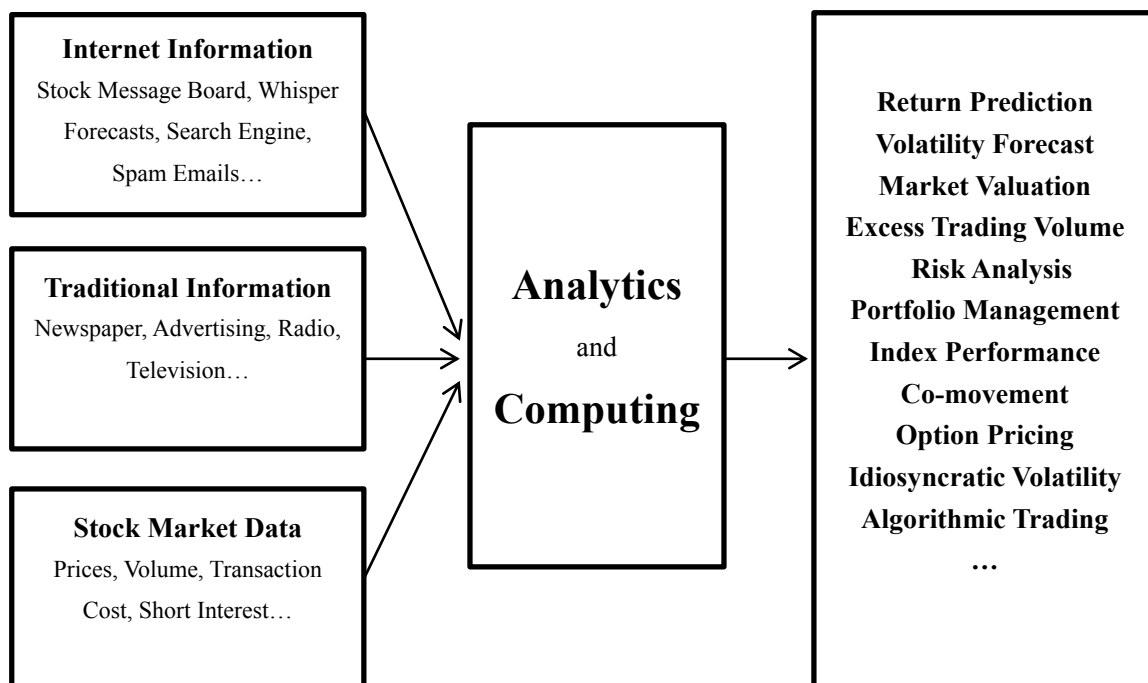
Figure 1.1 A simple illustration of the role of Internet Information in stock market

This figure illustrates the role of Internet Information in stock market. This definition of Internet Information is from the perspective of complementing the "data" for understanding the dynamics of stock market with analytics and computing.

2. An Overview of Stock Market Predictions Using Internet Information

An overview of the utilization of Internet Information for stock market predictions is given in this chapter. It is widely recognized that information plays a crucial role in stock market. The geometric volumes of Internet Information reflecting various aspects of investors' activities represent a new perspective for financial economist to address the predictability of stock market and a vital opportunity to address fundamental questions about the underlying causes and pricing mechanism. In this chapter, I will give an overview of stock market predictions using Internet Information, starting from the earliest discussion existed in mainstream newspapers on the predictive power of the Internet Information, to the interdisciplinary efforts from financial economists, econometricians, computer scientists and econophysics researchers.

2.1 Early Discussion: Information or Noise

The earliest discussion of using Internet Information for the stock market predictions could trace back to reporters and commenters from mainstream newspaper in the USA, e.g., *Seattle Times* (Batsell, 1998), *Dow Jones News Service* (Bennett, 1998), *Dallas Morning News* (Goldstein, 1998), *New York Times* (Harmon, 1998), *Wall Street Journal* (Maremount, 1998) and *Chicago Daily Herald* (Medill, 1998). All these articles raise a very intriguing topic: does the Internet Information have some predictive power for the stock market? Even if the Computer Business Review reported that "comments posted to a Yahoo! Finance message board have so devalued a Las Vegas, Nevada-based biotech stock that the company's chief executive officer held a conference call to reassure investors and analysts that fundamentals are sound", little is known about the relations between various sources of Internet Information and stock market performance, e.g., abnormal return, trading volume, breadth of ownership, intraday volatility and liquidity.

What is certain is that Internet Information has attracted the interest of financial economist. To the best of our knowledge, Wysocki (1998) is the first to document the relationship between message-posting volume of more than 3000 stock listed on Yahoo! Finance Message Board and the stock market performance. Wysocki finds that the overnight message-posting volume can be used to predict changes in next day's stock return and trading volume, concluding that the message-posting volume is related to the fundamentals of underlying companies. In contrast to this anecdotal evidence claiming the Internet Information indeed contain information relevant to fundamentals, Tumarkin and Whitelaw (2001) focus on the RagingBull.com discussion forum, examine the relationship between message board postings and the abnormal stock returns and find that the message board posting has no predicative power for abnormal trading volume and industry-adjusted returns. However, the interest in excavating the association between Internet Information and stock market performance has gained increasing popularity and various kinds of Internet Information are utilized in empirical studies. Bagnoli *et al.* (1999) collect the whisper forecast of quarterly earnings from the website of Wall Street Journal, fool.com and techstocks.com and they find that the whisper forecast spread on the Internet contains information that not conveyed in the official forecasts (e.g., the First Call analyst forecasts). Felton and Kim (2002) search postings on Yahoo! Enron Message Board for four years before Enron collapsed in 2001 and they find that there exist some descriptions of corporate culture and warnings for investors to dump their holdings by apparent insiders. Dewally (2003) investigates the market reaction to the online stock recommendations (archived in Deja-News website) and concludes that the there are no new information exchanged in the forums. Clarkson *et al.* (2006) examine the stock market reaction to takeover rumors in the HotCopper (an Australian stock market and politics forum) and indicate a statistically significant relationship between market performance and the posting of takeover rumors. Böhme and Holz (2006) correlate the unsolicited email (spam emails) from the Stock Spam Effectiveness Monitor (SSEM) archive with the stock market and find a significantly positive abnormal return on the spam campaigns event date. Hanke and

Hauser (2008) retrieve the spam emails from the Crummy database and find that this unsolicited emails have a significant impact on the excess stock return, turnover and intraday volatility. Hu *et al. (*2011) acquire data from BlogPulse.com and document a positive impact on the market valuation as well as one-way directional Granger causality from blogging to trading activity. Blankespoor *et al.* (2014) find that firms with additional information diffusion channel, i.e., use Twitter to send links to market participants, have greater abnormal depths and lower abnormal bid-ask spreads. Generally speaking, all these studies have shifted research focus from building more sophisticated prediction models to the sources of information themselves and recognized the role of Internet Information in stock market.

2.2 Investor Attention, Information Demand and Information Arrival

Apart from discovering the interesting empirical results with the utilization of Internet Information, financial economists begin to construct direct proxies to revisit some financial theories as well as validate some hypotheses. The majorities of these studies mainly rely on the unique data retrieved from search engine, e.g., Google and Baidu, and relate to investor's searching behavior or state of mind.

Da *et al.* (2011) construct a direct proxy for investor attention with the search frequency of stock names of Russell 3000 in Google Trends from 2004 to 2008. They find that an increase in investor attention can predict higher stock returns in the following 2 weeks and investor attention can also explain the large debut-day return and long-term underperformance. Bank *et al.* (2011) also use the Google search volume of company name as the proxy for investor attention and study the implications for German stock market. They find that an increase in search volume is associated with short-term higher stock returns and rise in liquidity and trading activity and attribute the improved market liquidity to the reduction of information asymmetry. Zhang *et al.* (2013) employ the search frequency of stock name in Baidu Index as the proxy for investor attention in China and conclude that investor attention increase the predictive power for abnormal return. Aouadi *et al.* (2013)

investigate this issue in French stock market and find that the search-based investor attention is a significant determinant of the market volatility and illiquidity. Vozlyublennaia (2014) investigates the relations between index performance and investor attention measured by Google search volume and find that there exists a significant short-term impact of increased attention on index return, but a long-term impact of return shocks on the change in attention. They further demonstrate that the increased investor attention reduces index return predictability and simultaneously improves the market efficiency. Ding and Hou (2015) employ the search frequency of stock ticker provided by Google Trends as the proxy for retail attention and find it can improve stock liquidity and enlarge the breadth of ownership. Dimpfl and Jank (2016) find that the Google search-based investor attention can Granger-cause volatility and incorporating the searches into the prediction model for realized volatility leads to more precise in-sample and out-of-sample forecast.

In a similar way, Drake *et al.* (2012) view the Google search volume as the information demand for public information and find that this information demand increases about 2 weeks before earnings announcement, spikes sharply on the announcement date and retains at high level for a period after the announcement. Therefore, they conclude that information diffusion is not instantaneous. Vlastakis and Markellos (2012) also view the Google search volume as the information demand and investigate the relations between information demand and the risk aversion and stock market activity. Gwilym *et al.* (2015) view the search frequency of "concept stocks" obtained from Google Insights as investors' speculative demand and this demand can predict subsequent price reversals.

Some of the other studies further elaborate the searching behavior of investor as the reflection of investor's state of mind. Joseph *et al.* (2011) elaborate the search frequency from Google Insights as "a set of beliefs about cash flow and investment risks", i.e., investor sentiment, and find that this sentiment can predict trading volume and abnormal stock returns. Following the intuition that investors react to

uncertainty by intensifying their search behavior, Dzielinski (2012) expounds the search frequency from Google Trends as the economic uncertainty and find this indicator is significantly correlated with stock return and volatility. Irresberger *et al.* (2015) construct a market-level and individual crisis sentiment with the search volume from Google Trends and find this sentiment can lead investor to devalue the bank stocks. Da *et al.* (2015) also construct the investor sentiment by Google Trends and find this FEARS (Financial and Economic Attitudes Revealed by Search) index can predict temporary increases in volatility, mutual fund flow and short-term price reversals. Besides, Zhang *et al.* (2014) employ the number of news appeared in Baidu News as direct proxy for information arrival, incorporate this proxy into the conditional variance equation of the GARCH model and find that the Internet Information can explain volatility clustering. This study is further extended by Shen *et al.* (2016) on separating the news into trading period and non-trading period. They show that both the lead information and the aggregate (the sum of lead and contemporaneous) information can explain the volatility clustering of individual stock and naturally provide alternative evidence for the Mixture of Distribution Hypothesis. In spite of a great quantity of work has been done in this stream, we should continue to explore the unfathomable information provided by search engine. The reason is that these microscopic and comprehensive data generated by hundreds of millions of Internet users can help us to propose new types of research questions and enable original research design that can reshape our understanding of the complex financial systems.

2.3　Machine Learning, Text Mining and Linguistic Classification

Based on the existing studies admitting that the Internet Information is not pure noise but related to fundamentals, interdisciplinary efforts from financial economists, econometricians, computer scientists and econophysics researchers bring some new analytics tools, e.g., machine learning, text mining and natural language processing, to extract sentiment form the content of Internet Information. The rationale behind

these efforts is that the content of Internet Information will have a materially effect on investors' decision-making process and subsequently affect the stock market performance. Naturally, the tonality of the content (positive or negative) is intrinsically correlated to the movement (upward or downward) in the stock prices.

Antweiler and Frank (2004) use the computational linguistics method to extract the bullishness index from more than 1.5 million postings on Raging Bull and Yahoo! Finance and they find the index can predict volatility both at intraday and daily frequency. Das *et al.* (2005) construct the sentiment and disagreement measures with the natural language processing technique from the Motley Fool stock message board and find this sentiment is related to the lagged stock returns. Das and Chen (2007) develop the voting scheme, containing the naïve classifier, vector distance classifier, discriminant-based classifier, adjective-adverb phrase classifier and Bayesian classifier, to extract sentiment from the Yahoo! Finance and show the constructed sentiment index is closely related to the Morgan Stanley High-Tech Index. Bollen *et al.* (2011) extract the collective mood states, including alert, sure, vital, kind, calm and happy, from Twitter and find this mood can predict the changes in closing value of the Dow Jones Industrial Average (DJIA). Zhang *et al.* (2011) construct a "collective hope and fear" index with the postings in Twitter and find a negative correlation between this index and S&P 500, NASDAQ and DJIA. Sabherwal *et al.* (2011) also construct the "online trader's credit-weighted sentiment index" from the Lion.com and find this index can predict the subsequent two trading days' returns. Zhang *et al.* (2012) download message board postings from Thelio!WallStreetPit and show the sentiment generated from text classifiers is an negative indicator of next-day's stock return. Chen *et al.* (2014) find that the retail investor's opinion extracted from Seeking Alpha can strongly predict future earnings surprise and stock returns. Siganos *et al.* (2014) examine the relations between Facebook's Gross National Happiness Index and trading behavior over 20 international stock markets and they find negative sentiment is related to increased volatility and trading volume. With the method of computational linguistics,

Sprenger *et al.* (2014) also extract the sentiment from Twitter with about 250,000 relevant tweets and find this tweet sentiment is closely related to trading volume and stock returns. They also conclude that the Twitter can amplify the diffusion of information. Studies that examine the impact of online information channel include Jones (2006) and Jin *et al.* (2016) among others.

Obviously, the research in this field has aroused the common interest from financial economics and computer scientists. Several issues are crucial for the quality of the empirical results, e.g., the natural language processing technique for non-English speaking countries, the accuracy of the classification methods and the reduction of the noise information from unstructured data.

2.4   Concluding Remarks and Future Research Directions

It can most reasonably be inferred from the above-mentioned empirical findings that the Internet Information will consistently play a more crucial role in crystalizing our understanding of the stock market. There are already signs in literature focusing on constructing proxies to investigate the financial theories. The Internet Information can be served as a unique source of data for empiricists to examine the untested theories. Examples include the relations between soft information (e.g., voice, gesture and facial expression) from managers and investors' perception, the dynamics asset pricing model based on the propagation path of information as well as the impact of biased information on expectation formation.

Another promising avenue for future research is to combine multi-sources of information. Most existing studies only adopt a single source of information. The information sharing mechanism is the way of coping with the low value density of Internet Information. The successful implementation of the sharing mechanism requires the basic research from computer scientists, mathematicians and econometricians. Meanwhile, the privacy problem should be highlighted.

It would be worth investigating the function of Internet Information as early warning signs of the contagion or interdependence. If the distributions and dynamics of Internet Information are correlated to the state of mind or the range of knowledge of hundreds of millions online users, it will make it possible to predict the macrobehavior emerged from investors' micromotives. To some extent, this point is somehow ignored as scholars are indulged into building more sophisticated models or just notice it without going into it deeply.

Last but not least, we firmly believe that the most challenging and promising avenue is to combine the Internet Information with the data generated by Agent-based model (ABM). These two sources of data will definitely go hand in hand, supplement and complement each other and help each other to understand the dynamics of the economic system as well as realize the data-driven decision-making.

2.5   Research Questions and Methodologies

Naturally, this dissertation is deeply motivated by employing Internet Information for academic research from the perspective of complementing the "data" for understanding the dynamics of stock market. In particular, four short essays are presented to illustrate the utilization of Internet Information. Figure 2.1 illustrates the research questions and methodologies of this dissertation. Starting from the hindsight on exploring more useful information, I first introduce a novel proxy for information flow and validate the SIAH in chapter 3. And then in chapter 4, I demonstrate that the familiarity-based investment is driven by information with a unique online information interactive channel. Chapter 5 shows that the investor sentiment constructed by investors' searching behavior can explain the excess comovement of stock returns. Chapter 6 illustrates the dependences between state of mind revealed by Twitter and the corresponding stock market performance. The methodologies mainly consist of data mining and regression analysis.

Internet Information

newly emerged, low cost, reveal behavior,
high frequent, diffusion path……

An Overview of Stock Market Predictions Using Internet Information

➔ Early Discussion: Information or Noise
➔ Investor Attention, Information Demand and Information Arrival
➔ Machine Learning, Text Mining and Linguistic Classification

Hindsight: The rationales behind all these studies are united together by the challenge of the efficient market hypothesis, which also shifted the academic attention from employing more sophisticated perdition models to exploring more useful Internet Information

➢ **Internet Information, Return Volatility and Trading Volume**
✓ **Methodologies:** data mining from Baidu News and regression analysis
● **Main Contribution:** introduce a novel proxy for information flow and validate the SIAH

➢ **The Impact of Information-based Familiarity on the Stock Market**
✓ **Methodologies:** manual collect from the website and regression analysis
● **Main Contribution:** prove the familiarity-based investment is driven by information

➢ **Online Sentiment Dynamics and Excess Comovement of Stock Returns**
✓ **Methodologies:** data mining from Baidu Index, PCA and regression analysis
● **Main Contribution:** explain the excess comovement with online sentiment

➢ **Euphoria: Daily Happiness and Stock Returns**
✓ **Methodologies:** data mining from Twitter and regression analysis
● **Main Contribution:** discover the dependences between state of mind and stock returns

Figure 2.1 Research questions and methodologies of this dissertation

3.   Internet Information, Return Volatility and Trading Volume

The information-volatility relations have long been the focus of financial empiricists. They mainly consider the trading volume as the proxy for the rate of information flow and examine these relations with the framework of Mixture of Distribution Hypothesis and Sequential Information Arrival Hypothesis. In this chapter, we advocate the search frequency of stock names in Baidu News as a direct proxy for the rate of information flow. This novel proxy not only overcomes the endogenous problem of the trading volume, but also portrays the information flow during the trading periods. Therefore, we examine the information-volatility relations in Chinese stock market and discover that the information is not instantly incorporate into the prices but requires some time to reach the ultimate equilibrium. Besides, we also observe the differences of order imbalance and large size of the trade in news and no-news periods, respectively.

3.1   Introduction

Return volatility can be the result of the rate of information flow, the reflection of the private information as well as the irrational behavior of the noise trader. Among these, the rate of information flow has been widely considered as the major factor uncovering the underlying mechanism on the changes of asset prices (Grossman and Stiglitz, 1980; Kyle, 1985 and Glosten and Milgrom, 1985; French and Roll, 1986 Ross, 1989; Bergemann *et al.*, 2015). However, the interactions and dynamics between the rate of information flow and return volatility still remain debatable for the reason that the intangible form of the information makes it hard to construct desired proxies. Existing literature mainly relies on trading volume (Lamoureux and Lastrapes, 1990 and Le and Zurbruegg, 2010) and its adjusted forms (Wagner and Marsh, 2005 and Fleming *et al.*, 2006 and Park, 2010), the number of firm-specific announcements (Kalev *et al.*, 2004) and number of transactions (McMillan and García, 2013) as proxies for the information flow. They usually incorporate the

constructed proxy into the conditional variance equation of the GARCH model (Bollerslev, 1986) and observe a significant reduction of the volatility persistence. In that sense, they provide evidences for the Mixture of Distribution Hypothesis (MDH), which claims that the changes of asset prices at a certain interval is associated with the rate of information flow (Clark, 1973; Epps and Epps, 1976; Harris, 1987 and Andersen, 1996). An alternative appealing explanation for the information-volatility relations is the Sequential Information Arrival Hypothesis (SIAH) (Copeland, 1976; Jennings *et al.*, 1981; Smirlock and Starks, 1988). The SIAH claims that there exist the lead-lag relations between the rate of information flow and return volatility.

The main purpose of this chapter is to examine the underlying mechanism, i.e., MDH or SIAH, between the rate of information flow and return volatility for Chinese stock market with a more appropriate proxy for the rate of information flow. This proxy is constructed by acquiring information from the Internet. In particular, we use the search frequency of the stock names from Baidu News as the direct proxy for the rate of information flow (a detailed introduction and illustration is given in Section 3.2.2) and uncover evidences on upholding the SIAH. With very few exceptions, Zhang *et al.* (2014) and Shen *et al.* (2016) also employ such proxy. But their studies focus on the explanatory power of the Internet Information, while we intent to address the nature, i.e., MDH or SIAH of the information-volatility in this chapter.

We contribute to the existing literature on the information-volatility relations in several aspects. Firstly, we use the information acquired from the Internet, it can naturally avoid the endogenous problem introduced by other proxies, e.g., trading volume and its adjusted forms. Besides, we can further divide the information into trading period information and non-trading period information and only the trading period information is considered as the information flow. This Internet Information is more satisfying proxy for the new information as assumed by MDH and SIAH

compared with the earning announcements. Secondly, we calculate four different measurements of volatility, i.e., the 1-minute, the 5-minutes, the GARCH and the range-based volatilities to examine the contemporaneous and causal relations between the information flow and return volatility. The volatilities constructed by intraday data (1-minute, the 5-minutes, the GARCH volatilities) are particularly suitable for examining the information-volatility relation. Because the stock market shows high speed of reaction to information, it will capture the valuable information contained in the intraday price movement. Thirdly, investigated by Lee and Rui (2000), the empirical study focusing on contemporaneous and causal relationships in Chinese stock market is more than one decade ago, both the investor structure and the trading rules have changed a lot. For example, the Qualified Foreign Institutional Investors (QFII) are allowed to trade stock in July 2003 and the all the stocks in B shares are changes to T+1 trading mechanism in December 2001. Reexamining the relations between the rate of information flow and return volatility will definitely provide some insights into the understanding of the pricing mechanism. Therefore, it also provide an alternative evidence on emerging stock market, other than a large amount of studies focusing on the developed market, e.g., the NYSE (Darrat *et al.*, 2003 and 2007), DAX, FTSE, CAC and TPX (Wagner and Marsh, 2005).

3.2   Data Description

There are mainly two sources of data in this chapter. The first refers to the capital data retrieved from RESSET Financial Research High Frequency Database. A brief introduction of Chinese stock market is given in the following section. The second is the Internet Information acquired from Baidu News representing the rate of information flow for individual stock.

3.2.1   Capital Data

The Chinese stock market composes of Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE), which were established on 19 December 1990 and 3 July 1991, respectively. Except for some national holidays (e.g., National Days,

Mid-autumn Festival, Dragon Boat Festival and Lunar Chinese New Year), both the exchanges are open five days a week (from Monday to Friday) from 9:30 in the morning to 15:00 in the afternoon with a trading break from 11:30 to 13:00 in GMT+8 time zone. Totally, there are 4 hours for trading on each trading day. There are three categories of shares, i.e., A shares, B shares and H shares. A shares are those of the local Chinese companies traded by RMB and primarily on Shanghai and Shenzhen stock exchanges. Only the Qualified Foreign Institutional Investors (QFII) with special permission are allowed to participate in this market. B shares refer to the companies listed in Shanghai, Shenzhen and Hong Kong stock exchanges and traded in U.S. Dollars with face value in RMB. Mainland Chinese investors are also allowed to trade B shares with legal foreign currency accounts. H shares usually represent the Chinese companies listed in Hong Kong stock exchange, and therefore can be traded freely by anyone. There are some companies that are dual-listed in Kong Hong stock exchange and either in Shanghai stock exchange or Shenzhen stock exchange. A commonly held belief is that the A shares dominate the market with both number of investor and capitalization.

There are some evidences on suggesting that Chinese stock market is significant different from other developed stock markets, e.g., the New York stock exchange and the London stock exchange. Firstly, according to a report released by Shanghai stock exchanges, individual investors account for a very high percentage of all the accounts in A shares: about 92.3% in 2012. Secondly, there exist price limits for the individual stocks where the prices cannot change more than 10% (-10%) from the opening prices at each trading day. Thirdly, for majorities of the stocks, short sale is still constrained. On 31 March 2010, the China Security Regulatory Commission (CSRC) approved the margin trading and securities lending program to remove the restrictions on short selling for selected stocks. Up to the end of 2015, only 892 (less than one third) stocks are allowed to be sold short, but the naked short sale is still prohibited for all the stocks on Shanghai and Shenzhen stock exchanges.

Our sample consists of the intraday data and daily data from 1 August 2011 to 31 August 2011 for 40 stocks of Shanghai Shenzhen CSI 300 Index (CSI 300 Index). The intraday data includes the tick data, 1-mintue, 5-minutes stock prices and trading volume. The daily data includes the GARCH-volatility, opening, closing, highest and lowest stock prices and trading volume. All the data are comes from the RESSET Financial Research High Frequency Database. The CSI 300 Index is the first index launched by both Shanghai and Shenzhen stock exchange with the aim of representing the overall performance of Chinese stock market and it is revised every six months with adding and deleting stocks from the index. The selected 40 stocks are the long-lived stocks that exist from the beginning date of the CSI 300 Index to end of December 2015. Therefore, these actively traded stocks are likely to have sufficient information flow per day to satisfy the conditions for the Central Limit Theorem employed by MDH (Osborne, 1959 and Lamoureux and Lastrapes, 1990).

Furthermore, we confined the sample period to August 2011 for the following three reasons. Firstly, it is due to the initial information we acquired from the Internet. According to the trading periods of Chinese stock market, the intraday Internet Information is divided into four subgroups, i.e., T1 (from 9:30 a.m.to 11:30 a.m.), T2 (from 11:30 a.m. to 13: 00 p.m.), T3 (from 13:00 p.m. to 15:00 p.m.) and T4 (from 15:00 p.m. to 9:30 a.m. in next trading day). T1 and T3 represent the trading period and T2 and T4 refer to the non-trading period. To control for the time interval of T1, we only consider the week that has five consecutive trading days, i.e., from Monday to Friday. This means that if one trading day, e.g., Wednesday, is missing in a certain week, we delete this week from the sample. The reason is that if Wednesday is missing, the information flow in T1 on Thursday is not appropriately acquired. In August 2011, there are 23 consecutive trading days and it naturally meets this criteria. Secondly, there are neither earning announcements nor trading halts for the selected 40 stocks in August 2011, which rule out the possibility that our results could be biased by other confounding factors. Thirdly, the macroeconomic conditions are normal and there is no major economic news in August 2011. In that

sense, the potential impact of macroeconomic environment (Engle and Rangel, 2008 and Zhang *et al.*, 2013) on the information flow is low.

3.2.2   Baidu News

The Internet Information employed in this chapter is from the Baidu News, which is a service provided by Baudu. According to a survey on the searching behavior of Chinese netizen released by China Internet Network Information Center (CINIC, https://www.cnnic.net.cn/), Baidu dominates the Chinese searching market with more than 80% shares. In that sense, Baidu News is an ideal source of Internet Information. Baidu News retrieves the news from more than 500 authoritative websites and provides 24/7 update service to its users. We search the stock name with Biadu News and use the reported number of news as the information flow for individual stock. Figure 3.1 gives an illustration of Baidu News with the search of Baoshan Iron & Steel Co. Ltd. We search the Chinese name of the stock and Baidu News reports the number of relevant news containing the company name. In this illustration, Baidu News reports that "Baidu News finds 8 related news containing the keyword you searched" (the Chinese sentence in the blue ellipse). We use the number of related news as the information flow for individual stock. Besides, given the "Advanced Settings" provided by Baidu News, we can restrict the news to a certain interval. As a consequence, we can obtain the intraday Internet Information from Baidu News. For each stock at a given trading day, we confine the searching results to T1 and T3. Therefore, we can obtain the Internet Information during the trading periods. Since Baidu News does not support the directly download service, a Java crawler script is written to automatically download the data. In particular, for the selected 40 stocks in the 23 trading days, there are 179 trading days (19.46%) without Internet Information. For all the morning trading periods of all the stocks, the no-news periods account for 30.87% (284 trading periods out of 920 trading periods). While the percentage of no-news periods for the afternoon trading periods is 52.40% (482 trading periods out of 920 trading periods).

Figure 3.1 An illustration of Baidu News

The figure illustrates how the information acquired from Baidu News with the search of Baoshan Iron & Steel Co. Ltd. We search the Chinese name of the stock and Baidu News reports the number of relevant news containing the name. In this case, Baidu News reportss that "Baidu News finds 8 related news containing the keyword you searched" (the Chinese sentence in the blue ellipse). We use the number of related news as information flow for individual stock.

3.3   Internet Information and Return Volatility

In the beginning, it is necessary to explore the empirical criterion of distinguishing the MDH from SIAH. MDH assumes that the investors receive the information simultaneously and this information is immediately reflected into the price changes. In that sense, the market shifts to the ultimate equilibrium without any intermediately partial equilibrium exists in the process. Therefore, the return volatility is proportional to the rate of information flow at a given interval and the contemporaneous relation is derived. Alternatively, SIAH posits that investors' reactions to new information are largely different and thus there exists intermediately partial equilibrium before the ultimate equilibrium. Therefore, the lead-lag relations between the rate of information flow and return volatility are derived. In this section,

we examine the MDH and SIAH by investigating the contemporaneous and causal relations between the rate of information flow and return volatility.

