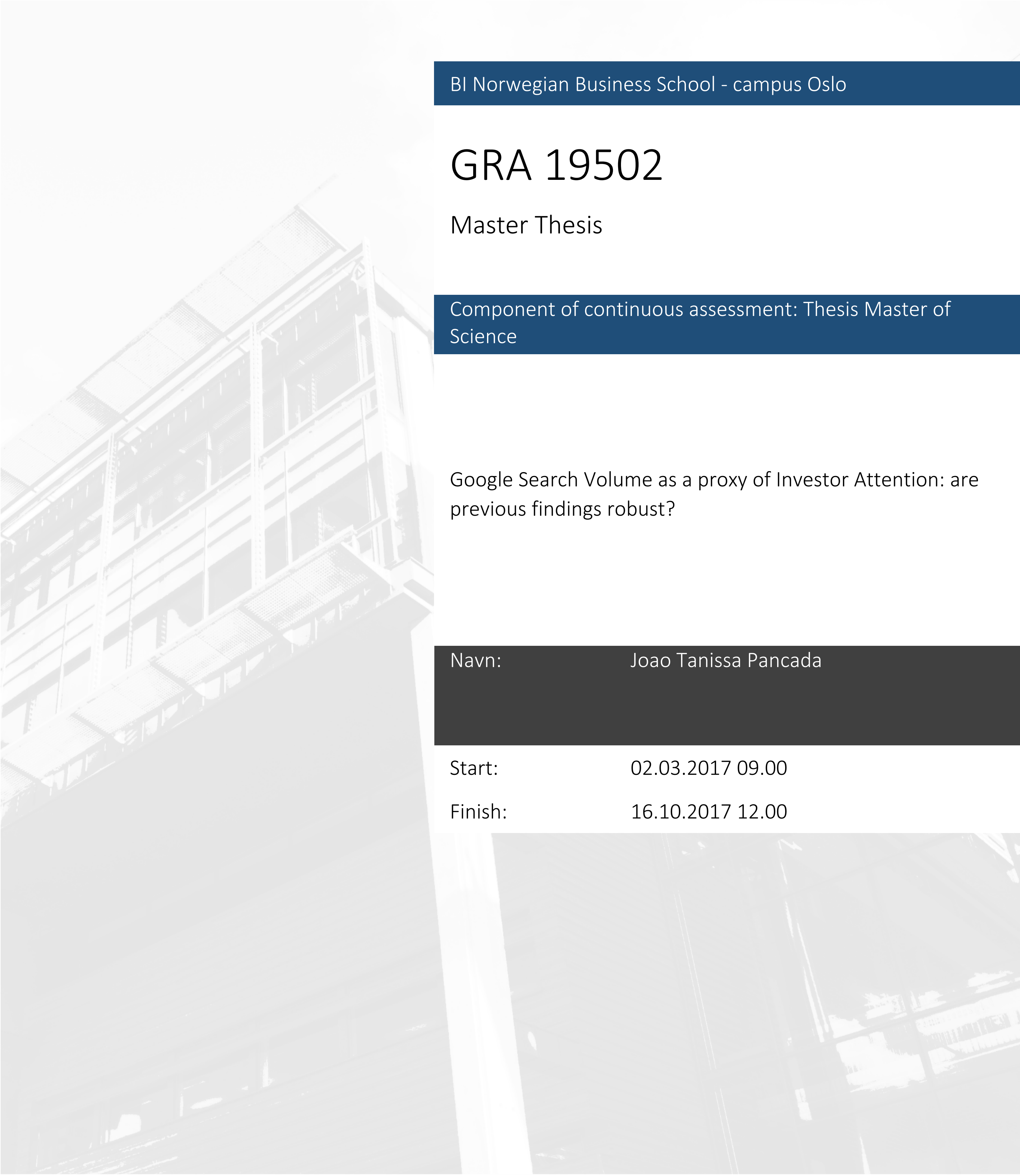


BI Norwegian Business School - campus Oslo

GRA 19502

Master Thesis

Component of continuous assessment: Thesis Master of Science

Google Search Volume as a proxy of Investor Attention: are previous findings robust?

Navn:	Joao Tanissa Pancada
Start:	02.03.2017 09.00
Finish:	16.10.2017 12.00

João Tanissa de Carvalho Pancada

Master Thesis

Google Search Volume as a proxy of Investor Attention: are previous findings robust?

Hand-in date:
16.10.2017

Campus:
BI Oslo

Examination code and name:
GRA 19502 Master Thesis

Programme:
Master of Science in Financial Economics

“This thesis is a part of the MSc programme at BI Norwegian Business School. The school takes no responsibility for the methods used, results found and conclusions drawn."

Abstract

I assess the robustness to different periods and panel models of several findings in the literature that uses Google search volume as an investor attention proxy. With all S&P 500 stocks between 2004 and 2016, I confirm that weekly search volume is persistent and increases are associated with high share turnover as well as earnings announcements. When the CCEMG estimator of Pesaran (2006) is used, I find evidence against Da et al. (2011) but consistent with Barber and Odean (2008). Even for large stocks, surges in people's attention predict positive abnormal returns one week ahead, which reverse within one year. I conclude the literature should adopt panel estimators more robust to the presence of both firm and time effects.