To measure return volatility, we adopt the 1-minute (5-minutes) stock price to calculate the standard deviation of 240 (48) prices as the 1-min volatility (5-mins volatility). The GARCH volatility is directly downloaded from the RESSET Financial Research High Frequency Database and the calculation method is the same as Bollerslev (1986) with the GARCH (1, 1) specifications. Besides, we also calculate the range-based volatility (Garman and Klass, 1980) with the opening, closing, highest and lowest stock prices for each stock (see table 3.4 for the model specifications). Table 3.1, table 3.2, table 3.3 and table 3.4 report the contemporaneous correlations between 1-min, 5-mins, GARCH volatility and range-based volatility, respectively. On average, less than 4 stocks have significant contemporaneous correlations. Table 3.5, table 3.6, table 3.7 and table 3.8 report the results of Granger causality test between 1-min, 5-mins, GARCH volatility and range-based volatility, respectively. Table 3.9 reports the summary statistics for the MDH and SIAH. It shows that more than 78.125% of stocks (on average, 31.25 out of 40) have a significant lead-lag relations in either direction, i.e., the Granger Causality from the information flow to return volatility or the vice versa. To sum up, the significant lead-lag relations between the information flow and return volatility in large number of the stocks in accordance with SIAH. Figure 3.2 and Figure 3.3 illustrate the difference in average 1-min volatility between news periods and no-news periods in the morning trading period and afternoon trading period, respectively. They show that the volatility in the news periods is significant higher than the volatility in the no-news periods. This finding accords with Darrat *et al.* (2007) that the volatility of stocks in NYSE during the public news periods is larger than that during the no-news periods.

Table 3.1 Contemporaneous correlations between 1-min volatility and Internet Information

Stock codes	Correlation coefficients	p-value	Stock codes	Correlation coefficients	p-value
000002	0.4891**	0.0179	000800	-0.0352	0.8733
000009	-0.1435	0.5135	000807	-0.0260	0.9064
000012	0.3567	0.0948	000825	-0.1036	0.6382
000024	0.4407**	0.0353	000839	0.2256	0.3007
000061	-0.0154	0.9442	000858	-0.0675	0.7595
000063	0.3611	0.0905	000878	-0.1496	0.4957
000069	-0.3179	0.1394	000898	0.0562	0.7990
000157	-0.1991	0.3624	000933	-0.1072	0.6263
000401	-0.0206	0.9256	000937	0.1771	0.4187
000402	0.3200	0.1366	600000	-0.0022	0.9922
000527	0.1390	0.5270	600005	-0.2075	0.3420
000528	-0.2735	0.2066	600009	0.1341	0.5420
000568	-0.2813	0.1935	600010	0.3081	0.1526
000625	-0.0047	0.9830	600015	0.1366	0.5342
000630	0.1804	0.4102	600016	0.0237	0.9144
000651	-0.0196	0.9293	600019	0.2163	0.3216
000709	-0.1729	0.4302	600028	0.0183	0.9340
000729	-0.2036	0.3515	600029	0.0832	0.7058
000778	0.0768	0.7277	600031	-0.1478	0.5010
000792	-0.1588	0.4694	600036	0.2185	0.3165

Notes: The columns containing the stocks are: Vanke Co. Ltd (000002), Baoan Group Co. Ltd (000009), CSG Hldgs A (000012), China Merchants Property (000024), Shenzhen Agricultural Products Co. Ltd (000061), ZTE Corporation (000063), Shenzhen Overseas Chinese Town Co. Ltd (000069), Zoomlion Heavy Industry Sci & Tch Co. Ltd (000157), Tangshan Jidong Cement Co. Ltd (000401), Financial Street Holding Co. Ltd (000402), GD Midea Holding Co. Ltd (000527), Guangxi Liugong Machinery Co. Ltd (000528), Lu Zhou Lao Jiao Co. Ltd (000568), Cq Changan Auto A (000625), Tongling Nonferrous Metals Group Co. Ltd (000630), Gree Electric Appliances Inc of Zhuhai (000651), Hebei Iron And Steel Co. Ltd (000709), Beijing Yanjing Brewery Co. Ltd (000729), Xinxing Ductile Iron Pipes Co. Ltd (000778), Qinghai Salt Lake Potash Co. Ltd (000792), Faw Car Co. Ltd (000800), Yunnan Aluminium Co. Ltd (000807), Shanxi Taigang Stainless Steel Co. Ltd (000825), CITIC Guoan Information Industry Co. Ltd (000839), Wuliangye Yibin Co. Ltd (000858), Yunnan Copper Co. Ltd (000878), Angang Steel Company Ltd (000898), Henan Shenhuo Coal & Power Co. Ltd (000933), Jizhong Energy Resources Co. Ltd (000937), Shanghai Pudong Development Bank Co. Ltd (600000), Wuhan Iron and Steel Co. Ltd (600005), Shanghai International Airport Co. Ltd (600009), Mongolia BaoTou Steel Union Co. Ltd (600010), Hua Xia Bank Co. Limited (600015), China Minsheng Banking Corp Ltd (600016), Baoshan Iron & Steel Co. Ltd (600019), China Petroleum & Chemical Corp (600028), China Southern Airlines Co. Ltd (600029), Sany Heavy Industry Co. LTD (600031) and China Merchants Bank Co. Ltd (600036). ** indicates correlation coefficient significant at 5% level.

Table 3.2 Contemporaneous correlations between 5-mins volatility and Internet Information

This table reports the contemporaneous correlations between 5-mins volatility and Internet Information. The 5-mins volatility is calculated as standard deviation of the stock prices in 5 minutes' interval. The stock prices are obtained from the RESSET Financial Research High Frequency Database.

Stock codes	Correlation coefficients	p-value	Stock codes	Correlation coefficients	p-value
000002	0.5026**	0.0145	000800	-0.0311	0.8879
000009	-0.0913	0.6788	000807	-0.0457	0.8361
000012	0.3973	0.0605	000825	-0.1044	0.6354
000024	0.4880**	0.0182	000839	0.2365	0.2773
000061	-0.0594	0.7877	000858	-0.0602	0.7851
000063	0.4298**	0.0407	000878	-0.1390	0.5271
000069	-0.3232	0.1325	000898	0.0482	0.8271
000157	-0.1762	0.4213	000933	-0.0821	0.7098
000401	-0.0128	0.9536	000937	0.1971	0.3674
000402	0.2977	0.1677	600000	0.0074	0.9733
000527	0.1470	0.5034	600005	-0.1947	0.3734
000528	-0.2672	0.2178	600009	0.1377	0.5310
000568	-0.2770	0.2007	600010	0.3144	0.1440
000625	0.0147	0.9469	600015	0.1343	0.5414
000630	0.1832	0.4029	600016	0.0397	0.8572
000651	-0.0093	0.9664	600019	0.2267	0.2981
000709	-0.1466	0.5043	600028	0.0099	0.9644
000729	-0.2079	0.3412	600029	0.0826	0.7078
000778	0.0782	0.7228	600031	-0.1602	0.4654
000792	-0.1504	0.4933	600036	0.2499	0.2502

Notes: ** indicates correlation coefficient significant at 5% level.

Table 3.3 Contemporaneous correlations between GARCH volatility and Internet Information

This table reports the contemporaneous correlations between GARCH volatility and Internet Information. The GARCH-volatility is directly downloaded from the RESSET Financial Research High Frequency Database and the calculation method is the same as Bollerslev (1986) with the GARCH (1, 1) specifications.

Stock codes	Correlation coefficients	p-value	Stock codes	Correlation coefficients	p-value
000002	-0.2711	0.2109	000800	0.0165	0.9403
000009	-0.0772	0.7261	000807	0.1255	0.5683
000012	0.6207***	0.0016	000825	0.4209**	0.0455
000024	-0.1498	0.4952	000839	0.1619	0.4606
000061	-0.5122**	0.0125	000858	-0.0739	0.7376
000063	0.0858	0.6970	000878	-0.1756	0.4228
000069	0.2985	0.1665	000898	-0.0312	0.8874
000157	0.2123	0.3308	000933	0.3202	0.1364
000401	0.4718**	0.0230	000937	-0.2497	0.2506
000402	0.0528	0.8109	600000	0.3842	0.0703
000527	0.4447	0.0335	600005	-0.0912	0.6789
000528	-0.2103	0.3354	600009	-0.1713	0.4345
000568	-0.0191	0.9309	600010	0.5201	0.0110
000625	-0.0015	0.9947	600015	0.6535***	0.0007
000630	0.4547**	0.0293	600016	0.6876***	0.0003
000651	-0.1329	0.5454	600019	0.0077	0.9724
000709	-0.2268	0.2981	600028	0.0856	0.6976
000729	-0.3860	0.0689	600029	0.1675	0.4449
000778	0.0096	0.9655	600031	0.4052	0.0551
000792	-0.0905	0.6814	600036	0.1228	0.5767

Notes: *** indicates correlation coefficient significant at 1% level.

** indicates correlation coefficient significant at 5% level.

Table 3.4 Contemporaneous correlations between range-based volatility and Internet Information

This table reports the contemporaneous correlations between range-based volatility and Internet Information. The range-based volatility is calculated as the method proposed by Garman and Klass (1980) with the opening, closing, highest and lowest stock prices for each stock.

$$RV_{i,t} = \frac{1}{2} hl_{i,t}^2 - (2ln^2 - 1)oc_{i,t}^2$$

where $hl_{i,t}$ is the difference in natural logarithms of the highest and lowest prices for index i on day t, $oc_{i,t}$ is the difference in natural logarithms of the opening and closing prices for index i on day t.

Stock codes	Correlation coefficients	p-value	Stock codes	Correlation coefficients	p-value
000002	0.8195***	0.0000	000800	-0.0734	0.7392
000009	0.0508	0.8181	000807	-0.0247	0.9108
000012	0.4104	0.0518	000825	-0.1386	0.5283
000024	0.6955***	0.0002	000839	0.2216	0.3094
000061	0.0462	0.8340	000858	-0.1508	0.4923
000063	0.6161***	0.0017	000878	-0.1136	0.6059
000069	-0.3202	0.1364	000898	0.1034	0.6388
000157	-0.1335	0.5436	000933	-0.1398	0.5247
000401	-0.0258	0.9069	000937	0.1667	0.4472
000402	0.3064	0.1551	600000	-0.1058	0.6308
000527	0.3767	0.0764	600005	-0.3133	0.1454
000528	-0.1686	0.4419	600009	0.4053	0.0551
000568	-0.3848	0.0698	600010	0.2983	0.1668
000625	-0.0373	0.8660	600015	0.0883	0.6886
000630	0.2435	0.2628	600016	-0.1036	0.6379
000651	-0.0357	0.8715	600019	0.2004	0.3592
000709	-0.0993	0.6522	600028	0.2439	0.2620
000729	-0.0987	0.6542	600029	0.0017	0.9940
000778	0.1154	0.6001	600031	-0.2895	0.1804
000792	-0.0649	0.7686	600036	0.2189	0.3155

Notes: *** indicates correlation coefficient significant at 1% level.

Table 3.5 Granger-causality between 1-mins volatility (1-V^2) and Internet Information (II)

This table reports the Granger-causality test (Granger, 1988) between 1-mins volatility (1-V^2) and Internet Information (II). Value in the table is the F-statistics and value in the parenthese is the critical value from the F-distribution. If F-statistics is greater than the critical value, the null hypothesis of no Granger Causality is rejected at certain significance level. The lag length selection is chosen using the Bayesian information Criterion.

Stock codes	Null: II does not granger cause1-V^2	Null: 1-V^2 does not granger cause II	Stock codes	Null: II does not granger cause1-V^2	Null: 1-V^2 does not granger cause II
000002	55.86 (5.064) ***	2.12 (5.092)	000800	6.07 (4.437) ***	4.01 (4.862)
000009	28.30 (4.556) ***	11.36 (5.064) ***	000807	8.70 (4.437) ***	10.11 (4.437) ***
000012	5.42 (4.437) ***	22.37 (4.556) ***	000825	23.78 (4.437) ***	6.10 (8.096)
000024	0.18 (8.096)	3.49 (4.437)	000839	15.23 (4.437) ***	46.84 (4.556) ***
000061	5.10 (4.437) ***	21.54 (5.064) ***	000858	11.79 (4.437) ***	1.75 (6.013)
000063	11.71 (5.064) ***	5.08 (5.092)	000878	7.04 (4.437) ***	2.80 (4.695)
000069	6.23 (4.437) ***	1.23 (8.096)	000898	15.20 (4.437) ***	0.36 (8.096)
000157	3.40 (4.437)	0.95 (8.096)	000933	4.56 (4.669)	3.34 (4.556)
000401	4.79 (8.096)	0.31 (8.096)	000937	3.70 (4.437)	0.78 (8.096)
000402	14.37 (4.437) ***	2.78 (5.092)	600000	2.71 (4.437)	6.93 (4.862) ***
000527	1.21 (8.096)	3.80 (4.773)	600005	24.28 (4.437) ***	2.65 (8.185)
000528	5.78 (4.669) ***	2.57 (6.112)	600009	7.54 (4.669) ***	4.45 (4.669)
000568	73.12 (5.064) ***	17.77 (4.862) ***	600010	65.01 (4.556) ***	1.34 (8.096)
000625	15.91 (4.437) ***	13.72 (6.013) ***	600015	15.96 (5.064) ***	3.82 (4.695)
000630	3.41 (4.437)	5.54 (4.695) ***	600016	3.56 (5.064)	4.55 (5.185)
000651	0.69 (8.096)	12.80 (4.437) ***	600019	8.56 (4.437) ***	0.68 (8.096)
000709	19.89 (4.437) ***	113.0 (5.064) ***	600028	14.59 (4.437) ***	2.70 (5.926)
000729	3.75 (4.437)	5.37 (5.035) ***	600029	7.04 (8.096)	5.59 (8.400)
000778	29.22 (4.437) ***	8.29 (4.862) ***	600031	3.71 (4.437)	1.60 (6.013)
000792	0.18 (8.096)	4.37 (5.926)	600036	1.79 (8.096)	1.95 (8.285)

Notes: *** indicates Granger-causality test significant at 1% level

Table 3.6 Granger-causality between 5-mins volatility (5-V^2) and Internet Information (II)

This table reports the Granger-causality test (Granger, 1988) between 5-mins volatility (5-V^2) and Internet Information (II). Value in the table is the F-statistics and value in the parenthese is the critical value from the F-distribution. If F-statistics is greater than the critical value, the null hypothesis of no Granger Causality is rejected at certain significance level. The lag length selection is chosen using the Bayesian information Criterion.

Stock codes	Null: II does not granger cause 5-V^2	Null: 5-V^2 does not granger cause II	Stock codes	Null: II does not granger cause 5-V^2	Null: 5-V^2 does not granger cause II
000002	23.62 (5.064)***	2.00 (5.092)	000800	6.42 (4.437)***	4.24 (5.417)
000009	24.38 (4.556)***	14.23 (5.064)***	000807	10.28 (4.437)***	10.02 (4.437)***
000012	4.81 (4.4374)***	23.63 (4.556)***	000825	22.29 (4.437)***	6.15 (8.0960)
000024	0.18 (8.096)	3.41 (4.437)	000839	16.06 (4.437)***	51.53 (4.556)***
000061	4.68 (4.4374)***	22.94 (5.064)***	000858	9.86 (4.437)***	1.79 (6.013)
000063	7.26 (5.0643)***	4.81 (5.092)	000878	7.23 (4.437)***	2.79 (4.695)
000069	5.61 (4.437)***	1.14 (8.096)	000898	14.82 (4.437)***	0.39 (8.096)
000157	3.33 (4.437)	1.02 (8.096)	000933	4.47 (4.669)	3.36 (4.556)
000401	5.09 (8.096)	0.22 (8.096)	000937	3.52 (4.437)	0.78 (8.096)
000402	13.01 (4.437)***	2.73 (5.092)	600000	2.74 (4.669)	6.97 (4.862)***
000527	0.98 (8.096)	3.83 (4.773)	600005	21.86 (4.437)***	2.77 (8.1849)
000528	5.46 (4.669)***	2.45 (6.112)	600009	7.59 (4.669)***	4.55 (4.669)
000568	60.73 (5.064)***	14.86 (4.862)***	600010	4.28 (4.437)	1.01 (8.096)
000625	17.31 (4.437)***	13.85 (6.013)***	600015	13.71 (5.064)***	4.03 (4.695)
000630	3.59 (4.437)	4.87 (4.695)***	600016	4.38 (5.064)	4.91 (5.185)
000651	0.71 (8.096)	12.73 (4.437)***	600019	8.62 (4.437)***	0.70 (8.096)
000709	18.60 (4.437)***	118.0 (5.064)***	600028	14.49 (4.437)***	2.77 (5.9259)
000729	3.65 (4.437)	5.28 (5.035)***	600029	7.12 (8.096)	5.56 (8.400)
000778	32.50 (4.437)***	8.20 (4.862)***	600031	3.48 (4.437)	1.50 (6.013)
000792	0.09 (8.096)	4.16 (5.926)	600036	1.96 (8.096)	2.01 (8.285)

Notes: *** indicates Granger-causality test significant at 1% level.

Table 3.7 Granger-causality between GARCH volatility (g-V^2) and Internet Information (II)

This table reports the Granger-causality test (Granger, 1988) between GARCH volatility (g-V^2) and Internet Information (II). Value in the table is the F-statistics and value in the parenthese is the critical value from the F-distribution. If F-statistics is greater than the critical value, the null hypothesis of no Granger Causality is rejected at certain significance level. The lag length selection is chosen using the Bayesian information Criterion.

Stock codes	Null: II does not granger cause g-V^2	Null: g-V^2 does not granger cause II	Stock codes	Null: II does not granger cause g-V^2	Null: g-V^2 does not granger cause II
000002	5.09 (8.29)	3.54 (4.44)	000800	3.61(4.44)	47.90 (4.86) ***
000009	23.98 (5.06) ***	69.84 (5.06) ***	000807	0.17 (8.10)	9.50 (4.67) ***
000012	3.88 (8.10)	22.83 (4.56) ***	000825	72.89 (5.06) ***	0.79 (8.10)
000024	4.13 (8.10)	2.94 (8.10)	000839	15.74 (4.67) ***	2.54 (4.56)
000061	0.14 (8.10)	45.87 (5.06) ***	000858	5.30 (4.44) ***	1.91 (8.18)
000063	351.0 (5.06) ***	2.43 (5.09)	000878	5.73 (4.44) ***	8.45 (4.69) ***
000069	20.23 (5.06) ***	4.99 (8.10)	000898	294.17 (5.06) ***	0.06 (8.10)
000157	84.55 (6.36) ***	1.12 (8.10)	000933	5.57 (4.56) ***	0.27 (8.18)
000401	11.94 (8.10) ***	2.62 (8.10)	000937	1.91 (5.93)	2.58 (5.93)
000402	59.21 (5.06) ***	2.67 (5.93)	600000	4.95 (5.06)	6.34 (4.86) ***
000527	5.44 (5.93)	50.88 (4.56) ***	600005	310.28 (5.06) ***	2.10 (8.18)
000528	4.72 (4.44) ***	4.57 (4.69)	600009	5.11 (8.10)	10.78 (4.44) ***
000568	6.26 (4.44) ***	30.22 (4.86) ***	600010	7.09 (4.56) ***	3.53 (4.67)
000625	6.52 (4.44) ***	22.03 (6.01) ***	600015	5.24 (5.93)	2.34 (6.11)
000630	2.85 (8.10)	11.33 (4.69) ***	600016	3.62 (8.10)	1.67 (8.18)
000651	0.03 (8.10)	2.44 (4.67)	600019	59.57 (5.06) ***	0.44 (8.10)
000709	5.34 (4.44) ***	107.0 (5.06) ***	600028	19.96 (5.06) ***	2.13 (5.93)
000729	44.69 (4.69) ***	8.84 (4.86) ***	600029	4.19 (4.44)	9.85 (4.86) ***
000778	3.81 (4.44)	11.21 (4.86) ***	600031	145.41 (5.06) ***	5.57 (8.18)
000792	3.30 (4.44)	1.27 (8.10)	600036	4.91 (8.10)	1.59 (8.29)

Notes: *** indicates Granger-causality test significant at 1% level.

Table 3.8 Granger-causality between range-based volatility (r-V^2) and Internet Information (II)

This table reports the Granger-causality test (Granger, 1988) between range-based volatility (r-V^2) and Internet Information (II). Value in the table is the F-statistics and value in the parenthese is the critical value from the F-distribution. If F-statistics is greater than the critical value, the null hypothesis of no Granger Causality is rejected at certain significance level. The lag length selection is chosen using the Bayesian information Criterion.

Stock codes	Null: II does not granger cause r-V^2	Null: r-V^2 does not granger cause II	Stock codes	Null: II does not granger cause r-V^2	Null: r-V^2 does not granger cause II
000002	9.02 (4.44) ***	1.82 (8.10)	000800	73.15 (4.44) ***	2.33 (8.40)
000009	76.88 (4.56) ***	3.16 (5.06)	000807	5.34 (4.44) ***	5.78 (4.44) ***
000012	7.39 (4.44) ***	22.39 (4.56) ***	000825	65.48 (4.44) ***	4.14 (8.10)
000024	8.64 (4.44) ***	45.15 (8.10) ***	000839	17.96 (4.44) ***	53.51 (4.56) ***
000061	88.91 (4.44) ***	31.36 (5.06) ***	000858	78.47 (4.56) ***	1.76 (8.18)
000063	4.90 (4.44) ***	5.83 (4.67) ***	000878	6.22 (4.44) ***	0.35 (8.29)
000069	10.36 (4.44) ***	1.77 (8.10)	000898	87.93 (4.44) ***	0.82 (8.10)
000157	19.99 (4.44) ***	0.78 (8.10)	000933	51.36 (4.44) ***	2.36 (8.18)
000401	4.33 (4.44)	4.41 (4.67)	000937	4.07 (4.44)	1.34 (8.10)
000402	42.92 (4.44) ***	4.19 (4.44)	600000	6.05 (4.44) ***	8.66 (4.86) ***
000527	26.74 (5.06) ***	4.84 (4.77) ***	600005	69.35 (4.86) ***	3.08 (8.18)
000528	21.65 (4.44) ***	5.57 (4.69) ***	600009	16.47 (5.06) ***	0.49 (8.10)
000568	0.46 (8.10)	16.93 (6.23) ***	600010	0.06 (8.10)	1.01 (8.10)
000625	115.86 (4.44) ***	18.44 (5.18) ***	600015	11.50 (5.06) ***	1.93 (8.29)
000630	26.42 (4.44) ***	4.40 (8.29)	600016	128.5 (5.06) ***	0.28 (8.18)
000651	31.43 (4.44) ***	2.91 (4.44)	600019	25.55 (4.44) ***	0.88 (8.10)
000709	1222.16 (4.69) ***	548.5 (5.06) ***	600028	9.85 (4.44) ***	1.60 (8.10)
000729	9.26 (4.44) ***	10.49 (4.86) ***	600029	90.45 (5.21) ***	4.24 (6.23)
000778	84.20 (4.44) ***	10.08 (4.86) ***	600031	17.78 (4.44) ***	1.89 (6.01)
000792	26.92 (4.56) ***	6.80 (8.10)	600036	4.02 (4.44)	1.98 (8.29)

Notes: *** indicates Granger-causality test significant at 1% level

Table 3.9 Summary statistics for the MDH and SIAH

This table reports the number of significant stocks for the contemporaneous and causal relations. II→V^2 denotes the Granger causality from Internet Information to return volatility and V^2→II denotes the reverse direction. V^2∼II denotes the Granger causality in either direction.

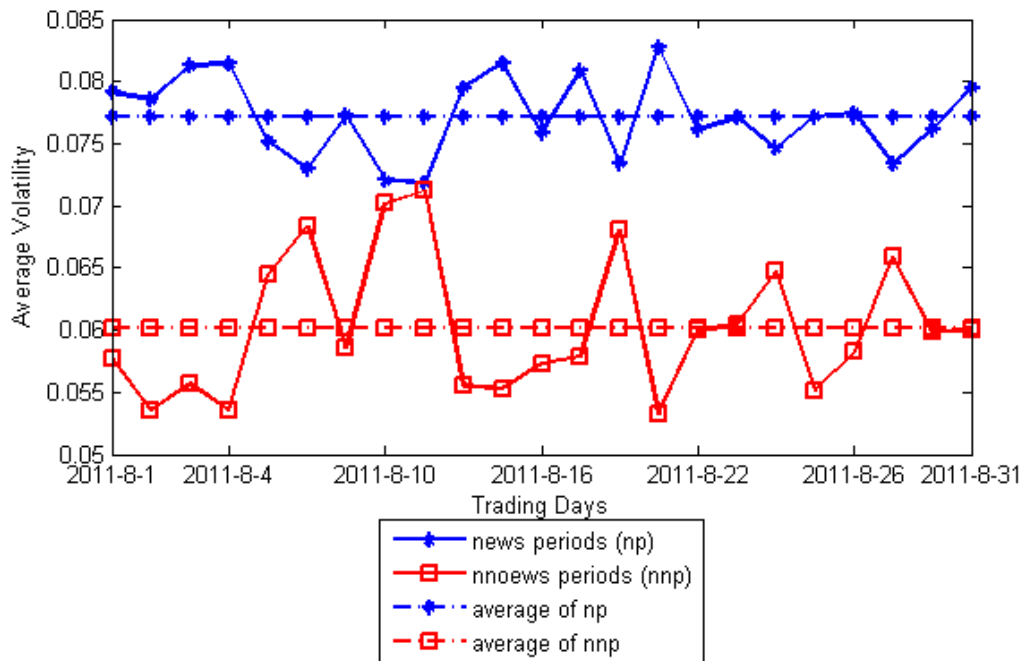| Relations	1-V^2	5-V^2	g-V^2	r-V^2
Contemporaneous correlations	2	3	7	3
Granger causality (II→V^2)	25	24	22	35
Granger causality (V^2→II)	13	13	16	14
Granger causality (V^2∼II)	29	28	32	36

Figure 3.2 Difference in average 1-min volatility between news periods and no-news periods in the morning trading period (9:30 a.m.-11:30 a.m.)
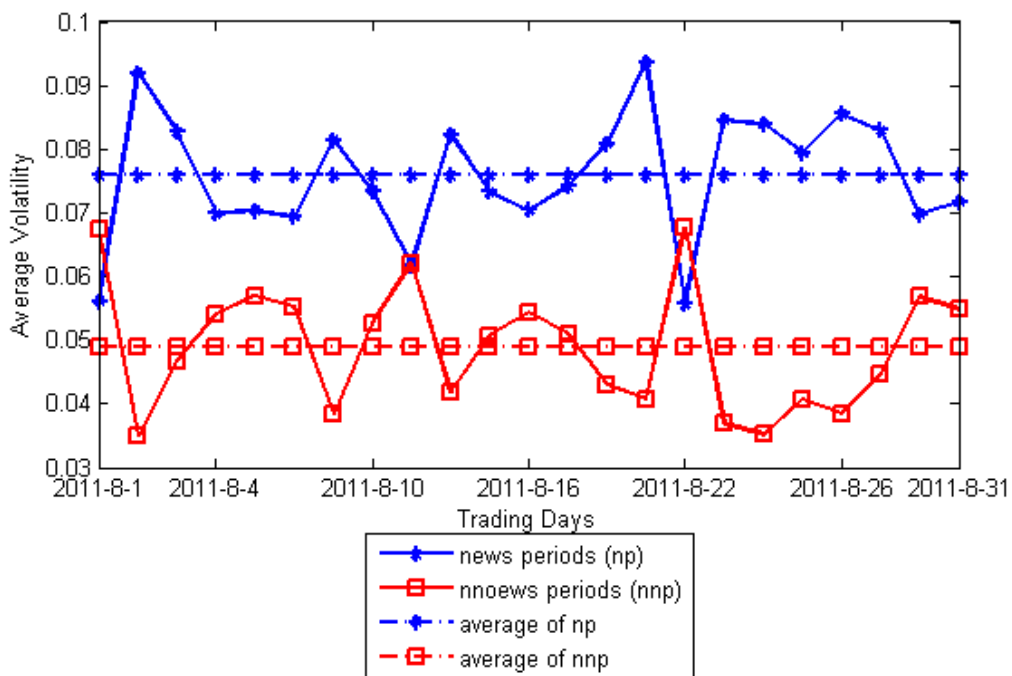
Figure 3.3 Difference in average 1-min volatility between news periods and no-news periods in the afternoon trading period (13:00 p.m.-15:00 p.m.)

3.4   Internet Information and Trading Volume

Just as the old Wall Street adage saying that "it takes volume to move prices", the trading volume is extensively expounded as the parsimonious representation of information in stock market and employed to provide evidences for various theoretical models on the information-volatility relations (Tauchen and Pitts, 1983; Varian, 1985; Admati and Pfleiderer, 1988; Harris and Raviv, 1993 and Wang 1994). Trading volume can be high either resulting from an extremely skewed distribution of seller-initiated trading and buyer-initiated trading, or due to a general trading interests among investors, which generate relatively even distribution. Intuitively, the reported trading volume of one million shares evenly split by 500, 000 seller-initiated shares and 500, 000 buyer-initiated shares may convey hugely different information from the one million shares are solely generated by the seller-initiated or buyer-initiated shares. The order imbalance can alter return by means of forcing the market maker to re-adjust their inventory as well as generating the autocorrelations in returns (Chordia *et al.*, 2002 and Chordia and Subrahmanyam, 2004). Besides, large size of the trade is always viewed as the signal of informed trading by various theoretical models (Grundy and McNichols, 1989 and Kim and Verrecchia, 1991) and its correlations with the return volatility are empirically investigated by Chan and Fong (2000). These are the reasons why the relations between the information and the characteristics and decomposition of the trading volume, i.e., the order imbalance and large size of the trade, worth investigating in this chapter.