Acknowledgments

I would like to thank several people that contributed directly or indirectly to the success of this thesis. First of all, I am glad that Professor Bruno Gerard accepted to be my supervisor. His prompt feedback and guidance were very helpful. I would have gone much further with his advice had I started earlier. Secondly, the unconditional support of my family throughout the process was crucial to finish the thesis on time. A special thanks goes to my brother, whose programming skills were vital to obtain the key variable of this thesis (Google search volume). Thirdly, I appreciate the comments and suggestions of many Statalist users to my econometric and Stata related doubts. In addition, I am thankful to several researchers whose Stata code I extensively used. Finally, I would like to thank Google Inc. for making available a measure of their historical search volume, which has sparked substantial research in many different fields.

Table of Contents

Index of Figures and Tables

1. Introduction

1.1 Motivation

We are currently living in the information age. The Internet, and subsequent innovations, have allowed a myriad of news and data to spread very fast and at a low cost across the world (Rubin and Rubin, 2010). Investors should not miss this constant information flow, since financial markets are known to be sensitive to new information (French and Roll, 1986). In the limit, as traditional asset pricing models assume, security prices reflect all available information, with new one being immediately incorporated once it is made public (Merton, 1987).

However, there is only so much information people can process each day. It takes time to obtain and then decide how to act on new data. Consequently, investors have to be selective and traditional models fail to consider these limitations (Peng and Xiong, 2006). Inattention can significantly impact asset prices and compromise market efficiency as the adjustment to new information takes longer than expected. For instance, Huberman and Regev (2001) take the extreme example of EntreMed, a US biotech company whose stock price soared 330% in one day. This rise followed a news piece from the New York Times on a research breakthrough that had been featured on Nature magazine 6 months before.

Measuring investor attention (or distraction) is no easy task. It is difficult to quantify the effort people put into, for example, collecting data or reading the news on a given company. Simply asking people is not a feasible strategy. Therefore, rather than looking at the demand for information, the literature has traditionally focused on indirect proxies that fit in one of two categories (Vlastakis and Markellos, 2012): a) supply side of information – e.g., Lou (2014) for advertising expenditures and Yuan (2015) for front-page articles; b) market data – e.g., Barber and Odean (2008) for abnormal turnover and extreme returns.

Nevertheless, as some (proprietary) databases are made public, there is a relatively recent trend that uses demand proxies. Starting with internet blogs (e.g., Antweiler and Frank, 2004), researchers have been trying to capture people's attention through their online behaviour. Once their actions or interests are *revealed*, an analogy to the concept of revealed preferences in microeconomics, we are one step closer towards measuring people's effort.

1.2 Google Search Volume as a demand proxy

One way of assessing the interests of many individuals at a point in time is to know what they search on Google. Ginsberg et al. (2009) are among the first to document the usefulness of Google's Search Volume Index (hereafter SVI) by showing that it can detect influenza outbreaks. Shortly after, Da et al. (2011) introduced SVI as a measure of aggregate investor attention in the US. They found it captures well retail activity and that high values help predict abnormal returns.

SVI is available to the general public through Google Trends since May 2006, but the database starts in 2004 (https://trends.google.com). When we type a set of words or 'search terms' into Google Trends, the platform returns a relative measure of their search volume history. To illustrate, monthly SVI is plotted for the term 'VRSK' in figure 1. It is based on searches from Google web users in the US during January 2004 to December 2016. 'VRSK' stands for the stock ticker of Verisk Analytics, a US data analytics company. Each monthly observation represents the ratio of search volume on 'VRSK' over the search volume on all terms in the US that month, scaled by the maximum monthly ratio from 2004 to 2016[1]. We can see there is practically no interest on that term prior to October 2009 but, suddenly, there is a search pattern that fluctuates over time. This is not a coincidence because the peak on that date represents the company's IPO month. Therefore, the word 'VRSK' gained a distinctive meaning after October 2009 and people started typing it regularly. This is clear evidence that SVI can capture people's attention.