Table 3.10 reports the definition of the order imbalance and large size of trade employed. All the transactions have been labeled as seller initiated or buyer initiated trade from the RESSET Financial Research High Frequency Database. We construct two measurements of the order imbalance, the order imbalance on the number of shares (OIBNUM) and the on the RMB value of the shares (OIMAMO). These definitions are in line with Chordia *et al.* (2002), Chordia and Subrahmanyam (2004),

Shenoy and Zhang (2007) and Bailey *et al.* (2009). Following Easley *et al.* (1997 a, b), we choose the 1,000 shares as the cut-off and designate the number of shares greater than or equal to 1,000 as the large size of trade. Two measurements are calculated as the LONUM denotes the number of trade has at least 1,000 shares and LOAMO as their RMB value of the trade.

Table 3.10 Variables definition

Variables	Definition
OIBNUM	The absolute value of the number of buy-initiated trades minus number of sell-initiated trades divided by the total number of trades
OIBAMO	The absolute value of the RMB value of buy-initiated trades minus RMB value of sell-initiated trades divided by the total RMB value of trades for the day
LONUM	The number of trades has at least 1000 shares
LOAMO	The RMB value of trade has at least 1000 shares

To control for the time-of the-day effect proposed by Amihud and Mendelson (1987) and Stoll and Whaley (1990), we exclude the opening 10 minutes and the closing 10 minutes, i.e., 9:30 a.m.-9:40 a.m. and 14:50 p.m.-15:00 p.m., to calculate the OIBNUM, OIBAMO, LONUM and LOAMO. Figure 3.4 illustrates the mean value of log trading volume for the news periods and no-news periods. In particular, the grand mean of the log trading volume of the no-news periods is 7.4693, compared to the grand mean of 7.3634 for the log trading volume in the news periods. This difference is highly significant at 1% (t-value=3.5223). Thus, more trading activities happen during the no-news period in which investors' trade solely on their private information. Figure 3.5, Figure 3.6, Figure 3.7 and Figure 3.8 illustrate the differences in daily OIBNUM, OIBAMO, LONUM and LOAMO between news periods and no-news periods, respectively. Table 3.11 summarizes the statistics for the differences between news period and no-news periods. The empirical results show that the LONUM and LOAMO in the no-news periods are significant larger than that in the news periods. The differences are highly significant at 1%. These findings suggest that the active trading in the no-news periods is generated by the large size of the trade. The difference between order imbalance in

32

news periods and no-news periods is less significant compared to the large size of trade. Only the OIBAMO significant at 10% (t-value=1.9226). As a whole, the trading activity in the no-news periods is probably driven by institutional investor.
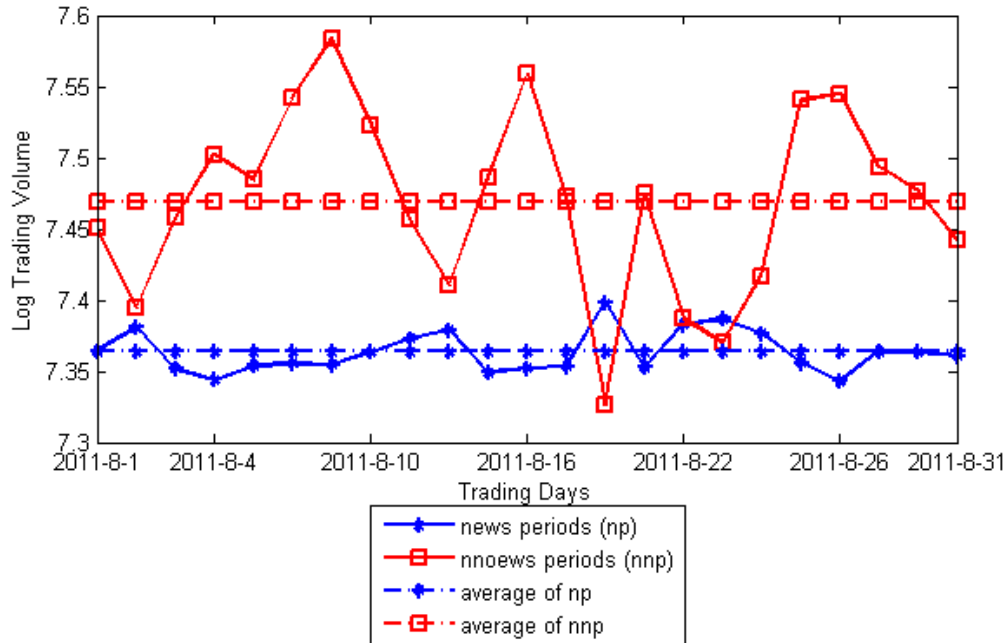
Figure 3.4 Differences in daily trading volume between news periods and no-news periods

This figure illustrates the mean value of the log trading volume for the periods with and without Internet Information and the grand means are also plotted. The two-sample t-test shows a significant larger trading volume in the no-news period than that of the news periods with t-value= 3.5223 and p-value=0.0019.

Table 3.11 Summary of the measurements of order imbalance and large size of trade in news and no news periods

This table reports the differences of OIBNUM, OIBAMO, LONUM and LOAMO in news periods and no-news periods. The two-sample t-test is performed to test the significances. The large size of trading in the no-news period is significant larger than that of in the news periods. But the difference of order imbalance is less significant compared to the large size of trade.

Variables	News periods	No news periods	Differences	t-value	p-value
OIBNUM	0.1932	0.1919	-0.0013	0.2502	0.8048
OIBAMO	0.1906	0.1765	-0.0141*	1.9226	0.0676
LONUM	0.4157	0.4484	0.0327***	-7.4701	0.0000
LOAMO	0.8822	0.8960	0.0138***	-6.6742	0.0000

Notes: *, *** indicates significant at 10% and1% level, respectively.

Figure 3.5 Differences in daily OIBNUM between news periods and no-news periods
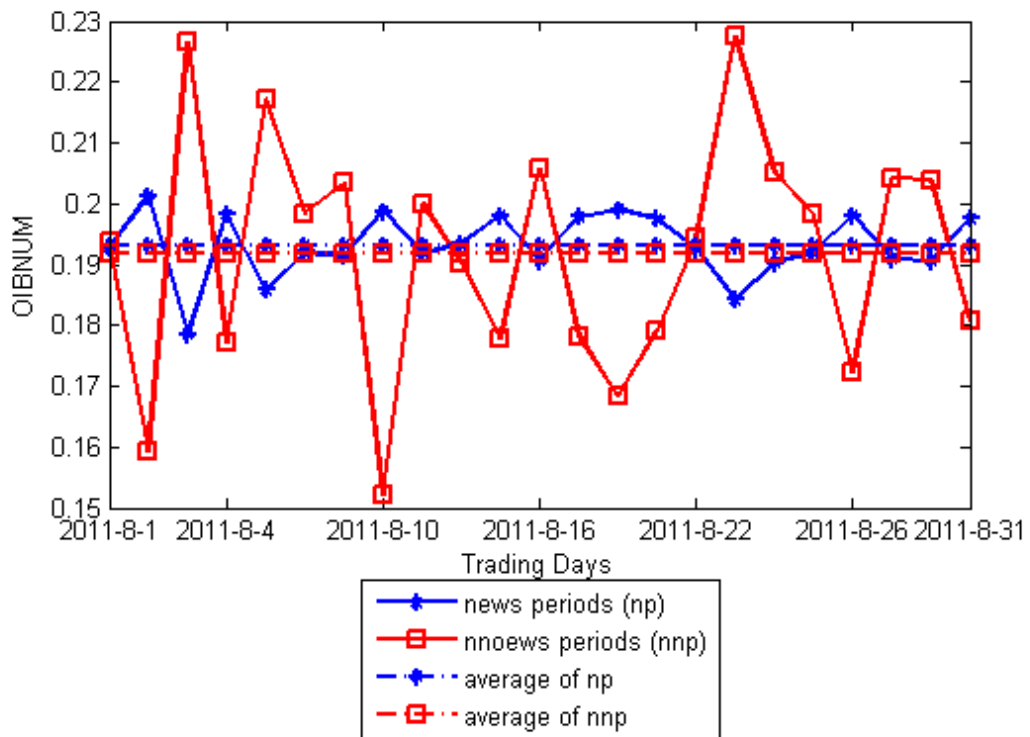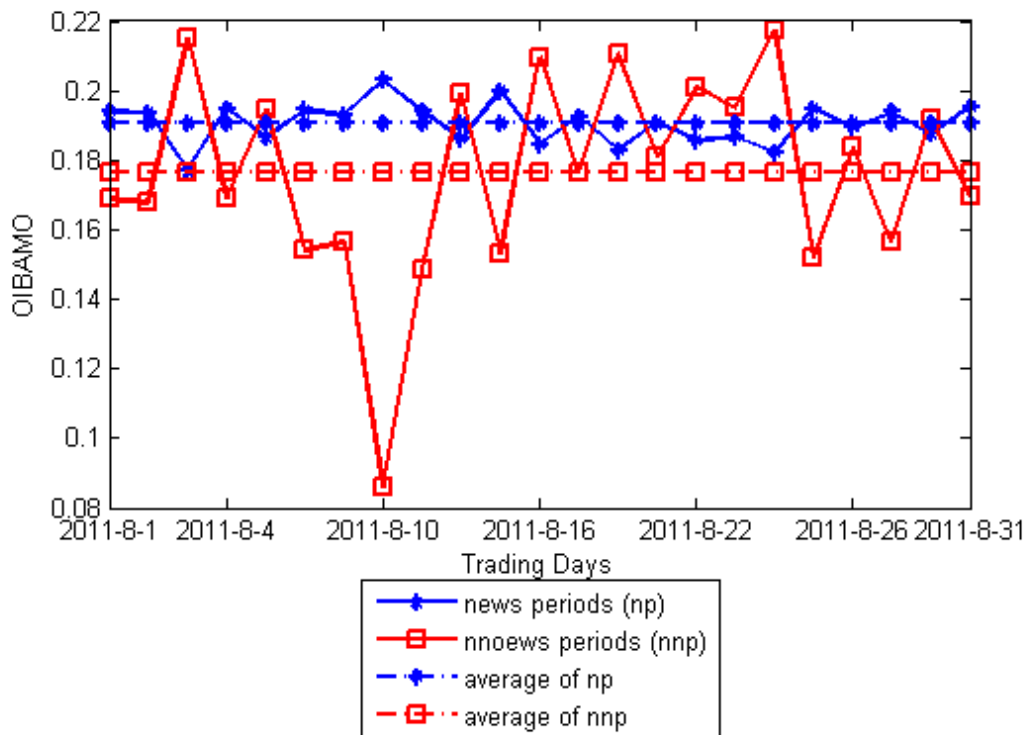

Figure 3.6 Differences in daily OIBAMO between news periods and no-news periods

34
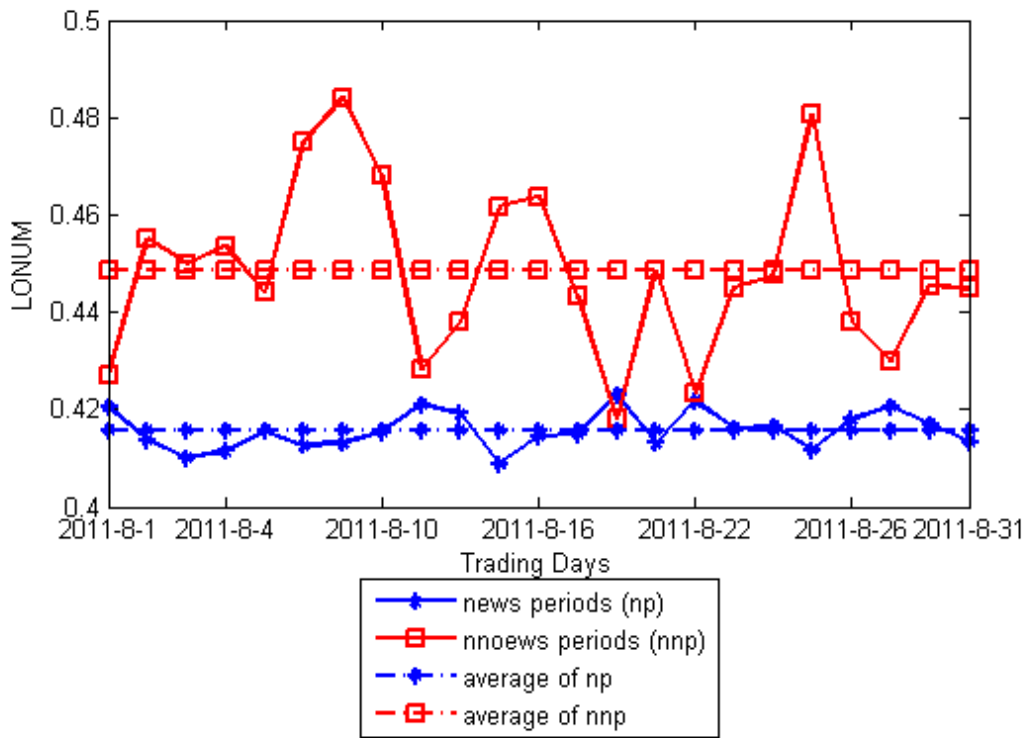
Figure 3.7 Differences in daily LONUM between news periods and no-news periods
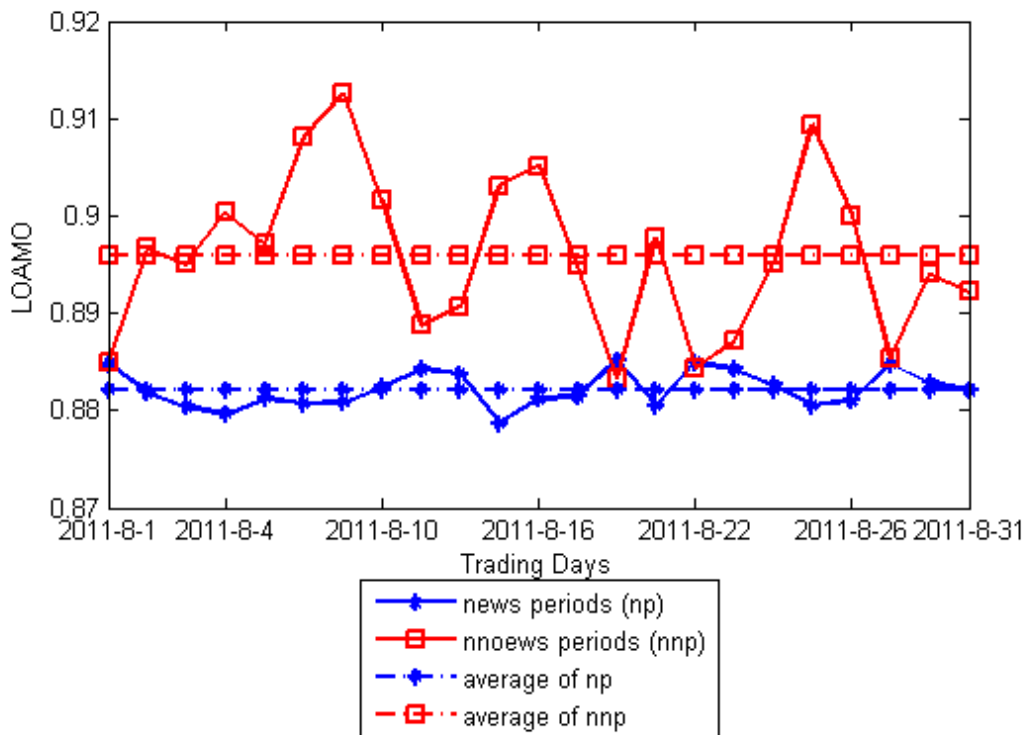
Figure 3.8 Differences in daily LOAMO between news periods and no-news periods

3.5 Robustness

As a robustness check, we further investigate the order imbalance and large size of trading with partitioned sessions, i.e., the morning trading session and the afternoon trading session. The time-of the-day effect is also controlled by excluded the trading periods 9:30 a.m.-9:40 a.m. and 14:50 p.m.-15:00 p.m., respectively. Table 3.12 reports the differences of order imbalance and large size of trade in news and no-news periods in both morning trading session and afternoon trading session. The empirical results confirm our previous findings that the large size of trading is larger in the no-news periods than that in the news periods. These differences are highly significant at 1%. Only the OIBAMO changes significantly in different news periods. In that sense, we can conclude that the more active trading in the no-news periods is generated by the large size of trade and trading activity is not solely driven by information.

Table 3.12 Summary of the measurements of order imbalance and large size of trade in news and no news periods with partitioned sessions

This table reports the differences of OIBNUM, OIBAMO, LONUM and LOAMO in news periods and no news periods with partitioned sessions, i.e., the morning trading session and the afternoon trading session. The two-sample t-test is performed to test the significances. The large size of trading in the no news period is significant larger than that of in the news periods. But the difference of order imbalance is less significant compared to the large size of trade.

Variables	News periods	No news periods	Differences	t-value	p-value
Panel A: morning trading session (9:40 a.m.-11:30 a.m.)					
OIBNUM	0.2009	0.2004	-0.0005	0.0879	0.9308
OIBAMO	0.1892	0.1759	-0.0133**	2.2100	0.0378
LONUM	0.4249	0.4471	0.0222***	-5.6861	0.0000
LOAMO	0.8835	0.8942	0.0107***	-5.2878	0.0000
Panel B: afternoon trading session (13:00 p.m.-14:50 p.m.)					
OIBNUM	0.2396	0.2333	-0.0063	0.9994	0.3285
OIBAMO	0.2653	0.2257	-0.0396***	5.4221	0.0000
LONUM	0.3979	0.4237	0.0258***	-9.4839	0.0000
LOAMO	0.8722	0.8859	0.0137***	-8.7868	0.0000

Notes: **, *** indicates significant at 5% and1% level, respectively.

3.6 Conclusions Remarks

In this chapter, we advocate the search frequency of stock names in Baidu News as the direct proxy for information flow and examine the information-volatility relations. With various measurements of return volatility, i.e., 1-min volatility, 5-mins volatility, GARCH volatility and range-based volatility, we find strong lead-lag relations between the information flow and return volatility and only a few stocks (less than 10%) exhibit the contemporaneous relations between information flow and return volatility. Generally speaking, these findings support the SIAH. Besides, we observe that the volatility in the news periods is significant higher than the volatility in the no-news periods. By decomposing the trading volume into the measurements of order imbalance and large size of trade, we further investigate the differences of OIBNUM, OIBAMO, LONUM and LOAMO in the news periods and no-news periods. The empirical results revel that there are more large size of trading in the no-news periods and the differences of order imbalance are less significant. Overall speaking, this chapter contributes to the existing literature by introducing a novel proxy for information flow, providing an alternative evidence of the SIAH in Chinese stock market and giving the first empirical results on the differences of order imbalance and large size of trading during news periods and no news periods.

This chapter is an initial step in illustrating the mechanism of the information-volatility, i.e., MDH and SIAH with Internet Information. Since Internet Information provide us with a large amount of firm-specific and market-wide information, we can further examine the market reaction to various categories of information and gain a paranoiac understanding of the market dynamics.

4. The Impact of Information-based Familiarity on the Stock Market

Is the familiarity-based investment a rational response to information or just a reflection of investors' nonfinancial attributes? Given the lacking of appropriate data to measure familiarity, the nature of the familiarity-based investment is less investigated. In this chapter, by utilizing a unique online information interaction channel organized by Shenzhen Stock Exchange, we show that the familiarity-based investment in Chinese stock market is driven by information diffusion and has a material impact on the subsequent liquidity and volatility.

4.1 Introduction

The legendary investors Peter Lynch and Warren Buffet, both ascribe their success to the strategy of "buy what you know". Is it pitfall or valuable advice for investors? Mounting evidence has shown the significant differences between optimal diversified portfolios suggested by traditional investment theories and the observed biased ones held by majority of real-world investors, i.e., investors tend to limit their investment in stocks which they are familiar with.

Previous literature has provided plenty of evidence for investors' familiarity-driven behavior in constructing portfolios across different countries (French and Porterba, 1991; Gehrig, 1993; Cooper and Kaplanis, 1994; Tesar and Werner, 1995; Kang and Stulz, 1997; Coval and Moskowitz, 1999; Coval and Moskowitz, 2001; Huberman, 2001; Seasholes and Zhu, 2010 and Døskeland and Hvide, 2011). Though these studies conclude several reasons for familiarity-driven behavior, i.e., investors tend to choose stocks that have geographic, professional or cultural proximities, the underlying nature of familiarity-driven investment have not been investigated among empirical studies. Theoretically, as biased behavior, familiarity may be caused by information or pure heuristic behavior (Heath and Tversky, 1991). In the case of information-based familiarity, the geographic,

professional and cultural closeness are cheap ways of acquiring information for investors, i.e., investors are found to have strong home-bias tendency in making investment since they actually have comparative advantages in acquiring local information compared with other foreign investors. While the behavioral story of familiarity posits that the aforementioned closeness merely makes the stocks more salient and recognizable to investors and the feeling of enjoying comparative advantages result from the behavioral-based familiarity is illusionary, i.e., investors may not benefits from familiarity-based investment since they do not have informational advantages.

It has become an increasingly important issue to explore the nature of familiarity since recent studies have shown that familiarity-based behavior is powerful enough to affect the stock market. For example, Loughran and Schultz (2005) find that there is a positive relationship between investors' familiarity with a certain stock and its liquidity. Consistently, using advertising cost as a proxy for investor's familiarity, Grullon *et al.* (2004) strongly advocate that, for a particular company, both the liquidity and shareholder base of its common stock will rise associated with an increase in its advertising expenditure. However, due to the limitation of dataset, to the best of our knowledge, Massa and Simonov (2006) is the only one that conducts an empirical study with a direct proxy known as professional change to investigate the nature of familiarity in one of the developed countries, i.e., Sweden. With a unique data set of Swedish investors broken down at disaggregate levels that involves individuals' wealth, income, professions and other demographic characteristics, they state that the nature of familiarity among Swedish households is information-based. It is obvious that the nature of familiarity among developed markets has started to attract academia's attention and needs deeper and more profound investigation, while the story of familiarity remains unfolded in developing countries.

Exploring such issues in Chinese stock markets is necessary and crucial since a

completely different picture may be drawn. Firstly, both market-level (e.g., property rights protections and government index) and firm-level (e.g., earnings management and accounting manipulation) information environment in Chinese stock market is less efficient and transparent compared with those of developed markets (Morck *et al.*, 2000; 2004; Kevin and Yuan, 2004 and Jin and Myres, 2006). Secondly, Chinese stock market possesses larger number of irrational individual investors who are subject to strong behavioral biases and thus make the market more speculative (Feng and Seasholes, 2008 and Lee and Liu, 2011). Thirdly, the information is not fully reflected due to the short sale constrain and price limit policy (Miller, 1977). Therefore, in this paper, we not only bridge this gap by investigating the nature of individuals' familiarity-based behavior in Chinese stock market, but also provide alternative evidence for the impacts of familiarity-based investment on the market liquidity and volatility.

We contribute to the existing literature in following aspects. Firstly, along with Massa and Simonov (2006), we fill the gap by conducting the first empirical study to test the nature of familiarity of individual investors in China. As a developing market, the Chinese stock market is more gambling-driven and involves a considerable number of unsophisticated individual investors compared with the developed markets. Secondly, current studies about familiarity either directly focus on exploring its nature (Massa and Simonov, 2006) or try to examine its possible impact on stock market performance with cross-country evidences (Loughran and Schultz, 2005; Grullon *et al.*, 2004; Graham *et al.*, 2009; Jacobs and Weber, 2011). Our unique data derived from the "familiarity shock" events that track individual investors' familiarity changes brought, not only allow us to directly test our hypothesis about the nature of investors' familiarity, but also provide us a rare opportunity to explore the short-term impact of familiarity on the liquidity and volatility. We believe that conducting such combined studies on both the nature and impact of familiarity rather than merely focusing on one aspect as previous studies may bring more valuable managerial implications for policy makers.

4.1.1 Evidence for Familiarity

Familiarity is often described as investors' tendency to purchase certain stocks and thus construct under-diversified portfolios due to different reasons, e.g., geographic proximity, professional closeness, cultural patriotism. Specifically, Huberman (2001) defines the concept of familiarity as "the general tendency of households' portfolios to be concentrated, of employees' tendency to own their employers' stocks in their retirement accounts, and to the home country bias in the international arena". French and Poterba (1991) observe that investors in the U.S, U.K as well as Japan are far more optimistic about domestic stocks than those foreign ones, thereby leading to the tendency to substantially overweight domestic stocks when constructing investment portfolios. Similarly, using data from five OECD countries (Canada, Germany, Japan, the U.K and the U.S), Tesar and Werner (1995) not only present strong evidence for investors' home bias behavior in making long-term international investment decisions, but also conclude that the increase in international investment positions resulting from lower transaction costs and policy relaxation is marginal and inadequate to meet the appropriate proportion predicated by theories. This phenomenon of eschewing foreign stocks due to geographic and cultural proximities has been defined as the home bias puzzle and has been found across different countries for both individual and institutional investors (Gehrig, 1993; Cooper and Kaplanis, 1994; Kang and Stulz, 1997; Lewis,1999) Moreover, following research conducted by Coval and Moskowitz (1999, 2001) amplifies the context of home bias by showing that the U.S institutional investors even tilt their stockholdings toward locally headquartered firms. Such "home bias at home" phenomenon has been dubbed local bias and attracted a variety of research thereafter. For example, having taken into account of geographic distribution of the seven U.S RBOC (Regional Bell Operating Companies), Huberman (2001) concludes that the households tend to hold more shares of the local RBOC than other non-local ones. More importantly, Grinblatt and Keloharju (2001) not only document that Finnish investors are more likely to hold and trade stocks whose headquarters locate nearby, but also observe a negative relationship between local bias and investors'

sophistication. Ivkovic and Weisbernner (2005, 2008) provide further evidence for the existence of local bias in the U.S market, and they start to unfold the underlying veils of local bias with the observed result of 3.2% additional returns for locally biased households. Seasholes and Zhu (2010) use two types of calendar time portfolios to test the existence of local bias, i.e., one is based on holdings and the other one is based on transaction records, and conversely, they find locally biased portfolios have underperformed returns compared with the optimal ones. As for stocks in different industries, Nofsinger and Varma (2012) find individual investors are four times more likely to buy stocks of local utility firms compared with that of utility companies out of their state of residence.

As is mentioned previously, familiarity manifests itself in more than home (local) bias in existing literature. Interestingly, employees have a strong willing to invest the retirement money into their company's stock regardless of its performance and other available options suggested in their pension plans (Benartzi, 2001; Huberman, 2001). Kahn (1997) documents that two thirds of 401(k) plans' asset of employees working at Mercury Finance are invested into Mercury's fast declining stock. Notably, participants taking the John Hancock-Gallup survey are even more optimistic about their employer's stock than domestic or local stocks, and they exhibit lowest trust for foreign stocks. Moreover, using the transaction data for all individual investors in Norwegian stock market, Døskeland and Hvide (2011) observe that investors still hold 11% of their portfolio in stocks of their professional industry after excluding the proportion invested in their employer's company stock, through which they demonstrate strong familiarity bias to stocks with professional proximity.

4.1.2 Discussion on Nature of Familiarity

The nature of familiarity stems from the famous psychological theory-"the Competence Hypothesis" proposed by Heath and Tversky (1991). The theory states that "holding judged probability constant, people prefer to bet in a context where

they consider themselves knowledgeable or competent than in a context where they feel ignorant or uninformed". In that sense, they suggest that the competence hypothesis may be applied to explain the puzzling fact that investors are often willing to sacrifice the benefits of diversification and tend to focus on stocks that they consider themselves familiar with. Since the feeling of competence in a situation depends on what has been known relative to what can be known, individuals may not have the ability to precisely realize the unknown information. That is, the bounded rational investors may consider themselves enjoying information advantages brought by geographic, cultural or professional proximities, whereas in fact their information situation has not been changed and their feeling of being competitive is merely illusionary. Such ambiguous interpretation of competence hypothesis has attributed the nature of familiarity to either information or pure heuristic behavior.

The information-based familiarity could be traced back to Merton (1987), holding the opinion that "investors buy and hold only those securities about which they have enough information". By showing that investors require higher expected returns for stocks that are less familiar to them, Merton implicitly attribute the "nature of familiarity" to information. However, theory of "ambiguity aversion" in the field of psychology explains investors' familiarity with pure heuristic behavior (Bossaerts *et al.*, 2010 and Ahn *et al.*, 2011). The theory suggests that individuals' ambiguity aversion tends to be excessively high under the comparative situation where they have the opportunities to evaluate the relative certainty and uncertainty simultaneously (Fox and Tversky, 1995). In most cases, investors in stock market are in the comparative situation and face both familiar stocks as well as unfamiliar ones at the same time; hence investors may exhibit intuitive intolerance to uncertain stocks and prefer familiar ones regardless of the diversification benefits. Accordingly, Boyle *et al.* (2012) develop a theoretical model to explain the nature of familiarity with the ambiguous aversion. However, empirical studies on exploring nature of familiarity are narrowed due to the limitation of data.

To the best of our knowledge, Massa and Simonov (2006) is the only study which directly investigates the nature of familiarity among investors. According to their hypothesis, the information-based familiarity should satisfy two criteria. Firstly, the performance of investors' holdings should changes after the "familiarity shocks". Secondly, the prices changes to "familiarity shocks" should differ across investors with different magnitude of familiarity. In particular, the stocks with high magnitude of familiarity among investors should have more profit than that of the stocks with low magnitude of familiarity among investors if the "familiarity shocks" are the events to increase the amount of information. To implement the first criterion, Massa and Simonov (2006) relate the magnitude of familiarity to investors' wealth by literally assuming that the wealthier investors are more willing to pay for private information and hence have more information compared with poorer investors who are likely dependent on cheap public information. In that sense, investors' preference on familiar stocks results from actual information advantages may help under-diversified investors earn more abnormal profit than others. While in the context of pure heuristic behavior, investors' tilt to more familiar stocks may result from a feeling of being competent or affinitive brought by geographic, professional, cultural proximities. Since competitive advantages may be illusionary and subjective to individuals, heuristic-based familiarity could not guarantee the profit for investors holding under-diversified portfolios. That is, investors, especially individual investors, may face less abnormal returns or even experience a loss through conducting under-diversified portfolios based on pure heuristic familiarity. In line with the hypothesis, Massa and Simonov (2006) observe more profit for wealthier investors who make familiarity-based investment, indicating that the nature of familiarity is information rather than pure heuristic behavior. Referring to the second criterion, the "familiarity shocks" are explained as the events that can alter the investors' proximity to familiarity (Massa and Simonov, 2006). For example, changes of investors' professions and locations may cause investors to lose their familiarity to previous situations and adapt to new positions. Therefore, investors who experience "familiarity shocks" are likely to change their investment decisions

according to the altered familiarity conditions to certain stocks and such behavior may lead to persistent price changes. On the other hand, if familiarity is pure heuristic, the "familiarity shocks" may not change investors' familiar feeling about certain stocks.

4.2   Data Description

There are mainly two sources of data in this chapter. The first data refers to the measurement of familiarity. The Shenzhen Stock Exchange organizes the "Approach your company, know your investment" activities, which invite the stock holders, i.e., individual investors to have a face-to-face talk with the owner of the company. Therefore, these activities can naturally be considered as the "familiarity shocks" for investors. The second data is the capital data and is retrieved from the RESSET Financial Research Database.

4.2.1   Representative Magnitude of Familiarity

The "familiarity shock" event refers to an activity named "Approach your company, know your investment" which literally changes investors familiarity to certain stocks. This activity is a part of the online education and investor relationship management organized by Shenzhen Stock Exchange. The purpose of the activity is to build the bridge between investors, especially individual investors, and listed companies, through which offers investors a rare opportunity to further explore their invested firms. In that sense, investors' familiarity to a certain stock involved in each event will increase if he actually takes part in the event. At the organizing stage for each event, Shenzhen stock exchange publishes an announcement containing the name and general description of the firm that is under exploration, both on its own website and major Chinese security companies' websites, generally including the CITIC Securities, Haitong Securities, Guangfa Securities, China Merchant Securities and Guotai Junan Securities, which are the top five security companies in China. Individual investors who are interested in the announced firm can then apply to take part in the upcoming activity through their belonged security companies. However,

only the investors who currently hold the stock of the announced company have the right to apply, and their belonged security firms are responsible for selecting the appropriate investors according to certain criteria. In most cases, the security firms first construct rankings in a descending order for the interested investors by taking into account of their profitability, financial savvy and sophistication, size of their investment as well as other related factors. Following that, top-tier ranked investors are selected and invited to the event. This nonperiodic activity is firstly organized on 25 May 2012 and there are 144 activities through 29 April 2015. During the activity, a selected group of investors are invited to go to the headquarters of the pre-decided firms, and the whole touring process may contain production or service line visits, senior management team or executive team introduction session as well as an interactive discussion session where the executive team will answer investors' proposed questions face-to-face.

Since the breadth of ownership of individual investors is distinct from others, it is inappropriate to use the raw number of participated individual investor as the proxy for familiarity. Therefore, the Representative Magnitude of Familiarity (RMF) is defined as the number of participated individual investors divided by the percentage of the individual investor to measure the magnitude of familiarity among individual investor for a given stock. Owing to the lacking of statistic of individual investor, we use the nearest ex ante announcement on the total number and percentage of institutional investor to calculate the percentage of individual investor. The total number and share of institutional investor are announced quarterly, i.e., on 31 March, 30 June, 30 September and 31 December in each year. For example, if the "Approach your company, know your investment" activity happens on 13 July, the announced number and percentage of institutional investor on 30 June are employed to calculate the percentage of individual investor. In order to eliminate the potential price impact of earning announcements, price limits and trading halts, the stocks with earning announcements, price limits and experiences trading halts one week around the activity event are excluded from the initial sample. This leaves us with

139 stocks in the final sample. Figure 4.1 illustrates the *RMF* for all the stocks. As is shown, the *RMF* for some stocks are distinct from others. The maximum, minimum and mean is 286.27, 22.21 and 82.17, respectively.
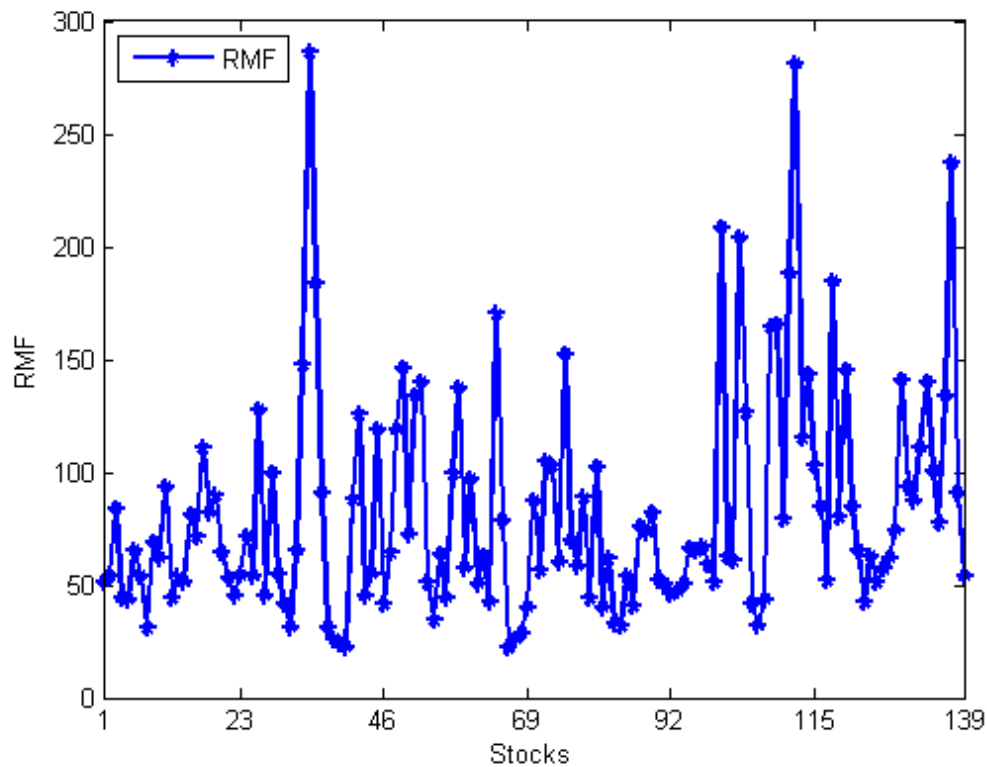
Figure 4.1 RMF of all stocks

This figure illustrates the *RMF* for all stocks in the final sample. The *RMF* is defined as the number of participated individual investors divided by the percentage of the individual investor. The maximum, minimum and mean is 286.27, 22.21 and 82.17, respectively. The two-sample t-test shows that the *RMF* is significantly different from the raw number of participated individual investor at 1% level with p-value=0.000 and t-value=5.174.

4.2.2 Capital Data

In order to give sufficient previous trading days for the event study methodology used in the following section, the sample period of capital data is enlarged to from 1 September 2011 to 30 September 2015. The capital data includes firm characteristics variables (firm age, market capitalization and PE ratio), Shanghai Shenzhen CSI 300 Index return, the liquidity measurement (turnover) and a variety of measurements of volatility (Garch-volatility, Ewma-volatility, Sma20-volatility, Sma60-volatility and Sma120-volatility). All these data comes from the RESSET Financial Research

Database. Table 4.1 shows the descriptive statistics for all the stocks experience the familiarity shocks. The maximum, minimum and median value of the firm age, market capitalization and PE ratio show that our sample containing firms with a wide range of characteristics. Besides, it is obviously that all the firms' characteristics variables are with skewed distributions. Consequently, the log-transformations are used for the empirical analysis.

Table 4.1 Summary statistics around ([-125, 50]) the familiarity shocks

Variables	Mean	Std. Dev	Min	Median	Max	Observations
Firm charcteristics						
Firm age (years)	8.8489	5.5269	2	6	23	24464
Market capitalization	11075	14477	862	5403	110030	24464
Individual returns	0.0023	0.0319	-0.1003	0.0018	0.9237	24464
PE ratio	51.7	59.5	-92.5	38.6	1213.2	24464
Liquidity measures						
Turnover	3.089	4.3563	0.0191	1.7478	95.0718	24464
Volatility measures						
Garch-volatility	0.0268	0.0074	0.0128	0.0258	0.0704	24464
Ewma-volatility	0.0258	0.0093	0.0080	0.0239	0.0710	24464
Sma20-volatility	0.0253	0.0101	0.0059	0.0233	0.0722	24464
Sma60-volatility	0.0258	0.0083	0.0096	0.0242	0.0639	24464
Sma120-volatility	0.0260	0.0071	0.0123	0.0247	0.0532	24464

Note: the market capitalization is in million RMB.

4.3 Main Findings: Information-based Familiarity

This section provides the main findings of this chapter. Firstly, the empirical results support the familiarity-based investment in Chinese stock market is driven by information. Secondly, this information driven investment has a material impact on the liquidity and volatility of underlying stocks.

4.3.1 Information-based Explanation

According to Massa and Simonov (2006), the crucial difference between information-based familiarity and heuristic-based familiarity is the role of information in the decision-making. In the case of information-based familiarity, familiarity serves as a way of providing more reliable information to investors. In the

case of heuristic-based familiarity, on the other hand, investors pay more attention to the information that is repeatedly mentioned to them. In that sense, an information-based story postulates that the activity reduce the cost for investors to acquire information and thus investor with more information would experience more price changes, whereas a heuristic story posits that participating this activity will increase the impression of the stock and thus priori information does not matter. Therefore, in order to prove the nature of familiarity-based investment is information-based, we need to provide the following two evidences. Firstly, if familiarity is information-based, investors with higher magnitude of familiarity to certain stocks should yield more returns compared with investors with lower magnitude of familiarity. Secondly, if the familiarity-based investment is driven by information, the "familiarity shock" will result in a persistent price changes without reversal.

A number of studies have documented the stock market reaction to various kinds of mass media and summarized two main hypotheses to the underlying mechanisms (Barber and Loeffler, 1993; Mathur and Waheed, 1995; Albert and Smaby, 1996; Liang, 1999 and Kerl and Walter, 2007). The Price Pressure Hypothesis (PPH) poses that news can generate temporary buying pressure in the highlighted stocks and this buying pressure causes the observed abnormal returns, which will reverse to their fundamental value in a relatively short period. The Information Diffusion Hypothesis (IDH) poses that the news reveals some relevant information about the fundamentals and thus the observed abnormal returns will not reverse to their fundamental value in a relatively short period. Both the PPH and IDH have gained increasing popularity and are considered as the theoretical foundations on investigating the stock market reaction to news is driven by information or psychological bias. In this section, we connect these hypotheses to investigating the nature of familiarity-based investment in Chinese stock market.

Furthermore, we divide the final 139 familiarity shocks into two subgroups, the

high RMF ($HRMF$) and low RMF ($LRMF$) based on RMF illustrated in Figure 4.1. To give a clear cut classification, we select the highest 60 familiarity shocks as the $HRMF$ subgroup and the lowest 60 familiarity shocks as $LRMF$ subgroup. In accordance with former criterion, if the familiarity-based investment is driven by information, we expect that both subgroups experience price changes after these familiarity shocks and the price changes in the $HRMF$ subgroup are larger than that of the $LRMF$ subgroup.

4.3.2 Cumulative Abnormal Return

To empirical examine the price changes after the familiarity shocks, we employ the event study methodology with the market model (Brown and Warner 1980; Brown and Warner 1985 and Boehmer *et al,.* 1991) to observe changes in returns of $HRMF$ subgroup and $LRMF$ subgroup. Therefore, the abnormal return for stock i on date t, AR_{it} is calculated as follows:

$$R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it} \quad t\text{=-125,\ldots,-26}$$
(4-1)

$$AR_{it} = R_{it} - (\alpha_i + \beta_i R_{mt}) \qquad\qquad t\text{=-25,\ldots,40}$$
(4-2)

where R_{it} is the return for stock i on date t, R_{mt} is the return of Shanghai Shenzhen CSI 300 Index on date t, both α_i and β_i are the ordinary least squares (OLS) estimates for the stock i's market model parameters. α_i and β_i are estimated over a period that extends from 125 trading days prior through 26 trading days prior to the familiarity shocks and AR_{it} is calculated for the days [-25, 40]. The choice of the estimation windows is consistent with Barber and Loeffler (1993) and Albert and Smaby (1996).

For the $HRMF$ subgroup and the $LRMF$ subgroup, the average abnormal return (AAR_t) for date t and the cumulative abnormal return from date t_1 to date t_2 ($CAR(t_1, t_2)$) are calculated as follows:

$$AAR_t = \frac{\sum_{i=1}^{N} AR_{it}}{N} \quad t\text{=-25,...,40} \tag{4-3}$$

$$\text{CAR}(t_1, t_2) = \sum_{t_1}^{t_2} AAR_t \quad t_1\text{=0, } t_2\text{ =40} \tag{4-4}$$

where N denotes the number of familiarity shocks in the $HRMF$ subgroup and the $LRMF$ subgroup. Therefore, in these settings, the empirical results are based on 60 familiarity shocks in the $HRMF$ subgroup and 60 familiarity shocks in the $LRMF$ subgroup.

Figure 4.2 illustrates the price changes in CAR after the familiarity shocks for the $HRMF$ subgroup and the $LRMF$ subgroup. Both of them experience significant increases in the CAR in the subsequent 40 trading days. The mean of the CAR are 0.0610 (t-value=17.32 and p-value=0.000) and 0.0439 (t-value=16.67 and p-value=0.000) for the $HRMF$ subgroup and the $LRMF$ subgroup, respectively. Besides, the CAR of the $HRMF$ subgroup is significant larger than that of the $LRMF$ subgroup at 1% level with the t-value=10.86 and p-value=0.000. In a further analysis, we also extend the subsequent event windows to 50 trading days for all the stocks that experience the familiarity shocks to observe the price changes. Figure 4.3 illustrates that the price changes is persistent and there is no return reversal. As argued by Albert and Smaby (1996), the price changes generated by the price pressure cannot be persistent in the subsequent 50 trading days. In that sense, we can conclude that the price changes are driven by information diffusion. These patterns suggest that the stocks with high familiarity among individual investors have larger price changes than that of stocks with low familiarity among individual investors and there are significant price changes after the familiarity shocks. Inferred from the above-mention criteria on investigating the nature of familiarity-based investment, the empirical results support the information-based familiarity in Chinese stock market.
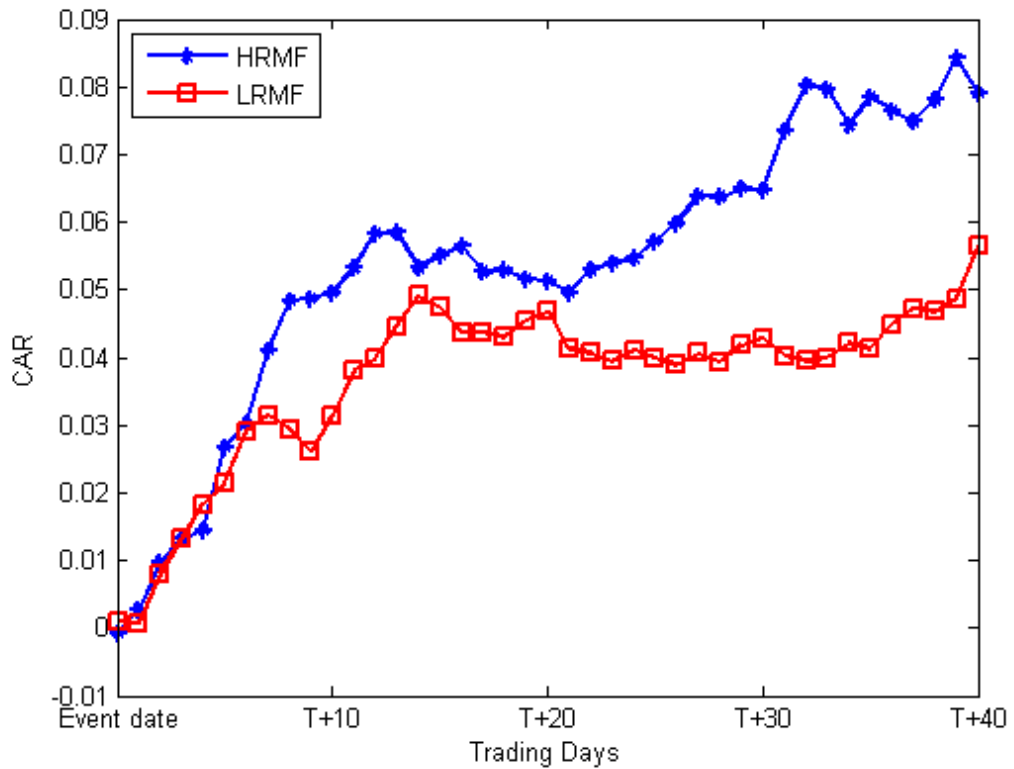
Figure 4.2 Cumulative abnormal return of *HRMF* subgroup and the *LRMF* subgroup

This figure illustrates the price changes in *CAR* after the familiarity shocks for the *HRMF* subgroup and the *LRMF* subgroup. The two-sample t-test is performed to test the statistical difference from zero for each subgroup. The mean of the *CAR* in the *HRMF* subgroup is 0.0610, which is significant different from zero at 1% level with t-value=17.32 and p-value=0.000. The mean of the *CAR* in the *LRMF* subgroup is 0.0439, which is also significant different from zero at 1% level with t-value=16.67 and p-value=0.000. Besides, the *CAR* of the *HRMF* subgroup is significant larger than that of the *LRMF* subgroup at 1% significant level with the t-value=10.86 and p-value=0.000.
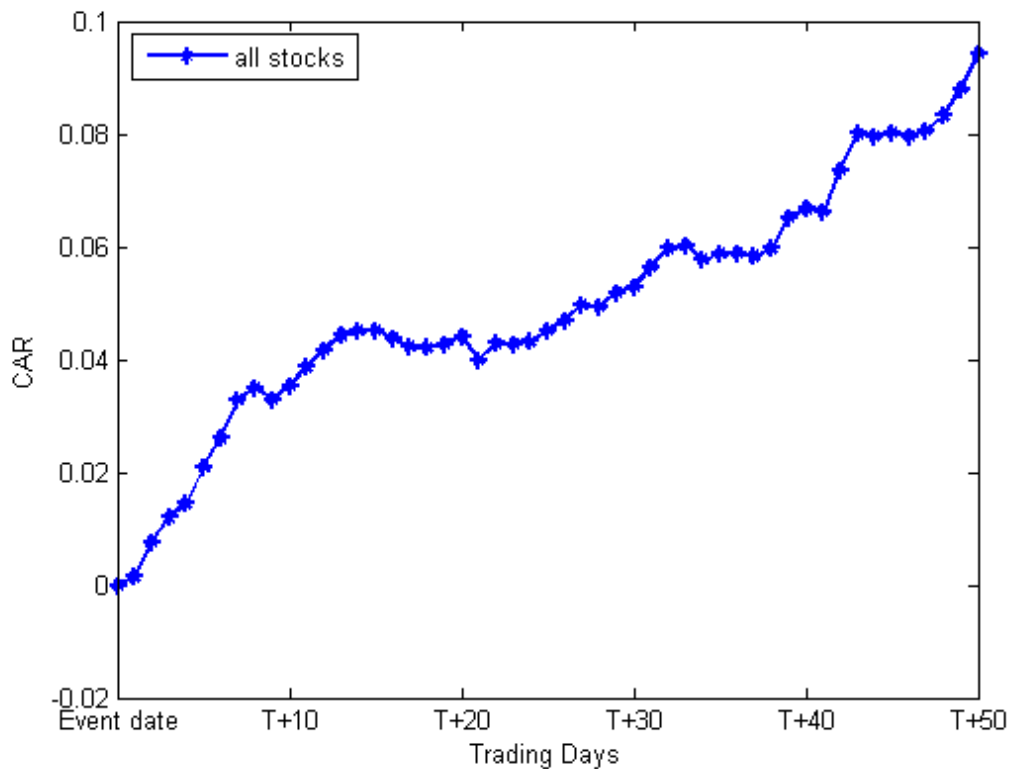
Figure 4.3 Cumulative abnormal return of all stocks

This figure illustrates the price changes in CAR after the familiarity shocks for all the stocks. The two-sample t-test is performed to test the statistical difference from zero for the price changes. The mean CAR is 0.0494, which is significant different from zero at 1% level with t-value= 16.24 and p-value=0.000.

4.4   The Effect of Familiarity on Liquidity

In this section, we further investigate the impact of the familiarity-based investment on the liquidity of the stocks that experience familiarity shocks. The intuition is that familiarity shocks may affect the liquidity through the investment by individual investor's trading behavior. Figure 4.4 illustrates this impact and the two-sample t-test documents a significant decrease in the subsequent 50 trading days. The turnover is decreased from 3.2463 to 2.9047 with the difference of 0.3416 at 1% significant level with t-value=7.7094 and p-value=0.000. Besides, we also perform the multivariate regression controlling for other factors that may affect the liquidity, including firm age, market returns, market capitalization, PE ratio. The familiarity shocks is set as a dummy variable and a negative coefficient (-0.3773) is

documented and significant at 1% level (t-value=-5.6499 and p-value=0.000).
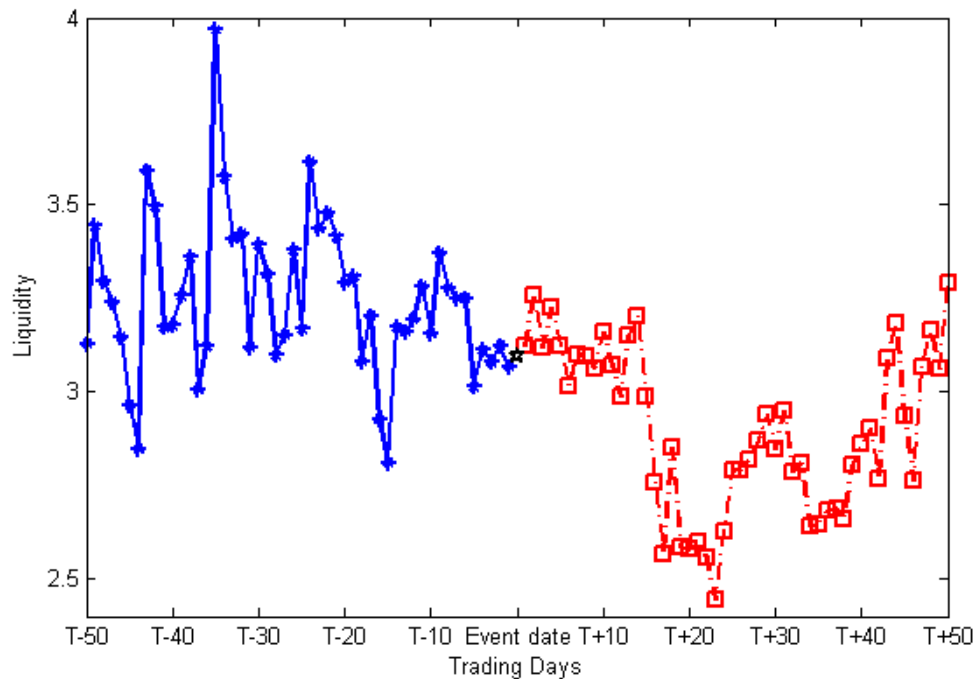
Figure 4.4 The impact of familiarity-based investment on liquidity

This figure illustrates the impact of familiarity-based investment on liquidity. The blue solid line with asterisk represents the turnover during previous 50 trading days. The red dashed line with square represents the turnover during subsequent 50 trading days. The black pentagram represents the turnover on the familiarity shocks events. The two-sample t-test is performed to test the statistical difference between the liquidity changes. The liquidity is decreased from 3.2463 to 2.9047 with the difference of 0.3416 at 1% significant level with t-value=7.7094 and p-value=0.000.

4.5 The Effect of Familiarity on Volatility

Following the similar intuition in previous section, we also examine the impact of familiarity-based investment on volatility. Several measurements of volatility are employed, including the volatility computed by the GARCH model, the exponentially weighted moving average volatility as well as the simple moving average with different moving days. Figure 4.5 illustrate this impact on Garch-volatility and the two-sample t-test documents a significant decrease in the subsequent 50 trading days. The Garch-volatility is decreased from 0.0263 to 0.0252

with the difference of 0.0011 at 1% significant level with t-value=6.7265 and p-value=0.000. Table 4.2 summarizes the volatility changes around the familiarity shocks events with alternative measurements. They all document a significant decrease in the subsequent trading days.
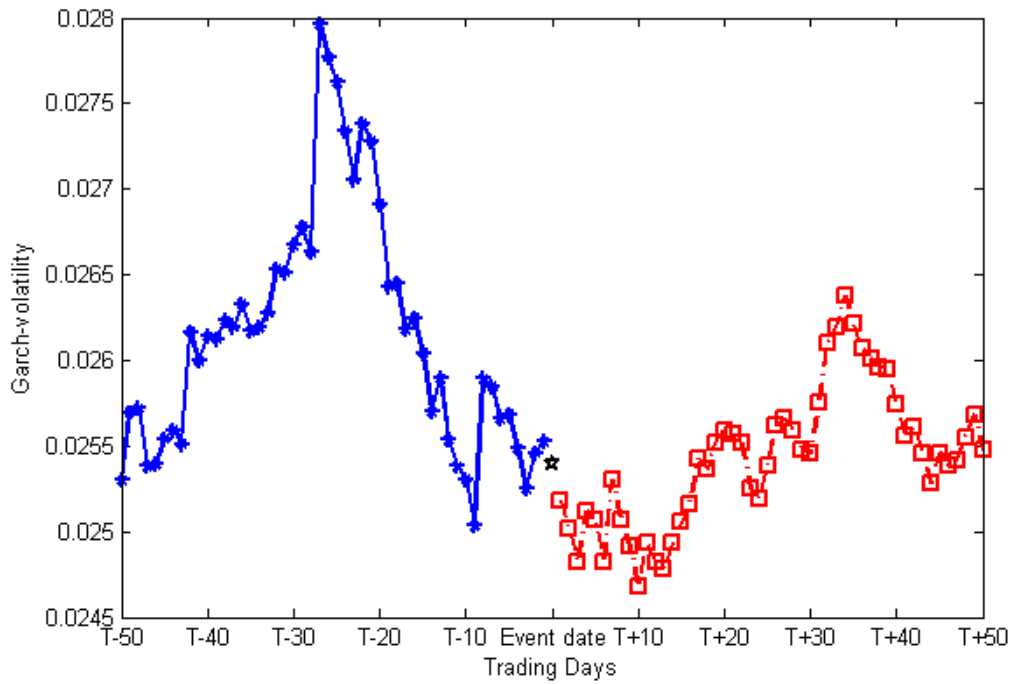
Figure 4.5 The impact of familiarity-based investment on volatility

This figure illustrates the impact of familiarity-based investment on volatility. The blue solid line with asterisk represents the Garch-volatility during previous 50 trading days. The red dashed line with square represents the Garch-volatility during subsequent 50 trading days. The black pentagram represents the Garch-volatility on the familiarity shocks events. The two-sample t-test is performed to test the statistical difference between the Garch-volatility changes. The Garch-volatility is decreased from 0.0263 to 0.0252 with the difference of 0.0011 at 1% significant level with t-value=6.7265 and p-value=0.000.

Table 4.2 Volatility changes around the familiarity shocks events

This table reports the volatility changes around the familiarity shocks events for various measurements. The Garch-volatility denotes the volatility computed with the GARCH model and the Ewma-volatility denotes exponentially weighted moving average volatility. The Sma20-volatility, Sma60-volatility and Sma120-volatility denote the simple moving average with 20, 60 and 120 days, respectively. The two-sample t-test is performed to test the statistical differences around the familiarity shocks. The Diff denotes the differences between the post and pre volatility and the DRate is the decreased rate.