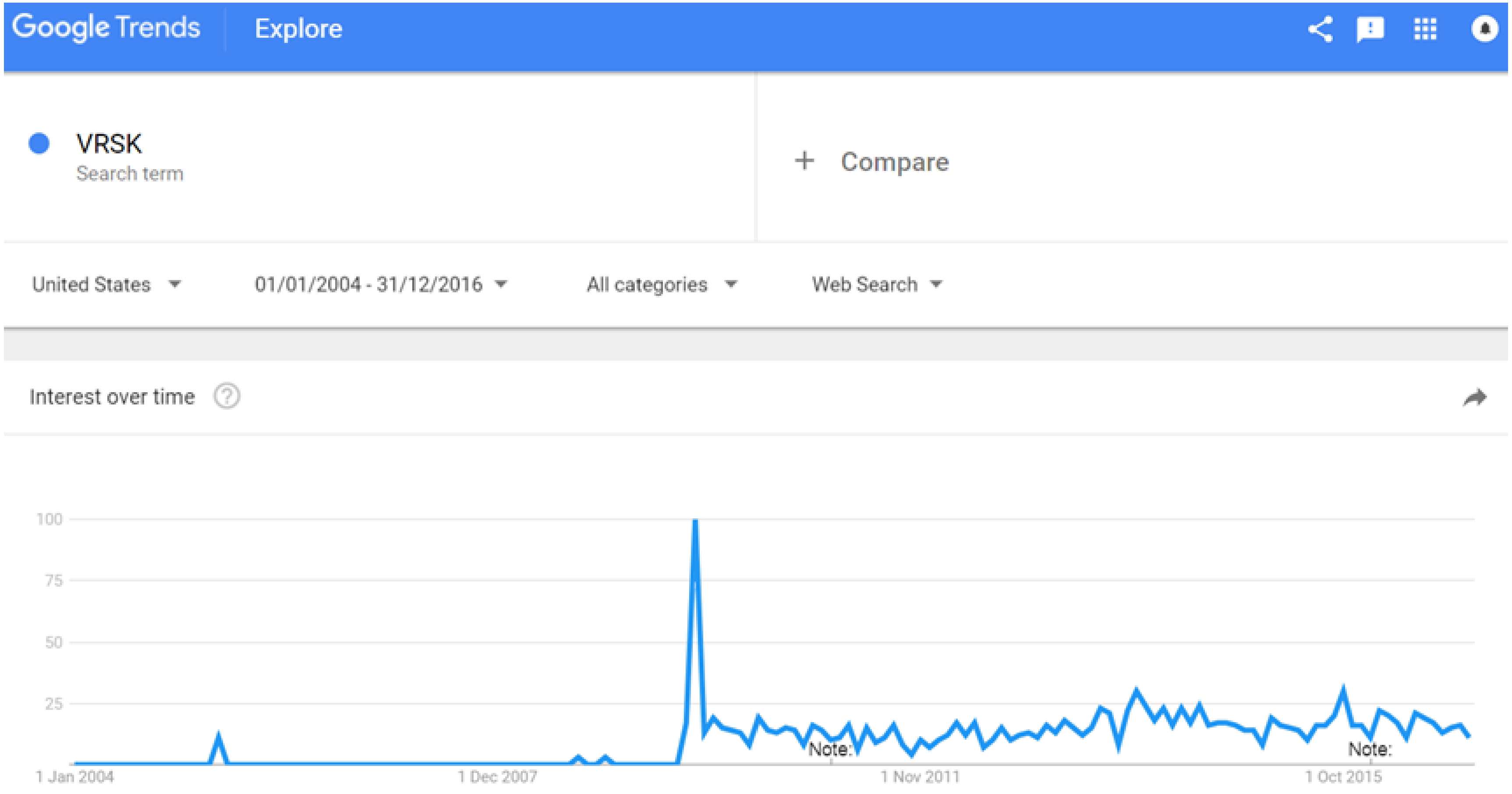
With the example of figure 1 in mind, there are two reasons to believe that aggregate search data from Google can be a good attention proxy. First, search frequency directly reveals whether people are interested in certain events (or companies) over time. This requires, however, that we specify the search term that (potential) investors often use. As discussed in section 4.2, I use each stock's ticker symbol. Second, Google continues to dominate the US market for web search engines, with a share fluctuating between 59% and 69% from 2008 to 2016[2]. Therefore, it is expected that Google searches represent the general search behaviour in the US.

[1] See section 1 of appendix A for a more detailed explanation of this two-step process.

[2] Source: Statista - https://www.statista.com/statistics/267161/market-share-of-search-engines-in-the-united-states/ (retrieved on September 10th, 2017)

Figure 1. Illustration of Google Trends Ouput

This figure shows the top half of the output from Google Trends when typing the term ‘VRSK’ on September 10th 2017. From January 2004 to December 2016, the chart plots monthly SVI for Google web searches on that term in the US. The series ranges from 0 to 100. Each monthly observation represents the ratio of search volume on ‘VRSK’ over the search volume on all terms in the US that month, scaled by the maximum monthly ratio from 2004 to 2016. See section 1 of appendix A for a more detailed explanation of how to interpret SVI. The notes that appear above the horizontal axis simply indicate two (out of many other) dates when the system was improved.



VRSK
Search term

United States
01/01/2004 - 31/12/2016
All categories
Web Search

1.3 Research problem and thesis contribution

The aim of this thesis is to examine in detail the consistency of several findings in the SVI literature to more recent time periods and robust panel estimators. This is done by testing the following three hypothesis:

1. Is SVI related but different from existing investor attention proxies?
   Hypothesis 1: *SVI is similar to existing proxies of investor attention.*

2. Does SVI mainly capture the attention of retail investors?
   Hypothesis 2: *SVI does not capture the search behaviour of retail investors.*

3. Is SVI in favour of the findings of Barber and Odean (2008)?
   Hypothesis 3: *SVI does not predict positive returns in the short-term for large stocks, which reverse within a year.*

Using all stocks that were ever part of the S&P 500 index between 2004 and 2016, the contribution of this thesis is twofold. I am the first testing hypothesis 1 under a new method of computing SVI and with 13 rather than 4 to 5 years of weekly data as in Da et al. (2011) and Drake et al. (2012). The same essentially applies to hypothesis 3 because Cziraki et al. (2017) use a lower frequency (i.e. monthly data) and do not entirely follow the method of Da et al. (2011).