	Measurements of volatility				
	Garch-volatility	Ewma-volatility	Sma20-volatility	Sma60-volatility	Sma120-volatility
Pre	0.0263	0.0250	0.0247	0.0255	0.0259
Post	0.0252	0.0237	0.0236	0.0240	0.0250
Diff	-0.0011	-0.0013	-0.0011	-0.0015	-0.0009
DRate	4.18%	5.2%	4.453%	5.88%	3.47%
t-value	6.7265	6.3923	3.2608	22.0233	26.4338
p-value	0.000	0.000	0.000	0.000	0.000

4.6   Robustness

To ensure the main findings on the information-based familiarity are not biased by the classification of the *HRMF* and the *LRMF* subgroups. We reconstruct the subgroups by choosing the highest (lowest) 30 to 65 familiarity shocks as the *HRMF* and the *LRMF* subgroups. Table 4.3 shows the results for the alternative classifications of the *HRMF* and the *LRMF* subgroups. As is shown, the mean *CAR* of *HRMF* is always larger than that of the *LRMF* subgroup (the value in the row of "Differences" are positive), which is significant at 1% level (the value in the row of "p-value" equal to 0.000). Besides, the *CAR* of *HRMF* and *LRMF* subgroups for all the classifications are significant from zero. Figure 4.6 illustrates the *CAR* of alternative classifications of the *HRMF* and *LRMF* subgroups. For all the classifications, ranging from highest (lowest) 30 to 65 familiarity shocks, the mean *CAR* of *HRMF* is always larger than that of the *LRMF* (value in the black solid line with pentagram are positive). Figure 4.7 further illustrates the t-value of alternative classifications of the *HRMF* and *LRMF* subgroups. The t-value of *HRMF*, *LRMF* and Differences are larger than the t-value denotes the significant

level at 1%. These results naturally uphold the familiarity-based investment in Chinese stock market is driven by information.

Table 4.3 Alternative classifications of the *HRMF* and *LRMF* subgroups

	Number of highest (lowest) familiarity shocks as the *HRMF* and *LRMF* subgroups							
	30	35	40	45	50	55	60	65
CAR of *LRMF*	0.0718 (0.0000)	0.0486 (0.0000)	0.0491 (0.0000)	0.0464 (0.0000)	0.0506 (0.0000)	0.0495 (0.0000)	0.0439 (0.0000)	0.0428 (0.0000)
CAR of *HRMF*	0.0870 (0.0000)	0.0756 (0.0000)	0.0790 (0.0000)	0.0730 (0.0000)	0.0664 (0.0000)	0.0622 (0.0000)	0.0610 (0.0000)	0.0570 (0.0000)
Differences	0.0152	0.0270	0.0299	0.0266	0.0157	0.0127	0.0171	0.0141
t-value	3.9056	8.0859	9.3964	9.7577	7.2877	8.3117	10.8589	10.9924
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note: the p-value equal to 0.000 denotes significant at 1% level.
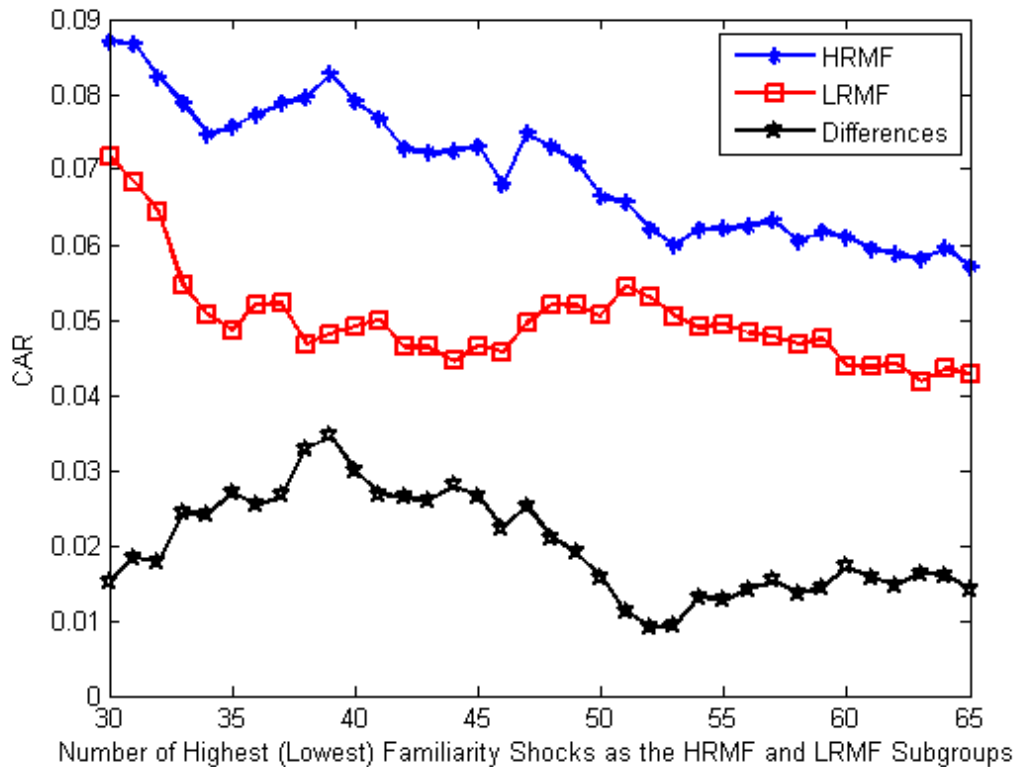
Figure 4.6 The *CAR* of alternative classifications of the *HRMF* and *LRMF* subgroups

The blue solid line with asterisk represents the mean of *CAR* of *HRMF* for all the classifications. The red solid line with square represents the mean of *CAR* of *LRMF* for all the classifications. The black solid line with pentagram represents the differences of *CAR* between *HRMF* and *LRMF* for all the classifications. The positive value in the black solid line with pentagram denotes the *CAR* of *HRMF* is always larger than that of *LRMF*.
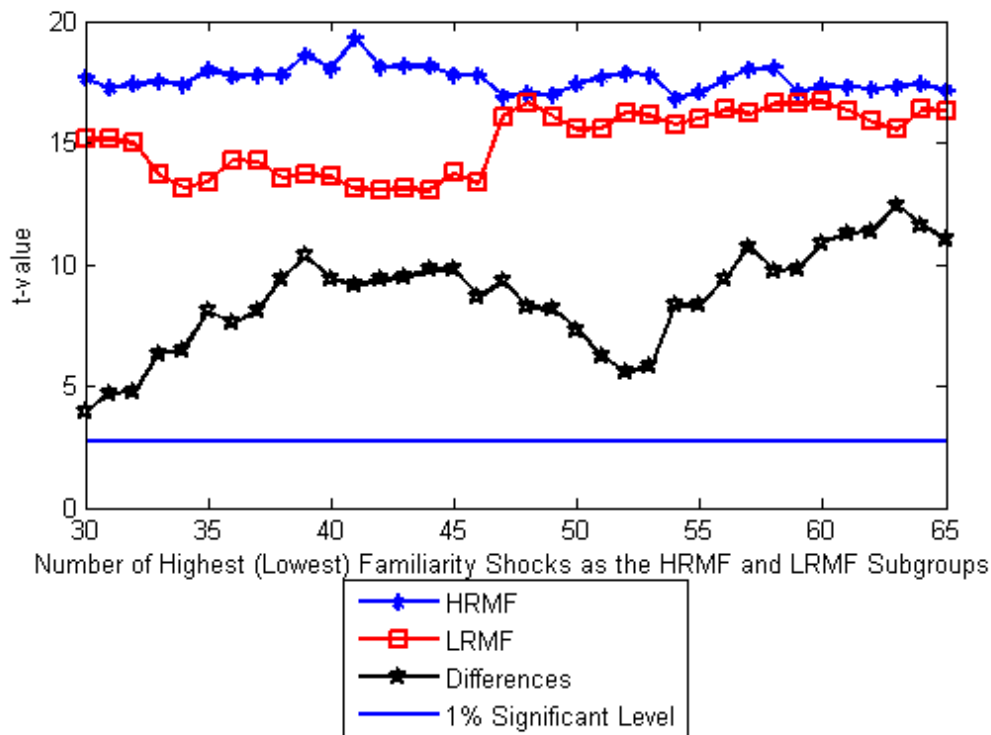
Figure 4.7 The t-value of alternative classifications of the HRMF and LRMF subgroups

The blue solid line with asterisk represents the t-value of   HRMF for all the classifications. The red solid line with square represents the-value of   LRMF for all the classifications. The black solid line with pentagram represents the t-value of the differences of CAR between HRMF and LRMF for all the classifications. All the t-values are larger than value of the blue solid line, which denotes the significant price changes at 1% level.

4.7   Discussions and Conclusions

Admittedly, the familiarity measurement used in the chapter could be expounded as the proxy for investor recognition. However, it is not true in this study and the reasons are in the followings. Firstly, the measurement of the Representative Magnitude of Familiarity (RMF) is defined as the number of participated individual investors divided by the percentage of the individual investors. This measurement is distinct from previous literature relying on the number search frequency of stock names with search engine (Da *et al.*, 2011 and Zhang *et al.*, 2013), the spending of advertising (Grullon *et al.*, 2004 and Lou, 2014) and institutional ownership (Lehavy and Sloan, 2008). As illustrated in the figure 4.1, the RMF is significantly different from the raw number of participated individual investor at 1% level (with

p-value=0.000 and t-value=5.1742). Secondly, we search the all the announcements on the website of Shenzhen stock exchange during the corresponding period and exclude the stocks with earning announcements, price limits and experienced trading halts one week around the "familiarity shocks". This reduces the potential grabbing-events as argued by Barber and Odean (2008) that extreme price changes can lead to individual investor attention. Thirdly, the empirical results show that there is no return reversal in the subsequent 50 trading days, which is different from the empirical findings support the theory of investor recognition in international stock market (Da *et al.*, 2011, Bank *et al.*, 2011 and Zhang *et al.*, 2013). Besides, the two-sample t-test also shows that there are no significant changes in the institutional ownership before and after the familiarity shocks (t-value=-0.0620 and p-value=0.9506). This also confirms the main hypothesis that the short-term price changes are driven by individual investor.

To sum up, we firstly examine the nature of familiarity-based investment in Chinese stock market. Connecting to the information diffusion hypothesis, we find significantly persistent price changes after the familiarity shocks and these changes of *HRMF* subgroup are larger than that of the *LRMF* subgroup. These findings prove that the familiarity-based investment is driven by information. Furthermore, we examine the impact of this familiarity-based investment and find significant decreases in both liquidity and volatility after the familiarity shocks.

5. Online Sentiment Dynamics and Excess Comovement of Stock Returns

The chapter investigates the explanatory power of investor sentiment to daily excess comovement of 90 stocks in Chinese stock market from March 1st 2011 to March 30th 2012. The excess comovement is defined as the significant correlation coefficient between any two pairs of stocks beyond what would be justified by the Fama-French three factors. We find the investor sentiment constructed by the Internet Information can provide explanatory power for the observed excess comovement and outperform other prevailing theories, e.g., information asymmetry, momentum trading, institutional trading and liquidity shocks.

5.1 Introduction

The covariance of assets is the foundation of portfolio and asset pricing theory that determines the expected return. Theoretically, under the CAPM framework, the covariance between two assets can be fully interpreted as the interactions between the price synchronicity with the market for each asset and the market variance. In that sense, the excess comovement between two assets is defined as the covariation beyond what can be justified by the fundamental factors. Recently studies begin to focus on the excess comovement of various categories of assets among different countries (Pindyck and Rotemberg, 1990 and 1993; Deb *et al.*, 1996; Karolyi and Stulz, 1996; Barberis and Shleifer, 2003; Barberis *et al.*, 2005; Robin and Nathan, 2007; Greenwood, 2008 and Kallberg and Pasquariello, 2008). In particular, the excess comovement is alternatively interpreted as the contagion during the financial crisis (Forbes and Rigobon, 2002; Bekaert *et al.*, 2005). Several theoretical models have been proposed to explain the observed excess movement. King and Wadhwani (1990) regard the excess comovement among countries as the consequence of the investors' mistakenly inference about the information from other market. Allen and Gale (2000) claim that the excess comovement comes from the financially fragile induced by a liquidity shock and then spread to others economies. In a multiple

assets model, Kodres and Pritsker (2002) show that the idiosyncratic shock in one country can cause the investors to optimally rebalance their portfolio's exposures to macroeconomics risk. Therefore, countries whose assets are driven by the same macroeconomics factors are intrinsically correlated. Veldkamp (2006) provides an information-based explanation and shows that the access to the low cost of the high-demand information makes investors trade similarly in regardless of the fundamentals. Using a multi-asset trading mechanism, Pasquariello (2007) show that the excess comovement can still be an equilibrium outcome when investors receiving heterogeneous information about the fundamental. Other similar studies include Fleming *et al.* (1998), Kyle and Xiong (2001), Yuan (2005) and Pavlova and Rigobon (2007).

Despite all the empirical findings and theoretical predictions, we further conjecture that the above-mentioned theories are not applicable for the case in Chinese stock market and the investor sentiment maybe the potential cause for the excess comovement. There are mainly three reasons for this conjecture. Firstly, the focus of this chapter is largely different from existing literature on the cross-country evidence with long-term time series data. We attempt to explain the short-term excess comovement with the newly-available high frequent proxy for investor sentiment. Secondly, theories on investor sentiment have proved that the sentiment cannot be cancelled out by limited arbitrageurs and the some psychological biases, e.g., overconfidence and self-attribution, can generate the changes in autocorrelations of stock return and volatility (De Long *et al.*, 1990 and Daniel *et al.*, 1998). In that sense, investor sentiment is intrinsically correlated with the excess comovement. Thirdly, compared to the developed market, the Chinese stock market is relatively closed with low percentage of foreign institutional investors. Therefore, the cross-countries rebalancing activity of risk diversification (Kodres and Pritsker, 2002) and information transmission (King and Wadhwani, 1990) theories are not suitable. Besides, the Chinese stock market is dominated by large percentage of individual investors who subject to strong psychological biases (Feng and Seasholes,

2008 and Lee and Liu, 2011). This unique investor structure and behavior in Chinese stock market should display a distinct picture from other markets.

Following this intuition, we empirically address these issues by combining the capital data with an Internet Information data. More specifically, we use the search frequency of stocks names in Baidu Index as the initial metrics for the investor sentiment. Recent studies have shown that this category of sentiment index constructed by Internet Information is extremely suitable to measure investor sentiment and outperforms other categories of proxies in the following aspect. Firstly, unlike the market-based proxies for investor sentiment (Lee *et al.*, 1991; Neal and Wheatley, 1998; Baker and Wurgler, 2006 and 2007), i.e., trading volume, discounts of closed-end funds, first-day return of the initial public offering, mutual fund flows, are the consequences of mixed economic factor other than investor sentiment. The Internet Information-based proxy is undoubtedly more exogenously and objectively. Secondly, compared with the monthly or quarterly survey-based proxies (Solt and Statman, 1988; Fisher and Statman, 2000; Brown and Cliff, 2004 and 2005 and Lux, 2011), e.g., Investor Intelligence by American Association of Individual Investors, Consumer Confidence by the Conference Board, UBS/Gallup survey and animusX-Investors sentiment, the Internet Information-based proxy possesses advantage in high frequency and revealing the sentiment of hundreds of millions online users.

In this chapter, we first provide empirical evidences on the existing of excess comovement among selected 90 stocks' returns in Chinese market and then we show that the excess comovement can be explained by the Internet Information-based Investor Sentiment Index. The choice of the selected stocks is reconciled by the availability of the Internet Information. The following sections proceeds as follows. Section 5.2 gives a detailed description about the construction of investor sentiment and capital data. Section 5.3 presents econometric measurement of excess comovement. The empirical results on existence of excess sentiment and explanatory

power of investor sentiment are given in section 5.4 and section 5.5, respectively. Section 5.6 is the robustness and section 5.7 concludes.

5.2 Data Description

There are mainly two sources of data. The first refers to the construction of investor sentiment with the online search data acquired from Baidu Index. The second is the capital data, including randomly selected 30 stocks in each of the ChiNext, the SME and the Main Board of Chinese stock market.

5.2.1 Internet Search-based Investor Sentiment Index

In this section, we construct a novel proxy for Chinese investor sentiment based on the search frequency of stock names provided by Baidu Index over the period March 1st 2011 to March 30th 2012. As illustrated in Chapter 3, Baidu is the largest Chinese-language search engine with hundreds of millions online user every day. As one of the most important components, Baidu Index is a keyword analysis tool, which employs the search queries from Baidu to calculate the search frequency of certain keyword. Its function is very similar to the Google Trends and Google Insights, but possesses advantage in providing us with more scientific, objective and authentic data in absolute form (Zhang *et al.*, 2013). Figure 5.1 illustrates the search frequency of the keyword "Spain (the name of the country in Chinese)". The Baidu Index reports the daily search frequency of this keyword and plots them in a chronological order. As for the search identifier for stocks, we simply use the stock name. The reason is that the Chinese names of the listed companies are mainly compounding words that have exclusive meanings. While the stock tickers may be recognized as the account numbers in some cases. The Baidu Index data employed in this chapter is from Zhang *et al.* (2013), but this study differs from them in the following two aspects. Firstly, we use the search frequency to compute a universal proxy for individual investor sentiment, other than use the raw search frequency for each stock. Secondly, we focus on the explanatory power of this constructed proxy for individual investor sentiment on the excess co-movement, while Zhang *et al.*

(2013) address the granger causality and correlations between searching behavior and stock market performance, i.e., the abnormal stock return and trading volume.

Figure 5.1 An illustration of Baidu Index

This figure illustrates the search frequency of the keyword "Spain (the name of the country in Chinese)". It plots daily aggregate search frequency and can be displayed in one week, one month, one quarters, one year and the full time range. The capital alphabets in "Search Frequency" curve correspond to the hottest news relevant to the keyword, which is given in chronological order. However, Baidu Index no longer provides the searching results for stock names and Baidu itself established a mutual fund on 30 October, 2014.

Recent studies have interpreted the search frequency of Google Trends as some reflections of individual investors' sentiment (Joseph *et al.*, 2011; Dzielinski, 2012; Da *et al.*, 2015 and Irresberger *et al.*, 2015). The rationale behind this interpretation can be traced back to the literature in economic psychology consistently support that the agents will react to their inner state of mood with increasing searching behavior (Stigler, 1961; Rothschild, 1974 and Loibl and Hira, 2009). In that sense, we advocate a new proxy for investor sentiment with search frequency of stock names in Baidu Index. In particular, we employ the widely used Principal Component Analysis to construct this sentiment proxy (PCA-sentiment) (Baker and Wurgler, 2006; Billio *et al.*, 2012 and Chen *et al.*, 2014). Firstly, we acquire the search frequency of stock name for 90 stocks from the Baidu Index over the period 1 March 2011 to 30 March 2012. Secondly, we calculate the eigenvalue and eigenvector of

the covariance matrix, i.e., search frequencies of 90 stocks and 267 trading days for each stock during the period. Thirdly, we construct the investor sentiment index as the linear combination of the 90 variables by employing the eigenvector associated with the largest eigenvalue as the corresponding weight. Figure 5.2 illustrates the PCA-sentiment index and the average value of the stock returns. The correlation analysis documents an insignificant correlation coefficient 0.0580 (p-value=0.3454). This insignificant correlation coefficient thus suggests that the PCA-sentiment is not an appropriate systematic factor to explain the portfolio returns (Baker and Wurgler, 2006) and can be employed to address the prices changes caused by idiosyncratic factors (Campbell *et al.*, 2001; Xu and Malkiel, 2003; Fu, 2009 and Zhang *et al.*, 2016).
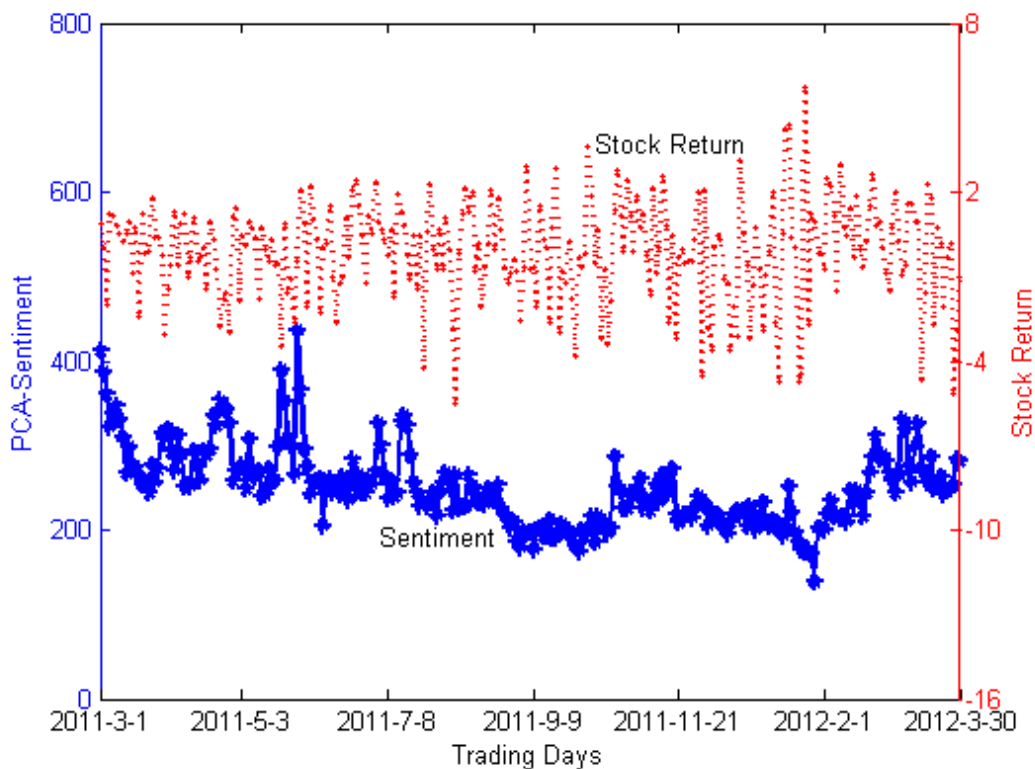
Figure 5.2 PCA-sentiment index and stock returns

This figure illustrates the PCA-sentiment index and the mean of the 90 stock returns. The blue solid line with asterisk represents the PCA-sentiment and the red dotted line represents the stock returns. The correlation analysis documents an insignificant correlation coefficient between the sentiment index and corresponding stock returns with the p-value=0.3454. The maximum, minimum and median value of the sentiment index is 436.21, 138.78 and 249.63, respectively. Please note that this PCA-sentiment is not the sentiment employed in section 5.5.

65

5.2.2 Capital Data

Our sample focuses on the ChiNext, the Small and Medium Enterprise (SME) Board and the Main Board of Chinese stock market. We randomly select 30 stocks from each board, covering the period from March 1st 2011 to March 30th 2012. Totally, there are 267 trading days. In particular, the ChiNext is a NASDAQ-style board, which offers an alternative capital platform for growing venture enterprises and the enterprises engaged in independent innovation. Both the ChiNext and the SME Board are important component of establishing a multi-tier capital market. The capital market data come from RESSET Financial Research Database, including the daily stock returns after dividend reinvestment, the trading volume, the PE ratio, the turnover, the Garch-volatility, the market returns, the Fama-French three factors and the quarterly number and percentage of institutional investors. The Shanghai Interbank Offered Rate (SHIBOR) is from the website (http://www.shibor.org/). Table 5.1 reports the descriptive statistics for all the 90 stocks in the sample. The maximum, minimum and median value of the trading volume, market capitalization, the turnover, PE ratio and the volatility show that our sample containing firms with a wide range of characteristics. Therefore, this sample can be considered as a parsimonious representation of Chinese stock market.

Table 5.1 Summary statistics of the sample stocks

This table reports the descriptive statistics of the 90 stocks in the sample. The capitalization and turnover are calculated on the tradable value. The Garch-volatility is computed with the GARCH model (Bollerslev, 1986). All these daily data are directly download from the RESSET Financial Research Database. All the numbers in the table are the average value of the daily variables over the sample period.

Variables	N	Mean	Median	Std.	Max	Min
Individual returns	23452	-0.0014	-0.0004	0.0280	0.1012	-0.1004
Trading volume	23452	$5.56*10^6$	$2.04*10^6$	$7.66*10^6$	$1.49*10^8$	33700
Capitalization	23452	$7.79*10^7$	$3.04*10^7$	$2.24*10^8$	$5.83*10^9$	$6.46*10^5$
PE ratio	23452	117.74	44.650	329.99	3119.3	-667.08
Turnover	23452	2.6196	1.5456	3.6437	58.943	0.0398
Garch-volatility	23452	0.0279	0.0269	0.0058	0.0649	0.0152

5.3   Econometric Measurement of Excess Comovement

In this chapter, we define the excess comovement among different stocks as covariations beyond what can be justified by common factors. Following Kallberg and Pasquariello (2008), we also employ the Fama-French three factors model, considering the market risk, book-market risk and small firm risk, is chosen as the common factors in the regression. The Fama-French factors have proved to be good proxies for risk factors in Chinese stock market (Lin *et al.*, 2012) and therefore, we construct the following regression model.

$$r_{it} - rf_t = \alpha_i + \beta_1 (r_{mt} - rf_t) + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_{it}$$
(5-1)

where r_{it} is the daily stock return of stock i on date t, r_{mt} is the CSI 300 stock index return on date t, rf_t is the three-months Shanghai Interbank Offered Rate (SHIBOR) on day t, SMB_t (small minus big) is the return on the portfolio for size factor on date t, HML_t (high minus low) is the return on the portfolio for the book-to-market factor on date t, ε_{it} is the regression error term.

In this econometrics setting, the comovement between any pairs of stock returns r_{kt} and r_{pt} is considered excess if after controlling for the Fama-French factors. The regression residuals ε_{kt} and ε_{pt} of stocks k and p are used to compute the excess correlation coefficient for any pairs of stocks $k \neq p$. To evaluate the time-varying characteristic of the excess comovement among the 90 stocks, we employ the rolling windows estimation to calculate the excess correlation coefficient. The advantage of this method is that it depicts the time-varying characteristic of excess comovement. Admittedly, the major drawback of the rolling windows estimation is that the choice of the window length may be arbitrary.

$$ecc_{kp,tm} = \frac{cov(\varepsilon_{k,tm}, \varepsilon_{p,tm})}{\sqrt{var(\varepsilon_{k,tm})var(\varepsilon_{p,tm})}}$$
(5-2)

where ecc_{kpm} is the excess comovement for stock k and p with the window length of m trading days at time t. The excess comovement among the selected stocks is then defined as the square unconditional, statistically significant coefficients. Each observation in t of the time series is the average value of the number of significant correlation coefficient of $[(90*89)/2]$. In particular, we employ the window length of 30 trading days and it gives 238 observations (number of t) in the time series. This primary choice of the window length is consistent with the focus on short-term excess comovement. However, we will alleviate this issue with alternative choices of window length and provide the results in 5.6.

In a similar way, we also construct a time-varying PCA-sentiment with the consideration of the comovement of investor sentiment towards each stock.

$$PCAsentiment_{t,m} = PCA_{t,m} * Comovement_{t,m}$$

(5-3)

where $PCAsentiment_{t,m}$ is the time-varying PCA-sentiment with the window length of m trading days at time t. $PCA_{t,m}$ is the loadings computed by PCA along the window length at time t. $Comovement_{t,m}$ is the comovement of Baidu Index with the window length of m trading days at time t. Consistent with $ecc_{kp,tm}$, we also employ the window length of 30 trading days and it gives 238 observations (number of t) in the time series. $PCAsentiment_{t,m}$ is the investor sentiment employed for the regression analysis in section 5.5.