The second contribution concerns a more rigorous econometric treatment of the underlying panel data. Recent research has shown that parameter consistency of standard methods can be significantly affected in the presence of cross-sectional dependence (e.g., Hjalmarsson, 2010; Graevenitz et al., 2016; Westerlund et al., 2017). The literature on SVI, and on investor attention in general, has greatly overlooked this issue. When testing hypothesis 1, researchers have simply used pooled OLS with time dummies and adjusted standard errors. Even Fama-MacBeth (1973) regressions, used by Da et al. (2011) and Cziraki et al. (2017) in hypothesis 3, may be problematic (e.g., Campello et al., 2013). I argue we should rather apply the CCEMG model of Pesaran (2006) because of the high cross-sectional dependence that is expected in financial data (Philipps and Moon, 1999; Pesaran, 2007). Finally, I suspect that the persistency of Google search, as seen in the VAR model of Da et al. (2011), might also be important to take into account, even in static models.

Overall, results indicate that the three hypothesis are rejected and it is very important to consider the structure of the data as well as different time periods.

In hypothesis 1, I do find that SVI is capturing a quite unique phenomenon as in Da et al. (2011) and Drake et al. (2012). However, in more recent periods, the explanatory power of existing attention proxies can be more than twice of what was previously assumed. Furthermore, only two proxies keep their anticipated relation with Google search over time. There is strong evidence that SVI increases when trading volume is high and when a company announces its quarterly earnings. This is also very good news about the usefulness of SVI as an attention proxy.

I also find that Google search is a persistent variable up to its 4th lag. Even though these lags appear to be correlated with the regressors, their omission does not materially influence the results. Non-stationarity is also not a major concern, but I find that it is essential to consider firm effects and cross-sectional dependence. In hypothesis 1, variables such as analyst coverage, the bid-ask spread, and absolute abnormal returns become statistically insignificant when more robust models are used. Most importantly, in hypothesis 3, the main conclusions dramatically change with the CCEMG estimator. I show that current week abnormal SVI predicts positive abnormal returns next week, which reverse within one year. This is robust to different time periods and ways of computing abnormal SVI. When 'noisy' tickers are removed, the effect is even stronger. The opposite conclusion, the one documented in Da et al. (2011), is found when using Fama-MacBeth (1973) regressions. Therefore, I support the predictions of Barber and Odean (2008) for large stocks.

Nevertheless, I confirm that SVI's effect is stronger among smaller stocks, as in Da et al. (2011), but not for those that retail investors are more likely to trade. In hypothesis 2, the change in idiosyncratic volatility is the only retail activity proxy that has a consistent positive relation with abnormal SVI over time. However, it is not robust against aggregate abnormal returns, which is expected given its indirect nature. The conclusion might change with a proprietary database on retail activity such as NYSE ReTrac, but that is left for future research.

This thesis proceeds as follows. Section 2 provides an overview of the background literature. Section 3 explains the methodology and section 4 describes the data. Section 5 explains each of the three hypothesis in detail and examines the results. Section 6 concludes and discusses where future research can be useful.

2. Literature Review

This section starts with the fact that investor attention is limited and it can affect financial markets. Traditional and demand proxies of attention are then outlined. This section finishes with an emphasis on the demand proxy used in this thesis, which is Google's Search Volume Index.

2.1 *Investor attention and its impact on financial markets*

The tradition in financial economics has been to assume that people act rationally in their investment decisions so that their utility is always maximized (Barber and Odean, 2011). This requires that investors keep analyzing the risk-return trade-off of all assets until they find their optimal portfolio, given their level of risk aversion. Under the Capital Asset Pricing Model of Sharpe (1964) and Lintner (1965), arguably the most famous model in asset pricing, all investors end up holding the same well-diversified portfolio of risky assets, which is the market portfolio. When new information is released to the public, and there are no frictions, prices react immediately as investors revise their risk-return expectations.

In reality, the adjustment may take time because frictions abound. One of the most pervasive is intrinsic to human nature, namely our limited cognitive resources (Kahneman, 1973). There are thousands of investment opportunities available and new information is being constantly produced about them. However, people do not have the time or the capacity to process and incorporate all that information in their investment decisions (e.g., Hirshleifer and Teoh, 2003; Peng, 2005).

Merton (1987) is the first to formally recognize the importance of investor attention in financial markets and to theorize its implications. His main contribution is that a firm's market value is increasing with investor recognition but expected returns are decreasing. The impact of a larger investor base has been validated empirically by Kadlec and McConnell (1994) and Chen et al. (2004), among others.

In a growing literature, other predictions have been proposed. For example, Peng and Xiong (2006) and Mondria (2010) postulate that limited attention leads people to act more on general signals (market or sector) rather than firm-specific, which increases stock co-movement. Hou et al. (2009) argue that the post-earnings announcement drift and the momentum anomaly are due to, respectively, investor underreaction and overreaction to information. Nieuwerburgh and Veldkamp (2010) suggest that when investors have to choose which information to collect before investing, their portfolio allocation is sub-optimally diversified.

Since it is clear that investor attention is limited, the next question is: where do investors focus their attention? The literature seems to be unanimous in that investors' attention is directed towards familiar or 'attention-grabbing' stocks. Seasholes and Wu (2007) find that investors are induced to buying stocks they did not own when prices hit upper limits on the Shanghai Stock Exchange. Likewise, Barber and Odean (2008) show that retail investors are net buyers of stocks that have recently caught their attention after periods of abnormal returns, turnover and news coverage. Corwin and Coughenour (2008) study NYSE's specialists and argue that these market makers dedicate more effort to their most active stocks, leading to less liquidity in the other securities they trade. Others have argued that limited attention leads investors to prefer the stock of domestic companies and of their employer, which can help explain the home bias puzzle (e.g., Coval and Moskowitz, 1999; Huberman, 2001).

The common issue among these researchers is that they can only provide indirect evidence of limited attention, since this is something difficult to measure or observe (Corwin and Coughenour, 2008). Based on Vlastakis and Markellos (2012), proxies of attention can be divided into three distinct groups: market data, supply of information, and information demand. The former two have been the tradition in the literature due to data limitations and are discussed in the next section. The latter, where SVI is included, has seen tremendous growth in recent years and is explored in sections 2.3 and 2.4.

2.2 Traditional proxies of investor attention

There is a vast literature of indirect proxies measuring investor attention. On the supply side of information, the two most popular measures have been news coverage and advertising expenditures. The reason is that both are associated with higher company visibility to the public, and hence to investors. The results are in general consistent with the predictions of Merton (1987).

The study of Thompson et al. (1987) is one of the first to document that both firm returns and their volatility change when firm-specific news are published on the Wall Street Journal. More recently, Ryan and Taffler (2004) present evidence that corporate news are important drivers of price changes and trading volume in the London Stock Exchange. Fang and Peress (2009) find in the cross-section that companies with no media coverage earn a return premium even after controlling for 5 common risk factors. Their results are stronger among smaller stocks and for companies with low analyst coverage, high idiosyncratic volatility, and high retail investor ownership. The importance of the media remains whether researchers take into account the content of news articles (e.g., Mitchel and Mulherin, 1994; Tetlock, 2007) or focus on the impact of market-wide news (e.g., Berry and Howe, 1994).

Grullon et al. (2004) and Frieder and Subrahmanyam (2005) show that more product advertising increases the number of individual and institutional investors. Chemmanur and Yan (2009) find that firms increase their levels of advertising when they aim at issuing equity. This spillover to the financial markets is found by Lou (2014) to have a temporary positive effect on returns due to retail investors. In turn, the researcher argues this is opportunistically considered by firm executives to either sell their company stock, issue new equity or finance acquisitions with stock.

In terms of market proxies, examples abound. Absolute and extreme stock returns have been used, respectively, by Corwin and Coughenour (2008) and Barber and Odean (2008). The drivers of high returns may catch investor attention as well as high returns themselves. Periods of high trading volume are also likely to be associated with higher investor interest (e.g., Gervais et al., 2001; Hou et al., 2009). When several firms announce their earnings at the same time, investors might not be able to follow all of them, which results in underreaction to the new information (e.g., Hirshleifer et al., 2009; Drake et al., 2012). Analyst coverage and stock liquidity have also been used (e.g., Da et al., 2011; Drake et al., 2012). The reason is that a company with more analysts and a more liquid stock is likely to have less information asymmetry. Therefore, investors do not need to spend as much effort processing company related information (Arbel, 1985; Leuz and Verrecchia, 2000).

2.3 *Demand proxies of investor attention*

The relevance of information demand was already well recognized in the theoretical literature (e.g., Kihlstrom, 1974; Grossman and Stiglitz, 1980). However, it was only possible to study this ‘new class’ of proxies with the advent of the internet and the subsequent release of proprietary data. Demand proxies are currently obtained from two different sources: online posting activity (e.g., internet forums, social media) and search volume (e.g., Google, Baidu).

Using more than 1.5 million posts from Raging Bull and Yahoo! Finance, Antweiler and Frank (2004) find intra-day and one day predictability for stock trading volume and volatility but not for returns. Das and Chen (2007) also do not find evidence for returns using Yahoo! messages. Later, Rubin and Rubin (2010) turned to the editing frequency of Wikipedia articles. They present evidence in favour of a positive relation where higher editing is associated with lower earnings surprises and less dispersion in analysts' forecasts.

With the emergence of social media, researchers started developing larger scale algorithms[3]. Bollen et al. (2011) show that daily mood changes on Twitter posts are correlated with changes in the DJIA Index over the next days. Siganos et al. (2014) use Facebook's sentiment index and find evidence that there is a positive contemporaneous relation between sentiment and stock market returns, which then reverses in the following weeks. Chen et al. (2014) analyse an exclusively investment-related website (Seeking Alpha). Using more companies and a longer period than previous studies, they show that people's opinions help predict future stock returns and anticipate earnings surprises.

The importance of studying investors' search patterns is motivated by the marketing literature. It is well-known that consumers spend (more or less) time searching for alternatives and understanding product features before making a purchase (e.g., Beatty and Smith, 1987). Thus, investor search can be considered a leading indicator of their investment decisions (Choi and Varian, 2009).