5.4   Empirical Results

Figure 5.3 illustrates the excess correlation coefficient of the ChiNext, SME Board and Main Board and the average value is 0.2299, 0.2125 and 0.2052, respectively. The two-sample t-test shows that the excess correlation coefficient of the ChiNext is significant larger than that of the SME Board (t-value=9.1067 and p-value=0.0000) and Main Board (t-value=13.1889 and p-value=0.0000) at 1% level. Besides, the

excess correlation coefficient of the SME Board is also significant larger than that of the Main Board at 1% level (the t-value=5.1207 and p-value=0.0000). From the perspective of informational efficiency, the less excess correlation coefficient suggests a more efficient market, i.e., the Main Board is the most efficient market compared with the ChiNext and the SME Board. These findings are consistent with (Zhang *et al.*, 2013) as well as intuitively reasonable since the ChiNext and the SME Board are considered as more speculative markets by conventional concept. And likewise, the explanatory power of the Fama-French three factors of the Main Board is strongest and the ChiNext is less likely to be driven by fundamentals.
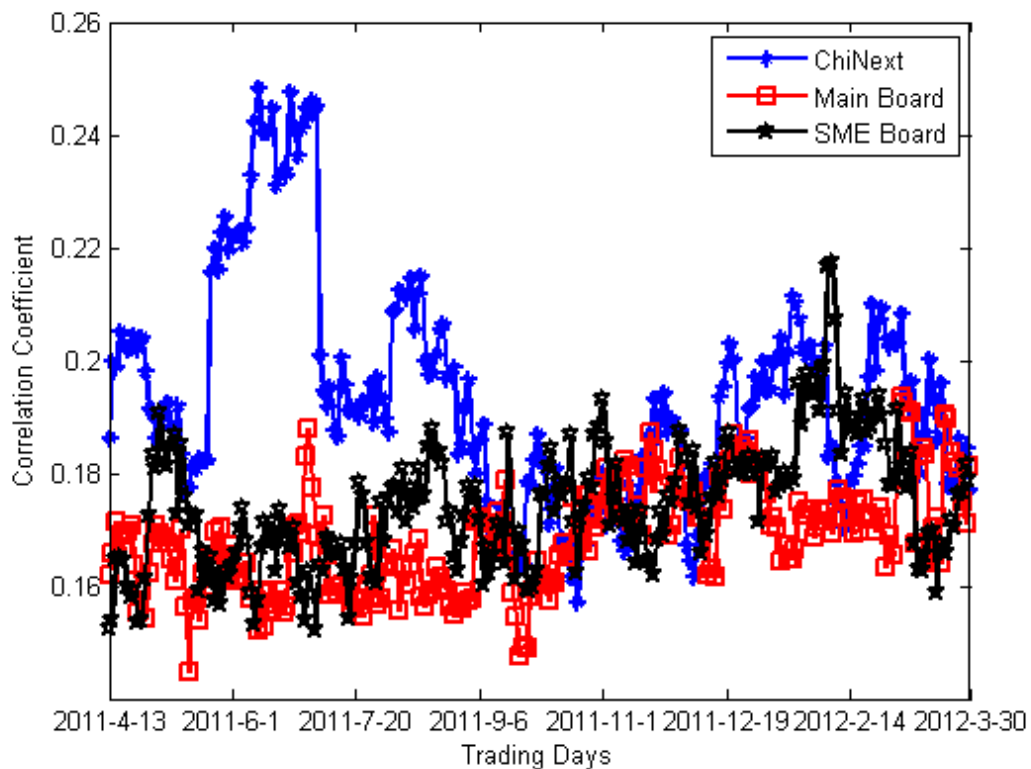
Figure 5.3 Excess comovements of the ChiNext, SME Board and Main Board

This figure illustrates the excess correlation coefficient of the ChiNext, SME Board and Main Board and the average value is 0.1937, 0.1741 and 0.1685, respectively. The two-sample t-test shows that the correlation coefficient of the ChiNext is significant larger than that of the SME Board (t-value=13.34 and p-value=0.0000) and Main Board (t-value=17.88 and p-value=0.0000) at 1% level. Besides, the correlation coefficient of the SME Board is also significant larger than that of the Main Board at 1% level with the t-value=5.9374 and p-value=0.0000.

Figure 5.4 illustrates excess comovement (the red line) and the comovement of the stock returns (the blue line). The difference between the comovement of the stocks' returns and the excess comovement is the comovement explained by the Fama-French three factors, i.e., the market factor, book-market factor and small firm factor. The mean value of the excess comovement is 0.2104 and it is highly significant different from zero (t-value=349.09 and p-value=0.0000). These findings suggest that there exists excess comovement in Chinese stock market. Besides, in order to rule out the spurious correlation coefficient from the Gaussian distribution, we calculate the percentage of the significant correlation coefficient of the matrix of the residuals and the random matrix, respectively. Figure 5.5 illustrates the percentage of the significant correlation coefficient. The blue line is the percentage of the significant correlation coefficient of the matrix of the residuals and its average value is 18.64%. The red line is the percentage of the significant correlation coefficient of the random matrix and its average value is 10.21%. This significant difference (t-value=35.47 and p-value=0.0000) confirms that the observed excess comovement in figure 5.4 is not driven by the Gaussian distribution.
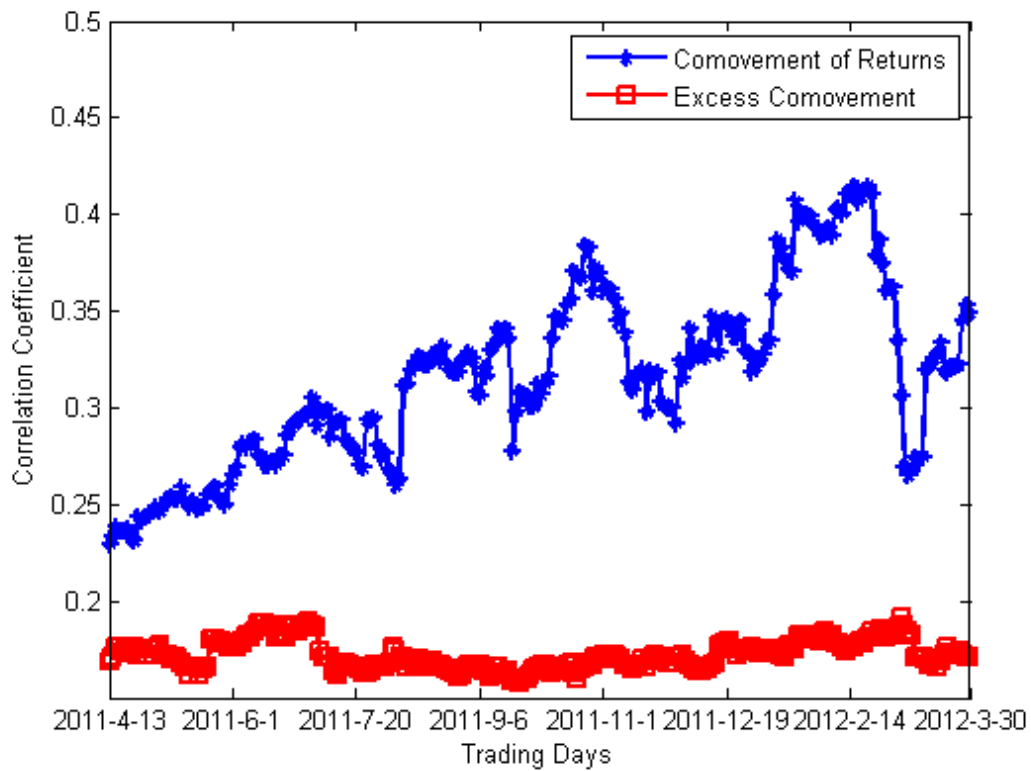
Figure 5.4 Excess comovement and the comovement of returns

This figure illustrates the excess comovement (the red line) and the comovement of the stock returns (the blue line). The difference between the comovement of the stocks returns and the excess comovement is the comovement explained by the Fama-French three factors. The mean value of the excess comovement is 0.1722 and it is highly significant different from zero (t-value=358.94 and p-value=0.0000).
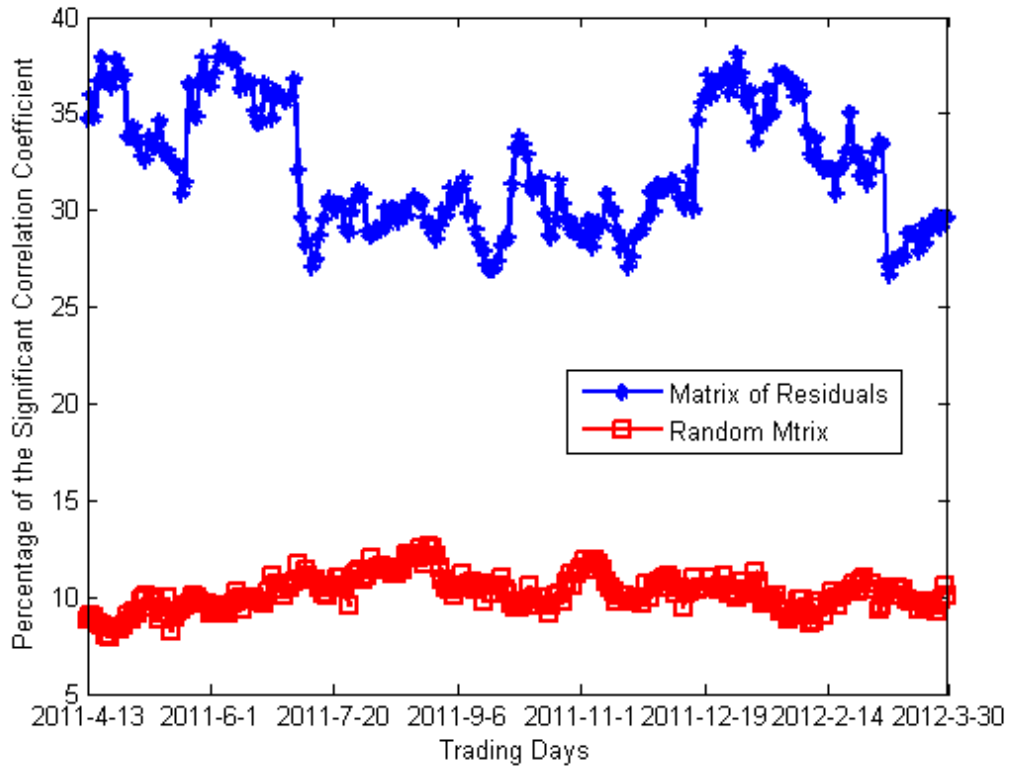
Figure 5.5 Percentage of the significant correlation coefficient

This figure illustrates the percentage of the significant correlation coefficient. The blue line is the percentage of the significant correlation coefficient of the matrix of the residuals and its average value is 32.11%. The red line is the percentage of the significant correlation coefficient of the random matrix and its average value is 10.21%. This difference is significant at 1% level (p-value=0.0000). The simulation of the random matrix is performed with MATLAB R2011b.

5.5 Explaining Excess Co-movement

The empirical results presented in section 5.4 indicate that a significant percentage of return comovement in Chinese stock market cannot be explained by the Fama-French factors as well as there exists some variations in different boards. We further explore the explanatory power of the PCA-sentiment and other prevailing explanations, i.e., the information asymmetry theory, the momentum trading theory, the institutional trading theory and the liquidity shocks theory. All the empirical results in this section are obtained from the following OLS regression model:

$$ecc_t = \alpha + \beta \times Factor_t + \varepsilon_t \tag{5-4}$$

where ecc_t is the observed excess comovement; α and β are the regression

coefficients; ε_t is the residual; $Factor_t$ are the $PCAsentiment$ and other proxies for the corresponding theories, respectively.

5.5.1 Information Asymmetry

As illustrated in section 5.1, the information asymmetry has long been considered as a source of the excess comovement (King and Wadhwani, 1990 and Kodres and Pritsker, 2002). The rationale of this explanation is that investor are unable to identify the systematic risk and the idiosyncratic risk and therefore leads to assets move together in spite of the fundamental value of each asset. We employ the market return as the common proxy for the information asymmetry of the market (Zhang, 2006 and Kallberg and Pasquariello, 2008). In particular, we use the volatility of SSE Composite Index (IndexV$_{SH}$), volatility of Shenzhen Component Index (IndexV$_{SZ}$) and volatility of CSI 300 Index (IndexV$_{HS300}$) as the proxies for information asymmetry. For the calculation of volatility, we employ the GARCH volatility, the exponentially weighted moving average volatility and the simple moving average with 20 and 60 days, respectively. Table 5.2 reports the explanatory power of the PCA-sentiment and the most satisfied proxy for information asymmetry, i.e., the GARCH volatility. The empirical results show that the information asymmetry is not a satisfying theory in explaining the excess comovement in Chinese stock market. However, the PCA-sentiment can explain 14.97% of the excess comovement with positive correlation.

Table 5.2 Comparisons with the information asymmetry explanation

This table reports the univariate regression for the coefficients of the explanatory power of PCA-sentiment and proxies for information asymmetry. We use the volatility of Shanghai Composite Index (IndexV$_{SH}$), volatility of Shenzhen Component Index (IndexV$_{SZ}$) and volatility of CSI 300 Index (IndexV$_{HS300}$) as proxies for the information asymmetry. For the calculation of volatility, we employ the GARCH volatility, the exponentially weighted moving average volatility and the simple moving average with 20 and 60 days, respectively. The information asymmetry proxies are from RESSET Financial Research Database. Value in the parenthesis is the p-value and F denotes the F-value of the regression model.

Variables	Regression Models			
Constant	0.1510***	0.1757***	0.1679***	0.1720***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
PCA-Sentiment	0.4111***			
	(0.0000)			
IndexG-V$_{SH}$		-0.2381*		
		(0.0762)		
IndexG-V$_{SZ}$			0.2551*	
			(0.0911)	
IndexG-V$_{HS300}$				0.0120
				(0.9619)
Observations	238	238	238	238
Adjusted R^2	14.97%	0.91%	0.79%	0.42%
	(F=42.71)	(F=3,1711)	(F=2.8783)	(F=0.0023)

Notes: *** indicates correlation coefficient significant at 1% level.

5.5.2 Momentum Trading

Momentum trading (Jegadeesh and Titman, 1993 and Carhart, 1997) is another potential explanation for the observed excess comovement. We employ the 10% profitable and unprofitable stocks of n-months momentum factor weighted by market capitalization (nM$_{10}$) and 30% profitable and unprofitable stocks of n-months momentum factor weighted by market capitalization (nM$_{30}$) as the proxies for the momentum factors. The holding period (n) equals to 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 18 and 24 months, respectively. Table 5.3 reports the explanatory power of the PCA-sentiment and the most satisfied proxy for the momentum trading, 6M$_{10}$ and 6M$_{30}$. The empirical results show that the momentum trading can only explain less than 5% the excess comovement in Chinese stock market. However, the

PCA-sentiment can explain 14.97% of the excess comovement with positive correlation.

Table 5.3 Comparisons with the momentum trading explanation

This table reports the univariate regression for the coefficients of the explanatory power of PCA-sentiment and proxies for momentum trading. We use the 10% profitable and unprofitable stocks of n-months momentum factor weighted by market capitalization (nM_{10}) and 30% profitable and unprofitable stocks of n-months momentum factor weighted by market capitalization (nM_{30}) as proxies for the momentum trading. The momentum factors are from RESSET Financial Research Database. Value in the parenthesis is the p-value and F denotes the F-value of the regression model.

Variables	Regression Models		
Constant	0.1510^{***}	0.1717^{***}	0.1716^{***}
	(0.0000)	(0.0000)	(0.0000)
PCA-Sentiment	0.4111^{***}		
	(0.0000)		
$6M_{10}$		-0.0592^{***}	
		(0.0003)	
$6M_{30}$			-0.0690^{***}
			(0.0005)
Observations	238	238	238
Adjusted R^2	14.97%	4.94%	4.53%
	(F=42.71)	(F=13.30)	(F=12.25)

Notes: *** indicates correlation coefficient significant at 1% level.

5.5.3 Institutional Trading

Anton and Polk (2014) show that shared ownership in mutual fund can explain the observed excess comovement. In that sense, we construct the number of trade (NUMofII) and the percentage of trade (PERCofII) by institutional investors as proxies for the institutional trading. Table 5.4 reports the explanatory power of the PCA-sentiment compared with the institutional trading theory. The empirical results show that the institutional trading is not a satisfying theory in explaining the excess comovement in Chinese stock market. However, the PCA-sentiment can explain 14.97% of the excess comovement with positive correlation.

Table 5.4 Comparisons with the institutional trading explanation

This table reports the univariate regression for the coefficients of the explanatory power of PCA-sentiment and proxies for institutional trading. We use the number of trade (NUMofII) and the percentage of trade (PERCofII) by institutional investors as proxies for the institutional trading. These factors are from RESSET Financial Research Database. Value in the parenthesis is the p-value and F denotes the F-value of the regression model.

Variables	Regression Models		
Constant	0.1510^{***}	0.1735^{***}	0.1724^{***}
	(0.0000)	(0.0000)	(0.0000)
PCA-Sentiment	0.4111^{***}		
	(0.0000)		
NUMofII		0.0000	
		(0.3448)	
PERCofII			-0.0136
			(0.9034)
Observations	238	238	238
Adjusted R^2	14.97%	0.00%	0.42%
	(F=42.71)	(F=0.8967)	(F=0.0148)

Notes: *** indicates correlation coefficient significant at 1% level.

5.5.4 Liquidity Shocks

Excess comovement has long been correlated with the changes of interest rate (Shiller, 1989 and Calvo, 1999). In their analysis, the excess comovement results from the trading activity of investor's liquidate behavior. Therefore, we employ the overnight SHIBOR, the 1-week SHIBOR and 1-month SHIBOR as the proxies for liquidity shocks. Table 5.5 reports the explanatory power of the PCA-sentiment compared with the liquidity shocks theory. It shows that the liquidity shocks can explain less than 9% the excess comovement. However, the PCA-sentiment can explain 14.97% of the excess comovement.

Table 5.5 Comparisons with the liquidity shocks explanation

This table reports the univariate regression for the coefficients of the explanatory power of PCA-sentiment and proxies for liquidity shocks. We use the overnight SHIBOR, 1-week SHIBOR and 1-month SHIBOR as proxies for the liquidity shocks. These factors are from RESSET Financial Research Database. Value in the parenthesis is the p-value and F denotes the F-value of the regression model.

Variables	Regression Models			
Constant	0.1510^{***}	0.1670^{***}	0.1643^{***}	0.1628^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
PCA-Sentiment	0.4111^{***}			
	(0.0000)			
$SHIBOR_{O/N}$		0.0015^{***}		
		(0.0003)		
$SHIBOR_{1w}$			0.0019^{***}	
			(0.0000)	
$SHIBOR_{1m}$				0.0018^{***}
				(0.0002)
Observations	238	238	238	238
Adjusted R^2	14.97%	4.83%	8.9%	5.18%
	(F=42.71)	(F=13.01)	(F=24.16)	(F=13.94)

Notes: *** indicates correlation coefficient significant at 1% level.

5.6 Robustness

5.6.1 Alternative Measurement of Investor Sentiment

As a robustness check, we construct an alternative investor sentiment proxy as the capitalization-weighted Internet Information-based investor sentiment index (Cap-sentiment). Figure 5.6 illustrates this Cap-sentiment and stock returns and the correlation analysis documents an insignificant correlation coefficient between them. We further employ this Cap-sentiment to explain the excess Comovement with model (5-4). The empirical results shows that it can explain 10.88% (t-value=5.4724 and p-value=0.0000) of the excess comovement.
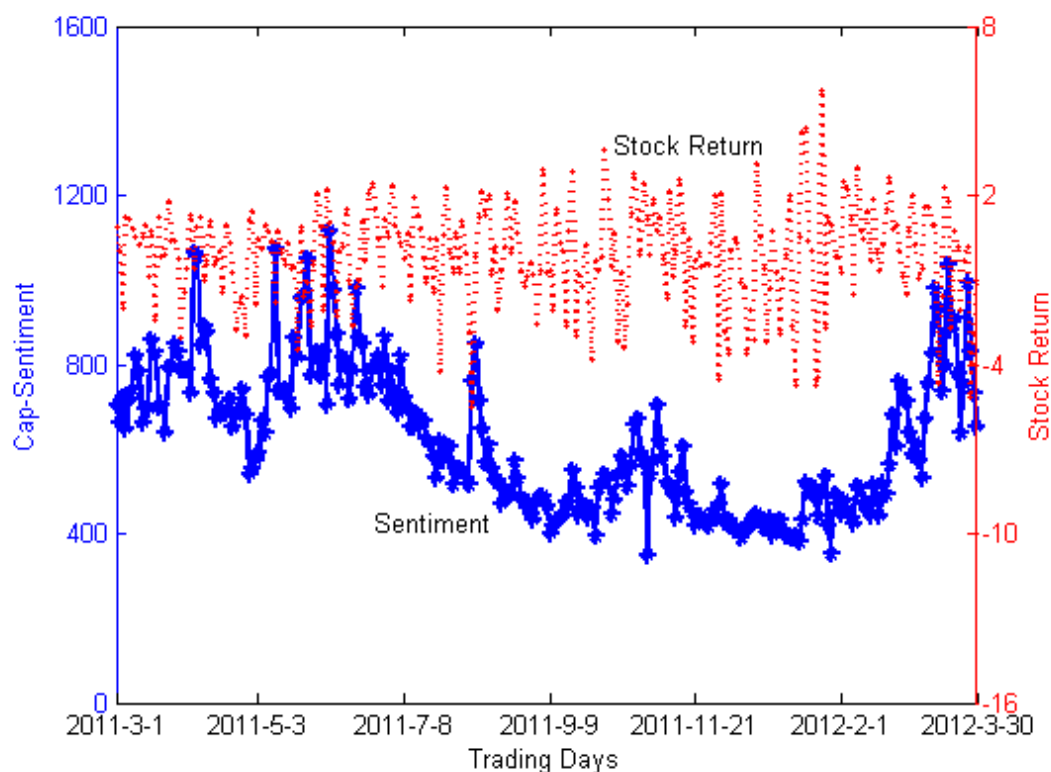
Figure 5.6 Cap-sentiment index and stock returns

This figure illustrates the Cap-sentiment index and the mean of the 90 stock returns. The blue solid line with asterisk represents the Cap-sentiment and the red dotted line represents the stock returns. The correlation analysis documents an insignificant correlation coefficient between the sentiment index and corresponding stock returns with the p-value=0.5997. The maximum, minimum and median value of the sentiment index is 1116.2, 349.9 and 587.0, respectively.

5.6.2   Alternative Choices of Window Length

To alleviate the issue of the arbitrary choices of window length, we further provide the explanatory power of the investor sentiment with alternative choices of window lengths, i.e., 35 trading days, 40 trading days, 45 trading days and 50 trading days. Meanwhile, due to the focus of this chapter is on the short-term excess comovement, the window length is confined to no larger than 50 trading days. Table 5.6 reports the explanatory power of investor sentiment on excess comovement with various window lengths. As is reported, the excess comovement decreases with the increase in the window lengths, but the explanatory power increases with the increase in the window lengths.

Table 5.6 Alternative choices of window length

This table reports the explanatory power of investor sentiment on excess comovement. Value in the rows of "PCA-sentiment" and "Cap-sentiment" are the explanatory powers for the corresponding excess comovement, which are obtained from the regression model (5-4). "Pct. of significant" denotes the percentage of significant correlation coefficient.

Variables	Window length			
	35 trading days	40 trading days	45 trading days	50 trading days
No. of observation	233	228	223	218
Excess comovement	0.1503	0.1336	0.1204	0.1095
PCA-sentiment	20.85%	21.16%	25.52%	29.81%
Cap-sentiment	15.17%	21.94%	26.03%	30.11%
Pct. of significant	32.32%	32.71%	33.15%	33.71%

5.7   Conclusions Remarks

In this chapter, we investigate one of the most fundamental aspects in asset pricing, the excess comovement of stock prices. The excess comovement is defined as the comovement beyond the degree which can be justified by common factors. In particular, we consider the Fama-French three factors as the common factors and calculate the square unconditional, statistically significant correlation of the residuals as the excess comovement. The empirical results show a significant excess comovement and there exist some variations in different Boards. Furthermore, we construct two proxies for investor sentiment with the Internet Information and prove that the investor sentiment is a satisfying explanation for the excess comovement.

To sum up, the exogenous investor sentiment proxy revealing hundreds of millions online users' state of mind shows potentials in completing our understanding of the stock market. More work can be performed to investigate the nature of excess comovement, i.e., information diffusion, private information or psychological bias. The Internet Information can undoubtedly provide some insights in measuring information flow as well as reflecting investors' state of mind.

6. Euphoria: Daily Happiness and Stock Returns

Euphoria is a psychological terminology, which is colloquially used to define the state of intense happiness. Given the unique daily happiness index extracted from Twitter, we empirically examine the relations between the daily happiness sentiment extracted from Twitter and the stock market performance in 11 international stock markets. By partitioning this happiness sentiment into quintiles from the least to the happiest days, we find that the happiness sentiment can provide additional explanatory power for index return in the most-happiness subgroup as well as the index return and the range-based volatility of the most-happiness subgroup are larger than those of other subgroup

6.1. Introduction

Traditional financial theory is based on the assumptions of investors' rationality and homogeneity that leaves no role for the irrational behavior in asset pricing. However, the behavioral economics has shifted the academic focus to investigate the relations between investor sentiment and asset prices. Meanwhile, market participants play an active role in generating "fashion" in stock market, which leads to investor's maniac behavior. All the facts recognize the role of sentiment in explaining and predicting the stock returns.

The existence of noise traders and psychological biases are viewed as the assumptions by most theoretical models on investigating the impact of investor sentiment on stock prices. De Long *et al.* (1990) firstly observe that irrational noise traders could not be offset by limited arbitrageurs; rather, with erroneous beliefs and diverse sentiment, they could actually affect the stock prices and earn higher expected returns. Barberis *et al.* (1998) develop a parsimonious model of investor sentiment to partially explain the underreaction and overreaction. Daniel *et al.* (1998) propose a model populated with overconfidence and self-attribution bias and show

that both of them can cause the changes in autocorrelations of returns and excess volatility. Hong *et al.* (1999) also develop a model with boundedly rational investors (news watchers and momentum traders) and show their explanatory power for the short-run continuation and long-run reversal of the prices. All these theoretical predictions provide a general framework for analyzing and interpreting the empirical results.

In order to investigate the implications of the above-mentioned theoretical predictions, scholars endeavor to construct various proxies for investor sentiment in their empirical work. In general, there are mainly three categories of proxies to measure investor sentiment according to the source of sentiment from which the proxy is extracted. The first category refers to the stock market-based proxies. This category includes the fluctuations in discounts of closed-end funds (Lee *et al.*, 1991), the net mutual fund redemptions (Neal and Wheatley, 1998), bid-ask spreads and turnover (Baker and. Stein, 2004), the combinations of the closed-end fund discount, NYSE share turnover, the number and average first-day returns on IPOs, the equity share in new issues and the dividend premium (Baker and Wurgler, 2006 and 2007) as well as the portfolio allocations to equity versus cash and fixed-income securities (Edelen *et al.*, 2010). As for the second category, empiricists use the survey-based proxies such as the Investor Intelligence by American Association of Individual Investors (Solt and Statman, 1988; Fisher and Statman, 2000; Lee *et al.*, 2002; Brown and Cliff, 2004 and 2005), Consumer Confidence by the Conference Board (CBIND) and Consumer Confidence by University of Michigan Survey Research Center (Lemmon and Portniaguina, 2006), UBS/Gallup survey (Qiu and Welch, 2006) and animusX-Investors sentiment (Lux, 2011). The third category is the proxies extracted from the news and social media content. This category includes sentiment extracted from the Raging Bull (Tumarkin and Whitelaw, 2001), Yahoo! Finance (Antweiler and Frank, 2004; Das and Chen, 2007 and Kim and Kim, 2014), Wall Street Journal (Tetlock, 2007), Twitter (Bollen *et al.*, 2011), Google Insights (Joseph *et al.*, 2011), Facebook (Siganos *et al.*, 2014) and Google Trends (Da *et al.*,

2015). These studies usually construct a time series proxy for investor sentiment and examine the contemporaneous and lead-lag relations between the constructed sentiment proxy and market variables (short-term returns, abnormal trading volume and excess volatility) as well as give mixed findings on the predictability and correlation coefficient.

In this chapter, we also attempt to examine the relations between the investor sentiment released by social media and the stock market activities. Naturally, it is also connected to plenty of studies on the dynamics between online behaviors and stock market movements. The main contribution of this chapter compared with others is to exploit a novel sentiment proxy, i.e., the daily happiness index extracted from Twitter, across 11 international stock markets, perform the contemporaneous correlation regression model with the consideration of trading volume as well as provide asymmetric impact of happiness on index return and range-based volatility with cross-sectional analysis.

The remainder of this chapter is organized as follows. Section 6.2 describes the daily happiness index and capital data. Section 6.3 6.4 and 6.5 performs the contemporaneous correlation, the Granger causality and the cross-sectional analysis between the daily happiness and major stock indexes. Section 6.6 sets forth the robustness test and section 6.7 concludes.

6.2. Data Description

6.2.1. Daily Happiness Index

We obtain raw daily happiness from the website (http://hedonometer.org/index.html). The raw daily happiness scores are derived from the natural language processing technique based on a random sampling of about 10% (50 million) of all messages posted in Twitter. To quantify the happiness of the language, the Amazon's Mechanical Turk service is used to score the level of happiness of selected words appeared on Twitter, e.g., laughter, joy, successful, winning, excellent, rainbow. To

avoid the spurious happiness events like "Thanksgiving" and "Christmas Eve" and the potential seasonality (the happiness rises during the end of the week), we regress the raw daily happiness on weekday dummies and holiday dummies and keep the residuals as the daily happiness index employed in this paper. The deseasonalized time series allows us to analyze the trend of the daily happiness independently of the predictable seasonal patterns and reveal the intrinsically changes in daily happiness. Table 6.1 reports the statistical property of the raw daily happiness and the daily happiness index employed in this paper. Figure 6.1 illustrates the raw daily happiness and figure 6.2 illustrates the daily happiness index, which spans from September 10th, 2008 to May 27th, 2015. Totally, there are 2451 and 1751 calendar days, respectively. In particular, both KPSS test and Augmented Dickey-Fuller test confirm the stationary property of the daily happiness index. For the empirical analysis in the following sections, the daily happiness index corresponds to the trading days of each stock market.

Table 6.1 Statistical property of raw daily happiness and daily happiness index

This table reports the statistical property of daily happiness. The raw daily happiness is the original index downloaded from the website (http://hedonometer.org/index.html) and the daily happiness index is residual of the regression model of raw daily happiness on weekday dummies and holiday dummies.

Index	N	Mean	Median	Std	Kurtosis	Max	Min
Raw daily happiness	1751	6.0092	6.0050	0.0428	6.0469	6.3560	5.8820
Daily happiness	1751	0.0000	-0.0038	0.0338	0.4023	0.1946	-0.1317
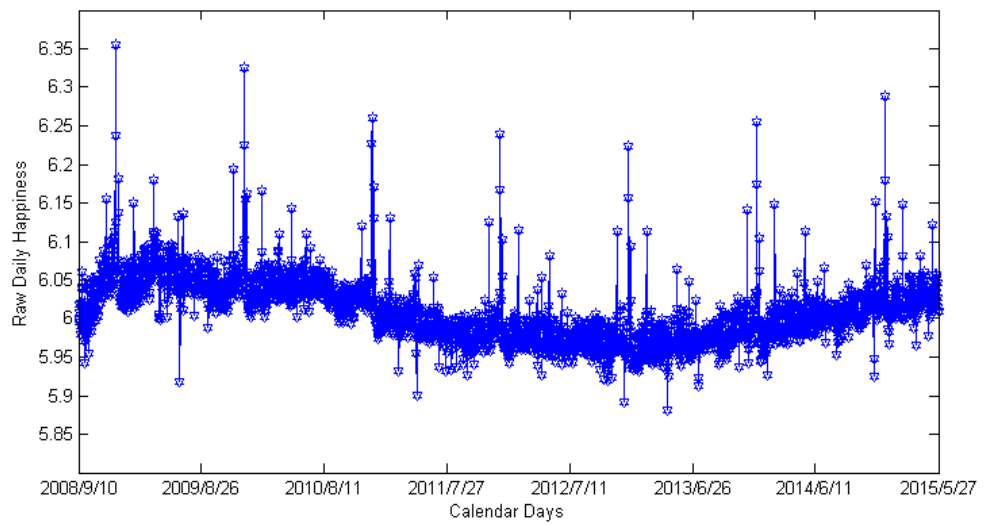
Figure 6.1 Raw daily happiness

This figure illustrates the raw daily happiness, which is the original index downloaded from the website (http://hedonometer.org/index.html).
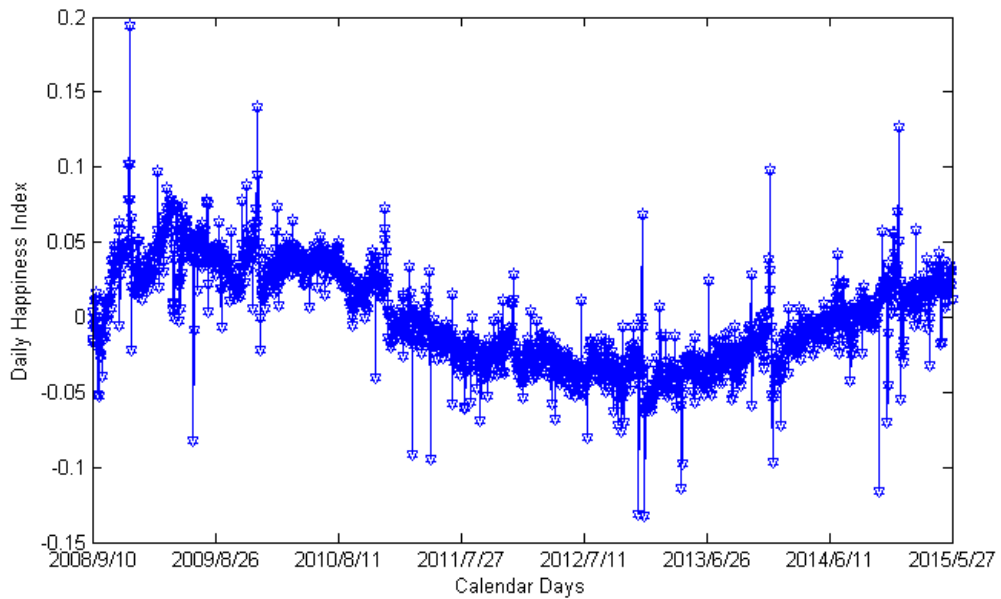
Figure 6.2 Daily happiness index

This figure plots the daily happiness index. We regress the raw daily happiness on weekday dummies and holiday dummies and keep the residual. The holiday includes: Thanksgiving, Christmas Eve, Christmas Day, New Year's Eve, New Year's Day, Valentine's Day, Easter, Father's Day and Mother's Day and Halloween.

6.2.2. Capital Data

Our analysis is conducted on 11 developed stock markets in the world, i.e., USA (S&P 500 and NASDAQ), Canada, France, Germany, UK, Hong Kong, South Korea, Japan, Australia and New Zealand. Almost all of these countries and region have large proportion of Twitter users. The capital market data are from YAHOO! Finance (http://finance.yahoo.com) including the opening price of the index, the highest price of the index, the lowest price of the index, the closing price of the index and the trading volume. The indexes include: S&P 500, NASDAQ Composite, S&P/TSX, CAC 40, DAX, FTSE 100, Hang Seng, KOSPI, Nikkei 225, S&P/ASX 200, and NZX 50 INDEX. All the sample periods are the same as that of the daily happiness index. Table 6.2 summarizes the statistics of the 11 stock indexes returns. It shows that the USA, UK, Australia and New Zealand have positive index returns, while other markets have negative index returns over the sample period. Obviously, all the index returns reveal the skewed pattern in their distributions. Therefore, the log-transformations are used in the following sections for the empirical analysis.

Table 6.2 Summary statistics of the index return

This table shows the descript statistics for the USA, Canada, French, Germany, UK, Hong Kong, South Korea, Japan, Australia and New Zealand. The sample period is September 10th 2008 to May 27th 2015. N denotes the number of observations in each stock market. Observations during non-trading days for each stock market are excluded for each stock market. All the statistics are calculated with the daily index returns.

|	Index	N	Mean	Median	Std	Kurtosis	Max	Min
USA	S&P 500	1688	0.00034	0.00081	0.00018	10.53	0.1024	-0.0912
USA	NASDAQ Composite	1689	0.00030	0.00106	0.01215	7.125	0.0613	-0.0792
Canada	S&P/TSX	1691	-0.00012	0.00042	0.01039	11.36	0.0715	-0.0789
France	CAC 40	1731	-0.00031	0.00025	0.00031	3.191	0.0694	-0.0747
Germany	DAX	1721	-0.00002	0.00032	0.01339	7.553	0.1114	-0.0734
UK	FTSE 100	1699	0.00017	0.00062	0.01299	9.041	0.0938	-0.0926
Hong Kong	Hang Seng	1702	-0.00048	-0.00073	0.01173	17.71	0.1216	-0.1162
South Korea	KOSPI	1722	-0.00052	-0.00025	0.01066	16.29	0.0860	-0.1135
Japan	Nikkei 225	1730	-0.00034	-0.00021	0.01250	17.18	0.1166	-0.1056
Australia	S&P/ASX 200	1726	0.00005	0.00022	0.00978	4.333	0.0465	-0.0764
New Zealand	NZX 50 INDEX	1737	0.00033	0.00076	0.00729	6.994	0.0581	-0.0494

6.3. Contemporaneous Correlation

6.3.1. Kendall Rank Correlation Coefficient

Given the non-normality property of index returns and corresponding daily happiness (table 6.1 and table 6.2), the Kendall Rank Correlation Coefficient (Abdi, 2007) is employed to calculate the correlation coefficient between index return and corresponding daily happiness. This correlation coefficient evaluates the similarity of daily happiness index and corresponding index return. The correlation coefficient takes value between +1 and -1. The negative value denotes daily happiness varies inversely to the index return. To observe the contemporaneous correlation between happiness sentiment and index return at various magnitudes of happiness, for each corresponding daily happiness index (excluding the non-trading days for certain stock market), we first classify it into quintiles from the least to the happiest days, i.e., the least-happiness subgroup, the 2-happiness subgroup, 3-happiness subgroup, 4-happiness subgroup and most-happiness subgroup. And then, we calculate the Kendall correlation coefficients between happiness sentiment and corresponding index return for each stock market index. In that sense, there are 5 Kendall correlation coefficients for each stock market index and there are 11 observations in each subgroup. Table 6.3 reports the Kendall correlation coefficients in different subgroups for all the 11 stock markets. To avoid the multiple comparisons problem, we employ the Bonferroni correction method to correct the p-values. The significance cut-off is set to α/n ($\alpha = 0.05$ and $n = 11$). The empirical results show that there exist positive correlation coefficients in the 4 and most-happiness subgroups. These findings suggest that the happiness sentiment is much more related to the stock market performance in the happier trading days.

Table 6.3 Kendall correlation coefficients between index return and happiness sentiment

This table reports the Kendall correlation coefficient for the 11 stock indexes and value in the parenthese is the p-value. All the p-values are corrected with the Bonferroni correction method and the significance level is denoted after this correction. N denotes the number of observation in each subgroup. We perform the two-sample t-test to detect the significant differences to zero. The results show that there are no significant correlation coefficients for the least subgroup, the 2 subgroup and the 3 subgroup. However, there exists significant correlation coefficients for the, the 4 subgroup (p-value=0.0270) and most subgroup (p-value=0.0120).

Index	N	Least	2	3	4	Most
S&P 500	337	0.07152	-0.01677	0.00358	0.11735**	0.04417
		(0.0510)	(0.6485)	(0.9225)	(0.0014)	(0.2279)
NASDAQ Composite	337	0.07812	0.02324	-0.03019	0.10857**	0.03889
		(0.0331)	(0.5274)	(0.4104)	(0.0031)	(0.2884)
S&P/TSX	338	0.08249	-0.02943	0.03317	0.09580	0.12185***
		(0.0242)	(0.4227)	(0.3651)	(0.0089)	(0.0000)
CAC 40	346	0.06089	-0.00845	0.03658	0.07139	0.02326
		(0.0923)	(0.8161)	(0.3121)	(0.0485)	(0.5200)
DAX	344	0.08245	-0.01378	0.04387	0.04295	0.07827
		(0.0230)	(0.7049)	(0.2268)	(0.2367)	(0.0308)
FTSE 100	339	0.11502**	0.01279	-0.00529	0.10861**	0.07527
		(0.0016)	(0.7273)	(0.8852)	(0.0030)	(0.0393)
Hang Seng	340	0.03945	0.09055	0.02865	0.06371	0.19418***
		(0.2797)	(0.0133)	(0.4327)	(0.0809)	(0.0000)
KOSPI	346	-0.05838	0.04556	-0.06197	0.10407**	0.14624***
		(0.1065)	(0.2089)	(0.0868)	(0.0040)	(0.0000)
Nikkei 225	345	-0.04566	-0.02563	-0.02481	0.05026	0.07790
		(0.2075)	(0.4805)	(0.4937)	(0.1654)	(0.0313)
S&P/ASX 200	347	0.01416	-0.04166	-0.01587	0.04846	0.13001***
		(0.6952)	(0.2499)	(0.6606)	(0.1798)	(0.0000)
NZX 50 INDEX	335	-0.03688	0.02061	-0.05311	-0.00092	0.10198**
		(0.3159)	(0.5761)	(0.1487)	(0.9805)	(0.0055)

Notes: *** indicates statistical significance at 1% level. ** indicates statistical significance at 5% level.

6.3.2. Multivariable Linear Regression

The price-volume relations have received considerable attention in finance literature (Karpoff, 1987; Lamoureux and Lastrapes, 1990 and Wang, 1994). Most of these studies view trading volume as the proxy for information arrival and test the predictive power of trading volume on index return. In order to explore the additional explanatory power of the happiness sentiment, we further investigate the

contemporaneous correlation between happiness sentiment and corresponding index return with the consideration of trading volume. In order to investigate the distinct explanatory power of daily happiness sentiment on index return, we choose to only focus on the least and most-happiness subgroups instead of emphasizing on the full sample. Therefore, we construct the following regression models:

$$r_{i,t} = log_{10}^{(\frac{p_{i,t}}{p_{i,t-1}})}$$
(6-1)

$$rl_{i,t} = \alpha_L + \beta_l DHL_{i,t} + \gamma_L \log(TVL_{i,t}) + \varepsilon_{i,t} \qquad (6\text{-}2)$$

$$rm_{i,t} = \alpha_M + \beta_M DHM_{i,t} + \gamma_M \log(TVM_{i,t}) + \varepsilon_{i,t}$$
(6-3)

where $p_{i,t}$ is the closing price for index i on day t, $DHL_{i,t}$ is happiness for index i on day t in the least-happiness subgroup, $DHM_{i,t}$ is happiness for index i on day t in the most-happiness subgroup, $TVL_{i,t}$ is the trading volume for index i on day t in the least-happiness subgroup, $TVM_{i,t}$ is the trading volume for index i on day t in the most-happiness subgroup, α_L, β_L, γ_L, α_M, β_M and γ_M are the regression coefficients and $\varepsilon_{i,t}$ is the regression error. The $DHL_{i,t}$, $DHM_{i,t}$, $TVL_{i,t}$ and $TVM_{i,t}$ are passed through the KPSS test and Augmented Dickey-Fuller test confirming the stationary property.

Table 6.4 reports the contemporaneous relations between happiness sentiment and corresponding index return in the least and most-happiness subgroups. The Bonferroni correction method is also employed to correct the p-values for avoiding the multiple comparisons problem. The empirical results show the significantly positive (p-value=0.0141) relations between happiness sentiment and stock returns, which is in line with existing literature on sentiment extracted from Facebook (Siganos *et al.*, 2014). These findings suggest that the happier the sentiment, the higher the return, and vice-versa.

Table 6.4 The contemporaneous relations

This table reports the contemporaneous relations between happiness sentiment and corresponding index return in the least and most-happiness subgroups. All the variables in equations (2) and (3) are passed through the KPSS test and Augmented Dickey-Fuller test confirming the stationary property. The estimated parameter represents the coefficient of regressing daily index return on our happiness sentiment from Twitter. Value in the parenthese is the p-value and the significant level is corrected with the Bonferroni correction method.

Index	N	β_L	β_M
S&P 500	337	-0.4093	3.1689***
		(0.0534)	(0.0000)
NASDAQ Composite	337	-0.3267	1.5659***
		(0.1446)	(0.0000)
S&P/TSX	338	-0.2466	0.0480
		(0.3469)	(0.1990)
CAC 40	346	0.0533	1.4880***
		(0.7860)	(0.0000)
DAX	344	-0.4769	0.3907**
		(0.0624)	(0.0027)
FTSE 100	339	-0.3008	1.7003***
		(0.1688)	(0.0000)
Hang Seng	340	-0.1643	1.0074***
		(0.3928)	(0.0000)
KOSPI	346	0.3470	-0.0152
		(0.1081)	(0.6721)
Nikkei 225	345	0.3205	0.0864
		(0.1542)	(0.2084)
S&P/ASX 200	347	-0.2886	-0.0481
		(0.3341)	(0.0856)
NZX 50 INDEX	335	0.0568	0.0266
		(0.7863)	(0.5972)

Notes: ***indicates statistical significance at 1% level. **indicates statistical significance at 5% level.

6.4. Granger Causality

Zhang *et al.* (2013) observe a bi-directional Granger causality pattern: the online searching behavior can granger causes the abnormal return and vice versa. To test the Granger causality between this novel happiness index return and corresponding index return, we construct the following regression models (Granger, 1988):

$$r_t = u_r + \sum_{i=1}^{p} a_i r_{t-i} + \sum_{j=1}^{q} \beta_j DH_{t-j} + \varepsilon_t \qquad\qquad (6\text{-}4)$$

$$DH_t = u_{DH} + \sum_{i=1}^{p} a_i DH_{t-i} + \sum_{j=1}^{q} \beta_j r_{t-j} + \varepsilon_t$$

(6-5)

where p and q denote the lag length, r_{t-i} and DH_{t-i} denote the index return and corresponding daily happiness index, respectively, a_i and β_j denote the coefficient, u_r and u_{DH} denote the intercept terms and ε_t is regression error. The lag length (p and q) selection is chosen using the Bayesian information Criterion. The r_t and DH_t are passed through the KPSS test and Augmented Dickey-Fuller test confirming the stationary property.

Table 6.5 reports the Granger causality results for the 11 indexes. As is shown, the daily happiness can Granger-cause the return for 63.64% (7 out of 11) indexes and no index return can Granger-cause daily happiness. This one direction pattern (from the daily happiness to index return) indicates that the amount of activity on social media is useful for predicting the stock market performance (Bollen *et al.*, 2011; Joseph *et al.*, 2011 and Zhang *et al.*, 2014).

Table 6.5 Granger-causality test between index return (r) and daily happiness (DH)

This table reports the Granger-causality test between index return and corresponding daily happiness. All the variables are passed through the KPSS test and Augmented Dickey-Fuller test confirming the stationary property. Value in the table is the F-statistics and value in the parenthese is the critical value from the F-distribution. If F-statistics is greater than the critical value, the null hypothesis of no Granger Causality is rejected at certain significance level.

Index	Null: r does not granger cause DH	Null: DH does not granger cause r
S&P 500	4.6212 (6.6501)	22.5569 (2.0494)[***]
NASDAQ Composite	4.5931 (6.6501)	18.1329 (2.0494)[***]
S&P/TSX	0.7191 (6.6500)	31.2633 (2.0923)[***]
CAC 40	0.5359 (6.6497)	0.2445 (6.6496)
DAX	0.4385 (6.6498)	0.0656 (6.6497)
FTSE 100	0.4450 (6.6500)	18.5176 (1.9761)[***]
Hang Seng	0.6269 (6.6499)	23.8712 (1.9158)[***]
KOSPI	1.0046 (6.6497)	1.6351 (6.6496)
Nikkei 225	1.1972 (6.6497)	13.4197 (1.8891)[***]
S&P/ASX 200	1.3848 (6.6496)	5.5818 (6.6495)
NZX 50 INDEX	0.6635 (6.6502)	16.7481 (1.9159)[***]

Notes: [***] indicates Granger-causality test significant at 1% level.

6.5. Cross-sectional Analysis

6.5.1. The Effect of Happiness on Index Return

Previous sections have showed time series positive correlation and Granger causality from happiness sentiment to index return. This section further analyzes the cross-sectional evidence for the index return and the intraday volatility. For each individual index return, we classify it into quintiles based on the corresponding magnitude of happiness sentiment and then calculate the average values in the quintiles for each stock market index (The calculation of the subgroups is the same as description in section 6.3.1). Table 6.6 shows the cross-sectional effect of happiness sentiment on index return. The most-happiness trading days have the largest positive index return (t-value=-3.3438 and p-value=0.0032) and the least-happiness trading days have the smallest negative index return (t-value=2.4572

and p-value=0.0233). A potential explanation for this discrepancy is that the peaks and troughs of happiness sentiment are related to extreme macroeconomic events. Looking at figure 6.2, the high happiness index period occurs coming out of the global financial crisis, when there were large positive returns in U.S. equities. The lower end of the trend includes fall 2011, during the European sovereign debt crisis and when S&P downgraded the U.S. credit rating. With a parsimonious representation of macroeconomic events from (Da *et al.*, 2015), we document a significantly negative between the happiness sentiment and macroeconomic events. Table 6.7 shows the pairwise comparisons of subgroups on index return. As is shown, the most-happiness subgroup is significantly different from the least, 2 and 3-subgroups at 1% level.

Table 6.6 The effect of happiness on index return

This table shows the effect of happiness sentiment on index return on subgroups. The two-sample t-test is performed to test the statistical difference from zero for each subgroup. The corresponding t-value and p-value for each subgroup are as followings: least-happiness subgroup (t-value=2.4572 and p-value=0.0233), 2-happiness subgroup (t-value=1.3585 and p-value=0.1894), 3-happiness subgroup (t-value=1.9121 and p-value=0.0703), 4-happiness subgroup (t-value=-0.6357 and p-value=0.5322) and most-happiness subgroup (t-value=-3.3438 and p-value=0.0032).

Index	Least	2	3	4	Most
S&P 500	-0.0004	0.0002	0.0000	0.0009	0.0010
NASDAQ Composite	0.0003	-0.0005	0.0002	0.0002	0.0013
S&P/TSX	-0.0003	-0.0002	-0.0005	0.0001	0.0003
CAC 40	-0.0007	-0.0001	-0.0005	-0.0002	-0.0004
DAX	-0.0006	-0.0003	0.0002	0.0006	0.0002
FTSE 100	-0.0006	0.0005	-0.0012	0.0013	0.0009
Hang Seng	-0.0014	-0.0001	-0.0011	-0.0004	0.0007
KOSPI	-0.0011	-0.0005	-0.0007	-0.0012	0.0009
Nikkei 225	-0.0003	-0.0005	-0.0013	-0.0004	0.0006
S&P/ASX 200	0.0004	-0.0001	0.0006	0.0001	-0.0002
NZX 50 INDEX	0.0001	0.0002	0.0002	0.0004	0.0004
1000*Mean	-0.4119	-0.1297	-0.3726	0.1296	0.5155

Table 6.7 Pairwise comparisons of subgroups on index return

This table repots the pairwise comparisons of subgroups. The two-sample t-test is performed to test the statistical differences among subgroups. As it showed, the most-happiness subgroup is significantly different from the least, 2 and 3-happiness subgroups at 1% level.

Index	Least	2	3	4	Most
Least		1.4628 (0.1591)	0.1530 (0.8799)	2.0517[*] (0.0535)	4.0721[***] (0.0005)
2			1.1193 (0.2763)	1.1519 (0.2629)	3.5580[***] (0.0020)
3				1.7807[*] (0.0902)	3.5742[***] (0.0019)
4					1.5096 (0.1468)

Notes: [***] indicates significant at 1% level. [*] indicates significant at 10% level.

6.5.2. The Effect of Happiness on Intraday Volatility

In line with some evidence on the relations between sentiment and volatility (Siganos *et al.*, 2014), we also examine these relations with the novel proxy for sentiment. The intraday volatility of each index is calculated as the range-based volatility (Garman and Klass, 1980).

$$V_{i,t} = \frac{1}{2}hl_{i,t}^2 - (2ln2 - 1)oc_{i,t}^2$$

(6-6)

where $hl_{i,t}$ is the difference in natural logarithms of the highest and lowest prices for index i on day t, $oc_{i,t}$ is the difference in natural logarithms of the opening and closing prices for index i on day t.

Table 6.8 and table 6.9 report the results for the effect of happiness on volatility. In the cross-sectional analysis, we observe that the volatility increases with the heightening magnitude of daily happiness. In the Most-happiness subgroup, the

volatility is significantly higher than that in the Least, 2 and 3-subgroups at 10% level.

Table 6.8 The effect of happiness on volatility

This table reports the effect of happiness sentiment on volatility on subgroups. All the value in the table is 1000 times larger than the actual value. The two-sample t-test is performed to test the statistical difference from zero for each subgroup. The corresponding t-value and p-value for each subgroup are as followings: least-happiness subgroup (t-value=7.6549 and p-value=0.0000), 2-happiness subgroup (t-value=7.0757 and p-value=0.0000), 3-happiness subgroup (t-value=8.1223 and p-value=0.0000), 4-happiness subgroup (t-value=8.1198 and p-value=0.0000) and most-happiness subgroup (t-value=7.8321 and p-value=0.0000).

Index	Least	2	3	4	Most
S&P 500	0.06631	0. 06819	0. 11419	0. 15181	0. 17060
NASDAQ Composite	0. 07026	0. 07351	0. 10932	0. 16040	0. 16448
S&P/TSX	0. 06809	0. 06883	0. 13686	0. 15767	0. 19406
CAC 40	0. 12125	0. 13239	0. 13521	0. 17263	0. 19418
DAX	0. 13510	0. 13956	0. 14679	0. 17063	0. 21191
FTSE 100	0. 08074	0. 08971	0. 10957	0. 12991	0. 16496
Hang Seng	0. 06170	0. 06785	0. 10184	0. 14517	0. 15562
KOSPI	0. 07034	0. 06522	0. 08199	0. 16652	0. 14787
Nikkei 225	0. 07692	0. 09174	0. 09603	0. 11204	0. 10609
S&P/ASX 200	0. 03336	0. 02523	0. 02675	0. 02608	0. 03076
NZX 50 INDEX	0. 02655	0. 02567	0. 02670	0. 03169	0. 03768
Mean	0.0737	0.0771	0.0987	0.1295	0.1435

Table 6.9 Pairwise comparisons of subgroups on volatility

This table reports the pairwise comparisons of subgroups on the range-based volatility. The two-sample t-test is performed to test the statistical differences among subgroups. As it showed, the volatility in the most-happiness subgroup is significant different from the least, 2 and 3-happiness subgroups at 10% level.

Index	Least	2	3	4	Most
Least		0.2331 (0.8180)	1.6107 (0.1229)	2.9958*** (0.0071)	3.3719*** (0.0030)
2			1.3223 (0.2010)	2.7141** (0.0134)	3.1150*** (0.0055)
3				1.5386 (0.1396)	2.0389* (0.0549)

		0.5751
4		(0.5716)

Notes: *** indicates significant at 1% level. ** indicates significant at 5% level. * indicates significant at 10% level.

6.6. Robustness

To ensure that our main results are not driven by the classification of subgroup, we re-estimate the models in section 6.3.1 and 6.3.2 with classifying the happiness sentiment into quantiles from the least to the happiest days, i.e., the least-happiness subgroup, the 2-happiness subgroup, 3-happiness subgroup and most-happiness subgroup. Table 6.10 reports the main results on Kendall correlation coefficients, β_L (β_M) as well as the impact of happiness sentiment on index return and range-based volatility. All the results are consistent with previous findings in sections 6.3 and 6.5.

Table 6.10 Main results on alternative subgroup classification

The definition and calculation method of the Kendall correlation coefficients, β_L (β_M), index return and range-based volatility are the same as in table 3, table 4, table 6 and table 8, respectively.

Variables	Least subgroup	Most subgroup	Difference	P-value
Kendall correlation coefficients	NaN	0.0612	0.0612	0.0299**
β_L or β_M	NaN	0.8265	0.8265	0.0170**
Index return	-0.2566	0.1521	0.4087	0.0198**
Range-based volatility	0.0743	0.1439	0.0696	0.0007***

Notes: NaN indicates no significant value. *** indicates significant at 1% level. ** indicates significant at 5% level.

6.7. Discussions and Conclusions

This chapter employs daily happiness index from Twitter as a novel proxy for sentiment and investigate its relations to index return across 11 international stock markets. By partitioning the happiness sentiment into quintiles from the least to the happiest days, we firstly observe the statistical property of daily happiness and show

the contemporary correlation coefficients between happiness and index return in the 4 and most-happiness subgroups are higher than that in the least, 2 and 3-happiness subgroups. Secondly, the explanatory power of happiness sentiment varies across magnitudes of sentiment: happiness sentiment contains more explanatory power in the most-happiness subgroup with the consideration of trading volume. Thirdly, the Granger causality test shows that happiness sentiment can granger-cause the changes in index return for seven stock markets. Fourthly, in the cross-sectional analysis, we find that the index return and the range-based volatility of the most-happiness subgroup are larger than those of other subgroups.

These findings suggest that there exists significant dependence between online sentiment and overall stock market performance. However, our paper only focuses on the developed stock markets. The emerging stock markets are ignored because of the lack of online sentiment data. This is mainly due to the limitations of natural language processing technique for non-English languages. For example, there are so many compounding and ambiguous words in Chinese; hence the classification accuracy for these words is not sufficient for financial studies. Investigating the connection between social media activities and the stock market performance has become a hot topic in econophysics, which needs multi-disciplinary efforts from computer scientists, financial economists, econometricians and econophysics scholars.