Mondria et al. (2010) started the finance literature on web search engines using AOL and link the US home bias to investor attention. Preis et al. (2010) and Da et al. (2010, 2011) followed with Google search volume, as discussed in the next section. Shi et al. (2012) use data from Baidu search engine and reach similar results as Barber and Odean (2008). Drake et al. (2015) use SEC's EDGAR search traffic and find that information demand is positively related to several corporate events. Lawrence et al. (2016) use Yahoo! Finance and show companies in the highest quintile of abnormal search outperform those at the bottom, an effect that has no reversal even after 1 year. The most recent contribution is that of Ben-Rephael et al. (2017), who for the first time directly study the behaviour of institutional investors through their searches on Bloomberg Terminal. They find that only retail investors are distracted by many news on the same day and firm-specific news, rather than market-wide information as suggested by Peng and Xiong (2006), are the relevant firm drivers for institutional investors.

[3] Practitioners have incorporated such algorithms in their trading strategies using Twitter.
Source: Business Insider - http://www.businessinsider.com/sell-signal-hedge-fund-unveils-secret-weapon-to-beat-the-market-twitter-2011-5 (retrieved on September 20th, 2017).

2.4 *Google Search Volume*

Google Trends was launched on May 2006 and Google Insights on August 2008[4]. Initial studies have used Google search to anticipate widespread diseases (e.g., Ginsberg et al., 2009), forecast unemployment claims (e.g., Askitas and Zimmermann, 2009) and private consumption (e.g., Choi and Varian, 2009).

The finance literature on SVI started with Preis et al. (2010), who find a positive correlation between changes in web search and trading volume for S&P 500 stocks. Several other papers document the same positive relation for individual companies and stock indices as well as with liquidity and volatility (e.g., Bank et al., 2011; Vlastakis and Markellos, 2012; Latoeiro et al., 2013; Andrei and Hasler, 2015). Vlastakis and Markellos (2012) are able to empirically show for the first time that information demand, given by Google searches, increases with the variance risk premium, which is proxying for the level of investor risk aversion.

Nevertheless, the most influential findings are those found in a series of papers from Da, Engelberg, and Gao, who use a comprehensive dataset of Russell 3000 stocks. Da et al. (2010) show that the momentum effect is stronger among the most searched stocks, especially for those in the winner’s portfolio. They argue that occurs because SVI is capturing the attention of retail investors. This hypothesis is successfully tested in Da et al. (2011) and Gwilym et al. (2016). In addition, Da et al. (2011) and Drake et al. (2012) show that existing proxies of attention explain little variation in Google search, with reported regression R^2 below 4%.

In terms of stock returns, several papers predict a positive relation with SVI in the short-term (e.g., Da et al., 2011; Bank et al., 2011; Joseph et al., 2011; Gwilym et al., 2016), including for stock markets (e.g., Da et al., 2014; Vozlyublennaia, 2014). The effect starts reversing after four or five weeks, which is consistent with investor sentiment predictions described in section 2.3. In fact, Da et al. (2014) builds a new investment sentiment index based on Google searches. However, this relation with returns seems to be only present among small stocks (Da et al., 2011; Bank et al., 2011; Cziraki et al., 2017). Finally, researchers have also shown that SVI increases substantially in the previous and at the same week of important corporate events such as IPOs, earnings announcements, and mergers announcements (Da et al., 2011; Drake et al., 2012; Siganos, 2013).

[4] They were merged in 2012 and the method of Insights is now the one of Google Trends. It used to be just the number of searches for a given term scaled by its time-series average (Da et al., 2011).

3. Methodology

The empirical analysis in this thesis is conducted through panel data regressions. Panel analysis is to a large extent a combination of both time-series and cross-sectional analysis (Wooldridge, 2009). The four sections that follow explain step by step the reasoning behind the statistical methods used. I leave to the references their underlying mathematical complexity and, when applicable, to appendix B their corresponding user-written Stata commands.

3.1 Panel regressions - background

In panel data we have two dimensions, firm and time, rather than only firm or time as in, respectively, cross-sectional and time-series analysis. Its great advantage is the ability to more easily solve the omitted variable bias problem (Hsiao, 2014). If the problem does not exist, we may find more efficient estimators than OLS.

When the data on all firms begins and ends at the same dates with no missing observations in between, the panel is said to be balanced. This is not, however, the case here because some firms: a) go bankrupt, b) are delisted after a take-over or merger, 3) only become listed after the sample begins. Having an unbalanced sample is not a major problem by itself because all tests and regressions conducted in this thesis can accommodate that. Based on Cameron and Trivedi (2005), an issue arises if leaving the sample, which is called attrition, is correlated with the regression residuals, i.e., attrition is non-random. This may occur in hypothesis 3 as firms leave the index for reasons related to the dependent variable (stock returns) such as going bankrupt. Section 4.1 outlines a few solutions.

The absolute (and relative) size of both dimensions determine which regression models are used because the asymptotic theory differs. In terms of absolute size, if N is large and T is small, we have the typical case in microeconomics. However, when T becomes large, the time-series properties of the data have to be considered, namely testing for stationarity (Wooldridge, 2009). There is no formal definition of what constitutes large N and T, but several papers that run Monte-Carlo simulations take values between 100 and 200 (e.g., Chudik and Pesaran, 2013; Everaert and Groote, 2016). Therefore, I assume large N and T because my sample (for most regressions) has $N = 739$ and $\overline{T} = 206$. This brings numerous advantages to the traditional fixed T, large N framework. Among others, it allows richer models that can control for cross-sectional dependence but at the cost of dealing with the aforementioned time-series dependence (Baltagi, 2009).

Following Cameron and Trivedi (2005), the most general panel regression is:

$$y_{i,t} = a_{i,t} + x_{i,t} * \beta_{i,t} + u_{i,t}, \quad \mathrm{i} = 1, \dots, \mathrm{N}, \quad \mathrm{t} = 1, \dots., \mathrm{T}, \tag{1}$$

where $y_{i,t}$ is the dependent variable, $x_{i,t}$ is the vector of regressors, $u_{i,t}$ is the error term, i refers to the firm dimension and t to the time dimension. Equation (1) cannot be estimated because there are far too many parameters as both the intercept ($a_{i,t}$) and the slope coefficients ($\beta_{i,t}$) vary by firm and over time.

The case of non-constant slopes is not considered in this thesis because: a) both dimensions are large, hence not feasible to estimate all coefficients; b) the interest on SVI has to do with its applicability in general rather than on any particular firm or time period[5]. Nevertheless, in most finance applications, the intercept should not be constant (Petersen, 2009). Therefore, alternatives to the Pooled OLS estimator, which requires constant slopes and intercepts, need to be found.

3.2 Panel regressions with varying intercepts

When the intercept changes by firm and/or over time, but that is not taken into account, the panel model takes the following general form (Bai, 2009):

$$y_{i,t} = \alpha + x_{i,t} * \beta + u_{i,t} \tag{2}$$

where $u_{i,t} = \theta_i F_t + v_{i,t}$. $v_{i,t}$ are individual specific idiosyncratic errors that are assumed iid and independent of the regressors. θ_i is a vector of factor loadings and F_t stands for a vector of unobserved common factors, which can be correlated with the regressors and over time. These last two terms together capture two effects. On the one hand, they consider the impact of time-varying variables that are constant across all firms (called time fixed effects) or heterogeneous such as macroeconomic shocks (Petersen, 2009). On the other hand, they capture firm-specific characteristics that are constant over time (called firm fixed effects) or time-varying such as the accumulation of human capital (Ahn et al., 2013). Time and firm effects have different implications for estimator consistency and valid statistical inference.

In terms of consistency, if these unobserved effects impact the dependent variable but are correlated with some regressors, we have an omitted variable bias. This is likely to occur when testing all hypothesis. For example, time effects such as recessions impact the returns of all firms in different ways (dependent variable of hypothesis 3) but also their advertising expenditures and sales (regressor). Those

[5] As far as I know, only Vlastakis and Markellos (2012) in the SVI literature estimate firm-specific regressions, but their sample only has 30 firms.

shocks also impact investor attention, leading more people to search certain companies like Bear Stearns in 2008 (dependent variable in hypothesis 1 and 2).

In terms of inference, the standard errors of the coefficients are underestimated when common factors are not considered, which may give rise to misleading statistical significance (Gow et al., 2010). Thompson (2010) shows that not controlling for time effects leads to cross-sectional error dependence, meaning that the residuals of any two firms are contemporaneously correlated. If these time effects are persistent, then the residuals of different firms at different points in time are also correlated. Finally, not controlling for firm effects leads to serial correlation. However, Thompson (2010) also shows that time and firm effects are only a concern when regressors are themselves, respectively, cross-sectionally correlated and autocorrelated.

The two previous paragraphs motivate the use of regression models that control for both time and firm effects. Traditionally, researches have used the two-way fixed effects model, which is a special case of equation (2) (Pesaran, 2006):

$$y_{i,t} = \alpha + x_{i,t} * \beta + u_{i,t} \tag{3}$$

where $u_{i,t} = \alpha_i + \gamma_t + v_{i,t}$. α_i stands for firm fixed effects whereas γ_t are time fixed effects. The former is controlled for by either (quasi) time demeaning the regression or with firm dummies. The latter requires either cross-sectionally demeaning the regression or time dummies.

As argued by Petersen (2009), when the time dimension is smaller than the number of firms, time dummies should be used to control for time fixed effects. Then, we either fully time demean the regression with the fixed effects model or partially do so using the random effects model. The choice depends on whether firm fixed effects cause an omitted variable bias, which can be tested with the robust form of the Hausman test to heteroskedasticity (Wooldridge, 2009). Under the null hypothesis, there is no bias, and the random effects model is more efficient. However, if time or firm effects are not fixed, but they are uncorrelated with the regressors, it is enough to further adjust the standard errors. With large N and T, firm effects are fully corrected when clustering by firm whereas time effects require clustering by time or Fama-Macbeth (1973) regressions (Petersen, 2009).

In section 5, I initially follow the methodology used in the previous literature. For hypothesis 1, Da et al. (2011) and Drake et al. (2012) use Pooled OLS with week dummies and robust standard errors clustered at the firm level to account for, respectively, time fixed effects and uncorrelated firm effects. In hypothesis 3, Da et al. (2011) use cross-sectionally demeaned Fama-MacBeth (1973) regressions and the Newey-West (1987) adjustment with 8 lags. Their purpose is to account for, respectively, time fixed effects, uncorrelated (firm-varying) time effects and uncorrelated firm effects.