Bibliography

Abdi H, The Kendall Rank Correlation Coefficient, N. Salkind (Ed.), *Encyclopedia of measurement and statistics*, Sage, Thousand Oaks 2007; 508-510

Admati AR, Pfleiderer P, A Theory of Intraday Patterns: Volume and Price Variability. *Review of Financial Studies* 1988;1; 3-40

Ahern KR, Sosyura D, Who Writes the News? Corporate Press Releases during Merger Negotiations. *Journal of Finance* 2014;69; 241-291

Ahn D, Choi S, Gale D, Kariv S, Estimating Ambiguity Aversion in a Portfolio Choice Experiment. *Quantitative Economics* 2014;5; 195-223

Albert Jr RL, Smaby TR, Market Response to Analyst Recommendations in the "Dartboard" Column: The Information and Price-Pressure Effects. *Review of Financial Economics* 1996;5; 59-74

Allen F, Gale D, Financial Contagion. *Journal of Political Economy* 2000;108; 1-33

Amihud Y, Mendelson H, Trading Mechanisms and Stock Returns: An Empirical Investigation. *Journal of Finance* 1987;42; 533-553

Andersen TG, Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility. *Journal of Finance* 1996;51; 169-204

Anton M, Polk C, Connected Stocks. *Journal of Finance* 2014;69; 1099-1127

Antweiler W, Frank MZ, Is all that talk just noise? The Information Content of Internet Stock Message Boards. *Journal of Finance* 2004;59; 1259-1294

Aouadi A, Arouri M, Teulon F, Investor Attention and Stock Market Activity: Evidence from France. *Economic Modelling* 2013;35; 674-681

Bagnoli M, Beneish MD, Watts SG, Whisper Forecasts of Quarterly Earnings per Share. *Journal of Accounting and Economics* 1999;28; 27-50

Bailey W, Cai J, Cheung YL, Wang F, Stock Returns, Order Imbalances, and Commonality: Evidence on Individual, Institutional, and Proprietary Investors in China. *Journal of Banking & Finance* 2009;33; 9-19

Baker M, Stein JC, Market Liquidity as a Sentiment Indicator. *Journal of Financial Markets* 2004;7; 271-299

Baker M, Wurgler J, Investor Sentiment and the Cross-Section of Stock Returns. *Journal of Finance* 2006;61; 1645-1680

Baker M, Wurgler J, Investor Sentiment in the Stock Market. *Journal of Economic Perspectives* 2007;21; 129-152

Bank M, Larch M, Peter G, Google Search Volume and Its Influence on Liquidity and Returns of German Stocks. *Financial Markets and Portfolio Management*

2011;25; 239-264

Barber BM, Loeffler D, The "Dartboard" Column: Second-Hand Information and Price Pressure. *Journal of Financial and Quantitative Analysis* 1993;28; 273-284

Barber BM, Odean T, All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies* 2008;21; 785-818

Barber BM, Odean T, The Behavior of Individual Investors. Working Paper, SSRN, 2011

Barber BM, Odean T, The Internet and the Investor. *Journal of Economic Perspectives* 2001;15; 41-54

Barberis N, Shleifer A, Style Investing. *Journal of Financial Economics* 2003;68; 161-199

Barberis N, Shleifer A, Vishny R, A Model of Investor Sentiment. *Journal of Financial Economics* 1998;49; 307-343

Barberis N, Shleifer A, Wurgler J, Comovement. *Journal of Financial Economics* 2005;75; 283-317

Batsell, J., Gossip Central---Internet Message Boards Can Leave Some Stocks Hanging by a Thread, *Seattle Times*, September 14, 1998

Bekaert G, Harvey C, Ng A, Market Integration and Contagion. *Journal of Business* 2005;78; 39-69

Benartzi S, Excessive Extrapolation and the Allocation of 401(k) Accounts to Company Stock. *Journal of Finance* 2001;56; 1747-1764

Bennett, J., Traffic on Financial Web Pages Rises When the Market Falls, *Dow Jones News Service*, September 1, 1998

Bergemann D, Heumann T, Morris S, Information and Volatility. *Journal of Economic Theory* 2015;158, Part B; 427-465

Billio M, Getmansky M, Lo AW, Pelizzon L, Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors. *Journal of Financial Economics* 2012;104; 535-559

Blankespoor E, Miller GS, White HD, The Role of Dissemination in Market Liquidity: Evidence from Firms' Use of Twitter™. *Accounting Review* 2014;89; 79-112

Boehmer E, Masumeci J, Poulsen AB, Event-study Methodology under Conditions of Event-induced Variance. *Journal of Financial Economics* 1991;30; 253-272

Böhme R, Holz T, The Effect of Stock Spam on Financial Markets. Working Paper, SSRN, 2006

Bollen J, Mao H, Zeng X, Twitter mood predicts the stock market. *Journal of Computational Science* 2011;2; 1-8

Bollerslev T, Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 1986;31; 307-327

Bossaerts P, Ghirardato P, Guarnaschelli S, Zame WR, Ambiguity in Asset Markets: Theory and Experiment. *Review of Financial Studies* 2010;23; 1325-1359

Boyle P, Garlappi L, Uppal R, Wang T, Keynes Meets Markowitz: The Trade-Off between Familiarity and Diversification. *Management Science* 2012;58; 253-272

Brown GW, Cliff MT, Cliff, Investor Sentiment and Asset Valuation. *Journal of Business* 2005;78; 405-440

Brown GW, Cliff MT, Investor Sentiment and the Near-term Stock Market. *Journal of Empirical Finance* 2004;11; 1-27

Brown SJ, Warner JB, Measuring Security Price Performance. *Journal of Financial Economics* 1980;8; 205-258

Brown SJ, Warner JB, Using Daily Stock Returns: The Case of Event Studies. *Journal of Financial Economics* 1985;14; 3-31

Calvo GA, Contagion in emerging markets: when Wall Street is a carrier. 1999;

Campbell JY, Lettau M, Malkiel BG, Xu Y, Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk. *Journal of Finance* 2001;56; 1-43

Carhart MM, On Persistence in Mutual Fund Performance. *Journal of Finance* 1997;52; 57-82

Chan K, Fong W-M, Trade Size, Order Imbalance, and the Volatility–Volume Relation. *Journal of Financial Economics* 2000;57; 247-273

Chan WS, Stock Price Reaction to News and No-news: Drift and Reversal after Headlines. *Journal of Financial Economics* 2003;70; 223-260

Chen H, Chong TTL, She Y, A Principal Component Approach to Measuring Investor Sentiment in China. *Quantitative Finance* 2014;14; 573-579

Chen H, De P, Hu Y, Hwang B-H, Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies* 2014;27; 1367-1403

Chiang TC, Qiao Z, Wong W-K, New Evidence on the Relation between Return Volatility and Trading Volume. *Journal of Forecasting* 2010;29; 502-515

Chordia T, Roll R, Subrahmanyam A, Liquidity and Market Efficiency. *Journal of Financial Economics* 2008;87; 249-268

Chordia T, Roll R, Subrahmanyam A, Order Imbalance, Liquidity, and Market Returns. *Journal of Financial Economics* 2002;65; 111-130

Chordia T, Subrahmanyam A, Order Imbalance and Individual Stock Returns: Theory and Evidence. *Journal of Financial Economics* 2004;72; 485-518

Clark PK, A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. *Econometrica* 1973;41; 135-155

Clarkson PM, Joyce D, Tutticci I, Market Reaction to Takeover Rumour in Internet Discussion Sites. *Accounting & Finance* 2006;46; 31-52

Cooper I, Kaplanis E, Home Bias in Equity Portfolios, Inflation Hedging, and International Capital Market Equilibrium. *Review of Financial Studies* 1994;7; 45-60

Copeland TE, A Model of Asset Trading Under the Assumption of Sequential Information Arrival. *Journal of Finance* 1976;31; 1149-1168

Coval JD, Moskowitz TJ, Home Bias at Home: Local Equity Preference in Domestic Portfolios. *Journal of Finance* 1999;54; 2045-2073

Coval JD, Moskowitz TJ, The Geography of Investment: Informed Trading and Asset Prices. *Journal of Political Economy* 2001;109; 811-841

Cutler DM, Poterba JM, Summers LH, What Moves Stock Prices? *Journal of Portfolio Management* 1989;15; 4-12

Da Z, Engelberg J, Gao P, The Sum of All FEARS Investor Sentiment and Asset Prices. *Review of Financial Studies* 2015;28; 1-32

Da ZHI, Engelberg J, Gao P, In Search of Attention. *Journal of Finance* 2011;66; 1461-1499

Dang TL, Moshirian F, Zhang B, Commonality in News around the World. *Journal of Financial Economics* 2015;116; 82-110

Daniel K, Hirshleifer D, Subrahmanyam A, Investor Psychology and Security Market Under- and Overreactions. *Journal of Finance* 1998;53; 1839-1885

Darrat AF, Rahman S, Zhong M, Intraday Trading Volume and Return Volatility of the DJIA Stocks: A Note. *Journal of Banking & Finance* 2003;27; 2035-2043

Darrat AF, Zhong M, Cheng LTW, Intraday Volume and Volatility Relations with and without Public News. *Journal of Banking & Finance* 2007;31; 2711-2729

Das S, Martínez-Jerez A, Tufano P, eInformation: A Clinical Study of Investor Discussion and Sentiment. *Financial Management* 2005;34; 103-137

Das SR, Chen MY, Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science* 2007;53; 1375-1388

De Bondt WFM, Thaler R, Does the Stock Market Overreact? *Journal of Finance* 1985;40; 793-805

De Long JB, Shleifer A, Summers LH, Waldmann RJ, Noise TraderRisk in Financial Markets. *Journal of Political Economy* 1990;98; 703-738

Deb P, Trivedi PK, Varangis P, The Excess Co-Movement of Commodity Prices Reconsidered. *Journal of Applied Econometrics* 1996;11; 275-291

Dewally M, Internet Investment Advice: Investing with a Rock of Salt. *Financial Analysts Journal* 2003;59; 65-77

Dimpfl T, Jank S, Can Internet Search Queries Help to Predict Stock Market Volatility? *European Financial Management* 2016;22; 171-192

Ding R, Hou W, Retail Investor Attention and Stock Liquidity. *Journal of International Financial Markets, Institutions and Money* 2015;37; 12-26

Døskeland TM, Hvide HK, Do Individual Investors Have Asymmetric Information Based on Work Experience? *Journal of Finance* 2011;66; 1011-1041

Drake MS, Roulstone DT, Thornock JR, Investor Information Demand: Evidence from Google Searches Around Earnings Announcements. *Journal of Accounting Research* 2012;50; 1001-1040

Dzielinski M, Measuring Economic Uncertainty and Its Impact on the Stock Market. *Finance Research Letters* 2012;9; 167-175

Easley D, Kiefer NM, O'Hara M, One Day in the Life of a Very Common Stock. *Review of Financial Studies* 1997a;10; 805-835

Easley D, Kiefer NM, O'Hara M, The Information Content of the Trading Process. *Journal of Empirical Finance* 1997b;4; 159-186

Edelen RM, Marcus AJ, Tehranian H, Relative Sentiment and Stock Returns. *Financial Analysts Journal* 2010;66; 20-32

Einav L, Levin J, The Data Revolution and Economic Analysis. *Innovation Policy and the Economy* 2014;14; 1-24

Engelberg JE, Parsons CA, The Causal Impact of Media in Financial Markets. *Journal of Finance* 2011;66; 67-97

Engle RF, Rangel JG, The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes. *Review of Financial Studies* 2008;21; 1187-1222

Epps TW, Epps ML, The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis. *Econometrica* 1976;44; 305-321

Fama EF, Efficient Capital Market: A Review of Theory and Empirical Work. *Journal of Finance* 1970;25; 383-417

Fang L, Peress J, Media Coverage and the Cross-section of Stock Returns. *Journal of Finance* 2009;64; 2023-2052

Felton J, Kim J, Warnings from the Enron Message Board. *Journal of Investing* 2002;11; 29-52

Feng L, Seasholes MS, Individual Investors and Gender Similarities in an Emerging Stock Market. *Pacific-Basin Finance Journal* 2008;16; 44-60

Fisher KL, Statman M, Investor Sentiment and Stock Returns. *Financial Analysts Journal* 2000;56; 16-23

Fleming J, Kirby C, Ostdiek B, Information and Volatility Linkages in the Stock, Bond, and Money Markets. *Journal of Financial Economics* 1998;49; 111-137

Fleming J, Kirby C, Ostdiek B, Stochastic Volatility, Trading Volume, and the Daily Flow of Information. *Journal of Business* 2006;79; 1551-1590

Forbes KJ, Rigobon R, No Contagion, Only Interdependence: Measuring Stock Market Comovements. *Journal of Finance* 2002;57; 2223-2261

Fox CR, Tversky A, Ambiguity Aversion and Comparative Ignorance. *Quarterly Journal of Economics* 1995;110; 585-603

French KR, Poterba JM, Investor Diversification and International Equity Markets. *American Economic Review* 1991;81; 222-226

French KR, Roll R, Stock Return Variances: The Arrival of Information and the Reaction of Traders. *Journal of Financial Economics* 1986;17; 5-26

Fu F, Idiosyncratic Risk and the Cross-section of Expected Stock Returns. *Journal of Financial Economics* 2009;91; 24-37

Garman MB, Klass MJ, On the Estimation of Security Price Volatilities from Historical Data. *Journal of Business* 1980;53; 67-78

Gehrig T, An Information Based Explanation of the Domestic Bias in International Equity Investment. *Scandinavian Journal of Economics* 1993;95; 97-109

Glosten LR, Milgrom PR, Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics* 1985;14; 71-100

Goldstein, A., Money Messages: Electronic Message Boards Are a Good Way to Get Investing Facts and Fiction, *Dallas Morning News*, August 3, 1998

Graham JR, Harvey CR, Huang H, Investor Competence, Trading Frequency, and Home Bias. *Management Science* 2009;55; 1094-1106

Granger CWJ, Some Recent Development in a Concept of Causality. *Journal of Econometrics* 1988;39; 199-211

Greenwood R, Excess Comovement of Stock Returns: Evidence from Cross-Sectional Variation in Nikkei 225 Weights. *Review of Financial Studies* 2008;21; 1153-1186

Griffin JM, Hirschey NH, Kelly PJ, How Important Is the Financial Media in Global Markets? *Review of Financial Studies* 2011;24; 3941-3992

Grinblatt M, Keloharju M, How Distance, Language, and Culture Influence Stockholdings and Trades. *Journal of Finance* 2001;56; 1053-1073

Grossman SJ, Stiglitz JE, On the Impossibility of Informationally Efficient Markets. *American Economic Review* 1980;70; 393-408

Grullon G, Kanatas G, Weston JP, Advertising, Breadth of Ownership, and Liquidity. *Review of Financial Studies* 2004;17; 439-461

Grundy BD, McNichols M, Trade and the Revelation of Information through Prices and Direct Disclosure. *Review of Financial Studies* 1989;2; 495-526

Gurun UG, Butler AW, Don't Believe the Hype: Local Media Slant, Local Advertising, and Firm Value. *Journal of Finance* 2012;67; 561-598

Gwilym Oa, Hasan I, Wang Q, Xie R, In Search of Concepts: The Effects of Speculative Demand on Stock Returns. *European Financial Management* 2015; DOI: 10.1111/eufm.12067

Hanke M, Hauser F, On the Effects of Stock Spam E-mails. *Journal of Financial Markets* 2008;11; 57-83

Harmon, A., The Market Turmoil: Investors on Line, *New York Times*, September 1, 1998

Harris L, Transaction Data Tests of the Mixture of Distributions Hypothesis. *Journal of Financial and Quantitative Analysis* 1987;22; 127-141

Harris M, Raviv A, Differences of Opinion Make a Horse Race. *Review of Financial Studies* 1993;6; 473-506

Hayek FA, The Use of Knowledge in Society. *American Economic Review* 1945;35; 519-530

Heath C, Tversky A, Preference and Belief: Ambiguity and Competence in Choice under Uncertainty. *Journal of Risk and Uncertainty* 1991;4; 5-28

Hu N, Liu L, Tripathy A, Yao LJ, Value Relevance of Blog Visibility. *Journal of Business Research* 2011;64; 1361-1368

Huberman G, Familiarity Breeds Investment. *Review of Financial Studies* 2001;14; 659-680

Huberman G, Jiang W, Offering versus Choice in 401(k) Plans: Equity Exposure and Number of Funds. *Journal of Finance* 2006;61; 763-801

Huberman G, Regev T, Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar. *Journal of Finance* 2001;56; 387-396

Irresberger F, Mühlnickel J, Weiß GNF, Explaining Bank Stock Performance with Crisis Sentiment. *Journal of Banking & Finance* 2015;59; 311-329

Ivković Z, Sialm C, Weisbenner S, Portfolio Concentration and the Performance of

Individual Investors. *Journal of Financial and Quantitative Analysis* 2008;43; 613-655

Ivković Z, Weisbenner S, Local Does as Local Is: Information Content of the Geography of Individual Investors' Common Stock Investments. *Journal of Finance* 2005;60; 267-306

Jacobs H, Weber M, The Trading Volume Impact of Local Bias: Evidence from a Natural Experiment. *Review of Finance* 2011;

Jegadeesh N, Titman S, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance* 1993;48; 65-91

Jennings RH, Starks LT, Fellingham JC, An Equilibrium Model of Asset Trading with Sequential Information Arrival. *Journal of Finance* 1981;36; 143-161

Jensen MC, Some Anomalous Evidence regarding Market Efficiency. *Journal of Financial Economics* 1978;6; 95-101

Jin L, Myers SC, R^2 around the World: New Theory and New Tests. *Journal of Financial Economics* 2006;79; 257-292

Jin X, Shen D, Zhang W, Has Microblogging Changed Stock Market Behavior? Evidence from China. *Physica A: Statistical Mechanics and its Applications* 2016;452; 151-156

Jones AL, Have Internet Message Boards Changed Market Behavior? *Info* 2006;8; 67-76

Joseph K, Babajide Wintoki M, Zhang Z, Forecasting Abnormal Stock Returns and Trading Volume using Investor Sentiment: Evidence from online search. *International Journal of Forecasting* 2011;27; 1116-1127

Kahn V. M., 1997, A 401(k) with One Big Gun Is One Big Risk, New York Times, April 6, 1997

Kahneman D, Tversky A, Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 1979;47; 263-291

Kalev PS, Liu W-M, Pham PK, Jarnecic E, Public Information Arrival and Volatility of Intraday Stock Returns. *Journal of Banking & Finance* 2004;28; 1441-1467

Kallberg J, Pasquariello P, Time-series and Cross-sectional Excess Comovement in Stock Indexes. *Journal of Empirical Finance* 2008;15; 481-502

Kang J-K, Stulz RM, Why Is There a Home Bias? An Analysis of Foreign Portfolio Equity Ownership in Japan. *Journal of Financial Economics* 1997;46; 3-28

Karolyi GA, Stulz RM, Why Do Markets Move Together? An Investigation of U.S.-Japan Stock Return Comovements. *Journal of Finance* 1996;51; 951-986

Karpoff JM, The Relation Between Price Changes and Trading Volume: A Survey.

Journal of Financial and Quantitative Analysis 1987;22; 109-126

Kerl AG, Walter A, Market Responses to Buy Recommendations Issued by Personal Finance Magazines: Effects of Information, Price-Pressure, and Company Characteristics. *Review of Finance* 2007;11; 117-141

Kevin CWC, Yuan H, Earnings Management and Capital Resource Allocation: Evidence from China's Accounting-Based Regulation of Rights Issues. *Accounting Review* 2004;79; 645-665

Kim O, Verrecchia RE, Market Reaction to Anticipated Announcements. *Journal of Financial Economics* 1991;30; 273-309

Kim S-H, Kim D, Investor Sentiment from Internet Message Postings and the Predictability of Stock Returns. *Journal of Economic Behavior & Organization* 2014;107, Part B; 708-729

King MA, Wadhwani S, Transmission of Volatility between Stock Markets. *Review of Financial Studies* 1990;3; 5-33

Klibanoff P, Lamont O, Wizman TA, Investor Reaction to Salient News in Closed-End Country Funds. *Journal of Finance* 1998;53; 673-699

Kodres LE, Pritsker M, A Rational Expectations Model of Financial Contagion. *Journal of Finance* 2002;57; 769-799

Kyle AS, Continuous Auctions and Iinsider Trading. *Econometrica* 1985;53; 1315-1335

Kyle AS, Xiong W, Contagion as a Wealth Effect. *Journal of Finance* 2001;56; 1401-1440

Lakonishok J, Shleifer A, Vishny RW, Contrarian Investment, Extrapolation, and Risk. *Journal of Finance* 1994;49; 1541-1578

Lamoureux CG, Lastrapes WD, Heteroskedasticity in Stock Return Data: Volume versus GARCH Effects. *Journal of Finance* 1990;45; 221-229

Le V, Zurbruegg R, The Role of Trading Volume in Volatility Forecasting. *Journal of International Financial Markets, Institutions and Money* 2010;20; 533-555

Lee CF, Rui OM, Does Trading Volume Contain Information to Predict Stock Returns? Evidence from China's Stock Markets. *Review of Quantitative Finance and Accounting* 2000;14; 341-360

Lee CMC, Shleifer A, Thaler RH, Investor Sentiment and the Closed-end Fund Puzzle. *Journal of Finance* 1991;46; 75-109

Lee DW, Liu MH, Does More Information in Stock Price Lead to Greater or Smaller Idiosyncratic Return Volatility? *Journal of Banking & Finance* 2011;35; 1563-1580

Lee WY, Jiang CX, Indro DC, Stock Market Volatility, Excess Returns, and the Role

of Investor Sentiment. *Journal of Banking & Finance* 2002;26; 2277-2299

Lehavy R, Sloan R, Investor Recognition and Stock Returns. *Review of Accounting Studies* 2008;13; 327-361

Lemmon M, Portniaguina E, Consumer Confidence and Asset Prices: Some Empirical Evidence. *Review of Financial Studies* 2006;19; 1499-1529

Lewis KK, Trying to Explain Home Bias in Equities and Consumption. *Journal of Economic Literature* 1999;37; 571-608

Liang B, Price Pressure: Evidence from the "Dartboard" Column. *Journal of Business* 1999;72; 119-134

Lin J, Wang M, Cai L, Are the Fama–French Factors Good Proxies for Latent Risk Factors? Evidence from the Data of SHSE in China. *Economics Letters* 2012;116; 265-268

Liu P, Smith SD, Syed AA, Stock Price Reactions to the Wall Street Journal's Securities Recommendations. *Journal of Financial and Quantitative Analysis* 1990;25; 399-410

Loibl C, Hira TK, Investor Information Search. *Journal of Economic Psychology* 2009;30; 24-41

Lou D, Attracting Investor Attention through Advertising. *Review of Financial Studies* 2014;27; 1797-1829

Loughran T, Schultz P, Liquidity: Urban Versus Rural Firms. *Journal of Financial Economics* 2005;78; 341-374

Lux T, Sentiment Dynamics and Stock Returns: the Case of the German Stock Market. *Empirical Economics* 2011;41; 663-679

Maremount, M., Predeal Trading in U.S. Surgical Puts Spotlight on Cyberinvestors, *Wall Street Journal*, May 28, 1998

Massa M, Simonov A, Hedging, Familiarity and Portfolio Choice. *Review of Financial Studies* 2006;19; 633-685

Mathur I, Waheed A, Stock Price Reactions to Securities Recommended in Business Week's "Inside Wall Street". *Financial Review* 1995;30; 583-604

McMillan DG, García RQ, Does Information Help Intra-Day Volatility Forecasts? *Journal of Forecasting* 2013;32; 1-9

Medill, G., Chicago Firm Wants to Know What Yahoo! Left Messages, *Chicago Daily Herald*, October 12, 1998

Merton RC, A Simple Model of Capital Market Equilibrium with Incomplete Information. *Journal of Finance* 1987;42; 483-510

Meschke, F. J., CEO Interviews on CNBC. Working Paper, Arizona State University,

2004

Miller EM, Risk, Uncertainty, and Divergence of Opinion. *Journal of Finance* 1977;32; 1151-1168

Mitchell ML, Mulherin JH, The Impact of Public Information on the Stock Market. *Journal of Finance* 1994;49; 923-950

Morck R, Yeung B, Yu W, The Information Content of Stock Markets: Why Do Emerging Markets Have Synchronous Stock Price Movements? *Journal of Financial Economics* 2000;58; 215-260

Neal R, Wheatley SM, Do Measures of Investor Sentiment Predict Returns? *Journal of Financial and Quantitative Analysis* 1998;33; 523-547

Nofsinger JR, Varma A, Individuals and Their Local Utility Stocks: Preference for the Familiar. *Financial Review* 2012;47; 423-443

Osborne MFM, Brownian Motion in the Stock Market. *Operations Research* 1959;7; 145-173

Park B-J, Surprising Information, the MDH, and the Relationship between Volatility and Trading Volume. *Journal of Financial Markets* 2010;13; 344-366

Pasquariello P, Imperfect Competition, Information Heterogeneity, and Financial Contagion. *Review of Financial Studies* 2007;20; 391-426

Pavlova A, Rigobon R, Asset Prices and Exchange Rates. *Review of Financial Studies* 2007;20; 1139-1180

Pindyck RS, Rotemberg JJ, The Comovement of Stock Prices. *Quarterly Journal of Economics* 1993;108; 1073-1104

Pindyck RS, Rotemberg JJ, The Excess Co-Movement of Commodity Prices. *Economic Journal* 1990;100; 1173-1189

Pound J, Zeckhauser R, Clearly Heard on the Street: The Effect of Takeover Rumors on Stock Prices. *Journal of Business* 1990;63; 291-308

Qiu L, Welch I, Investor Sentiment Measures. National Bureau of Economic Research, 2006

Robin MG, Nathan S, Trading Patterns and Excess Comovement of Stock Returns. *Financial Analysts Journal* 2007;63; 69-81

Ross SA, Information and Volatility: The No-Arbitrage Martingale Approach to Timing and Resolution Irrelevancy. *Journal of Finance* 1989;44; 1-17

Rothschild M, Searching for the Lowest Price When the Distribution of Prices Is Unknown. *Journal of Political Economy* 1974;82; 689-711

Sabherwal S, Sarkar SK, Zhang Y, Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News.

Journal of Business Finance & Accounting 2011;38; 1209-1237

Seasholes MS, Zhu N, Individual Investors and Local Bias. *Journal of Finance* 2010;65; 1987-2010

Shen D, Zhang W, Xiong X, Li X, Zhang Y, Trading and Non-trading Period Internet Information Flow and Intraday Return Volatility. *Physica A: Statistical Mechanics and its Applications* 2016;451; 519-524

Shenoy C, Zhang YJ, Order Imbalance and Stock Returns: Evidence from China. *Quarterly Review of Economics and Finance* 2007;47; 637-650

Shiller RJ, Comovements in Stock Prices and Comovements in Dividends. *Journal of Finance* 1989;44; 719-729

Shiller RJ, Do Stock Prices Move too Much to be Justified by Subsequent Changes in Dividends? *American Economic Review* 1981;71; 421-436

Siganos A, Vagenas-Nanos E, Verwijmeren P, Facebook's Daily Sentiment and International Stock Markets. *Journal of Economic Behavior & Organization* 2014;107, Part B; 730-743

Simon HA, A Behavioral Model of Rational Choice. *Quarterly Journal of Economics* 1955;69; 99-118

Smirlock M, Starks L, An Empirical Analysis of the Stock Price-Volume Relationship. *Journal of Banking & Finance* 1988;12; 31-41

Solomon DH, Selective Publicity and Stock Prices. *Journal of Finance* 2012;67; 599-638

Solomon DH, Soltes E, Sosyura D, Winners in the Spotlight: Media Coverage of Fund Holdings as a Driver of Flows. *Journal of Financial Economics* 2014;113; 53-72

Solt ME, Statman M, How Useful Is the Sentiment Index? *Financial Analysts Journal* 1988;44; 45-55

Sprenger TO, Tumasjan A, Sandner PG, Welpe IM, Tweets and Trades: the Information Content of Stock Microblogs. *European Financial Management* 2014;20; 926-957

Stigler GJ, The Economics of Information. *Journal of Political Economy* 1961;69; 213-225

Stoll HR, Whaley RE, The Dynamics of Stock Index and Stock Index Futures Returns. *Journal of Financial and Quantitative Analysis* 1990;25; 441-468

Tauchen GE, Pitts M, The Price Variability-Volume Relationship on Speculative Markets. *Econometrica* 1983;51; 485-505

Tesar LL, Werner IM, Home bias and High Turnover. *Journal of International*

Money and Finance 1995;14; 467-492

Tetlock PC, All the News That's Fit to Reprint: Do Investors React to Stale Information? *Review of Financial Studies* 2011;24; 1481-1512

Tetlock PC, Does Public Financial News Resolve Asymmetric Information? *Review of Financial Studies* 2010;23; 3520-3557

Tetlock PC, Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance* 2007;62; 1139-1168

Tetlock PC, Saar-Tsechansky M, Macskassy S, More than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance* 2008;63; 1437-1467

Tumarkin R, Whitelaw RF, News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal* 2001;57; 41-51

Varian HR, Divergence of Opinion in Complete Markets: A Note. *Journal of Finance* 1985;40; 309-317

Veldkamp LL, Information Markets and the Comovement of Asset Prices. *Review of Economic Studies* 2006;73; 823-845

Vlastakis N, Markellos RN, Information Demand and Dtock Market Volatility. *Journal of Banking & Finance* 2012;36; 1808-1821

Vozlyublennaia N, Investor Attention, Index Performance, and Return Predictability. *Journal of Banking & Finance* 2014;41; 17-35

Wagner N, Marsh TA, Surprise Volume and Heteroskedasticity in Equity Market Returns. *Quantitative Finance* 2005;5; 153-168

Wang J, A Model of Competitive Stock Trading Volume. *Journal of Political Economy* 1994;102; 127-168

Wysocki P, Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. Working Paper, University of Michigan Business School, 1998

Xu Y, Malkiel B, Investigating the Behavior of Idiosyncratic Volatility. *Journal of Business* 2003;76; 613-645

Yuan K, Asymmetric Price Movements and Borrowing Constraints: A Rational Expectations Equilibrium Model of Crises, Contagion, and Confusion. *Journal of Finance* 2005;60; 379-411

Zhang W, Li X, Shen D, Teglio A, R^2 and Idiosyncratic Volatility: Which Captures the Firm-specific Return Variation? *Economic Modelling* 2016;55; 298-304

Zhang W, Shen D, Zhang Y, Xiong X, Open Source Information, Investor Attention, and Asset Pricing. *Economic Modelling* 2013;33; 613-619

Zhang X, Fuehres H, Gloor PA, Predicting Stock Market Indicators through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences* 2011;26; 55-62

Zhang XF, Information Uncertainty and Stock Returns. *Journal of Finance* 2006;61; 105-137

Zhang Y, Feng L, Jin X, Shen D, Xiong X, Zhang W, Internet Information Arrival and Volatility of SME PRICE INDEX. *Physica A: Statistical Mechanics and its Applications* 2014;399; 70-74

Zhang Y, Swanson PE, Prombutr W, Measuring Effects On Stock Returns Of Sentiment Indexes Created From Stock Message Boards. *Journal of Financial Research* 2012;35; 79-114