However, I believe we should not ignore the issue of inconsistent estimates arising from the correlation between the regressors and the residuals. As illustrated with the recession example, time effects for these sort of regressions are not constant and they may be correlated with the regressors. In addition, it is not clear whether firm effects are correlated with the regressors, and so it is prudent to verify it with the Hausman test. Finally, as explained in section 3.4, it may be a good idea to include the lagged values of SVI in hypothesis 1 to control for their potential correlation with the regressors.

Fortunately, with large N and T, all these potential problems can be mitigated by estimating a ‘factor-augmented regression’ that assumes the more complex case of equation (2). Therefore, in a second stage, section 5 shows how the results of hypothesis 1 and 3 change with a new estimation method.

3.3 Factor-augmented regressions

There are two ways of augmenting a panel regression and control for unobserved common factors: directly estimate them through principal components (PC); filter out their effect by including the cross-sectional averages of the dependent and independent variables as additional regressors. The former method is known as the interactive fixed effects or PC estimator of Bai (2009) whereas the latter is the common correlation effects (CCE) estimator of Pesaran (2006). The CCE estimator is divided into two, the CCE Mean Group (CCEMG) and the CCE Pooled (CCEP) estimators. Under constant slopes, the former has been shown to be preferred by Pesaran (2006) and Chudik and Pesaran (2013).

The overall evidence for my sample favours the CCEMG over the PC estimator. As explained in the next few paragraphs, this preference holds whether factors are strong or non-strong, but there are two exceptions (Pesaran, 2015). If the number of strong factors is not fixed, i.e., they increase with N asymptotically, neither estimator works. However, when there is a fixed number of weak factors, both estimators are equally valid. As defined in Chudik et al. (2011), the existence of strong factors causes strong cross-sectional dependence (CD) whereas weak factors cause weak CD. This distinction is rather technical but the important thing to keep in mind is that only strong CD "*can pose real problems*" (Pesaran, 2015 p. 1091).

In the presence of a fixed number of strong factors, the PC and CCEMG estimators have an inherent bias. The coefficients are shown to be $\sqrt{\mathrm{NT}}$ consistent and asymptotic normality requires $T < N < T^2$ (Westerlund and Urbain, 2015). The last condition is not an issue here because $206 < 739 < 206^2$. However, a rank condition has to be met for the CCEMG estimator, namely that the number of factors is smaller or equal to the number of regressors plus one (Pesaran, 2006). Otherwise, the convergence rate is only $\sqrt{\mathrm{N}}$. Nonetheless, according to Chudik and Pesaran (2015), the most conservative study in terms of necessary (strong) fixed factors "*estimate as many as seven factors*" (p. 394). Since my regressions have more than 6 regressors, this rank condition should not be a great concern.

When the number of non-strong/weak factors rises with N, the PC approach can be severely biased as shown in Bai and Ng (2008) and Chudik et al. (2011). However, the CCEMG estimator "*remains consistent and asymptotically normal* (…) *under certain conditions on the loadings of the factor structure*" (Chudik et al., 2011 p. 47). Namely, we have to assume uncorrelated factor loadings (θ_i), a relatively strong assumption (Sarafidis and Wansbeek, 2012). Despite that, Reese and Westerlund (2015) suggest that when $N > T$, which is the case here, the CCEMG estimator is strictly preferred.

The PC estimator has three additional unique drawbacks. First, as shown by Westerlund and Urbain (2013), its effectiveness is much more sensitive to the assumed error structure. Second, tests are needed to determine how many factors should be estimated and "*this can introduce some degree of sampling uncertainty into the analysis*" (Chudik and Pesaran, 2013 p. 18). This is an unnecessary issue because there is no interest in the factors themselves. At last, even though factors can be persistent, they cannot be integrated of order one (Kapetanios et al., 2010). Thus, they have to be stationary, which in practice is not clear whether it holds.

Note that recently, Karabiyik et al. (2017) have argued that some derivations associated with the CCE estimator are incorrect and hence "*many statements in the* [CCE] *literature are actually yet to be proven*" (p. 62). Those researchers modify the existing proofs and reach the same results as Pesaran (2006) provided that both N and T are large, and N is 'sufficiently greater' than T. Given that in this thesis N is more than three times T, this last condition should also hold.

3.4 Dynamic panel regressions

We have a dynamic model when the lagged value of the dependent variable ($y_{i,t-1}$) is included as a regressor. In that case, the estimators considered so far should be biased because they rely on the strict exogeneity assumption that is violated (Wooldridge, 2009).

The literature on SVI has mostly focused on static regressions, but there are two compelling reasons to adopt the dynamic version when ASVI is the dependent variable (hypothesis 1 and 2). First, it would be interesting to confirm whether ASVI is persistent. For example, weeks with abnormally high search interest may be followed by another week of high search. Second, a static model cannot be used if $y_{i,t-1}$ leads to an omitted variable bias problem. To understand how this can happen, let us take the case of absolute (abnormal) returns and assume ASVI is persistent. Weeks with high search volume may be associated with high returns in the following week. This can happen if we are not controlling for certain events or announcements that take time to be fully incorporated into stock prices. One example is the impact of index additions and deletions as shown by Chen et al. (2004). Therefore, $y_{i,t-1}$ impacts the dependent variable and is correlated with a regressor (returns) but it is unobserved. This causes the classical omitted variable bias, which invalidates inference.

Solving this issue depends on the model being used. Everaert and Groote (2016) extend previous studies and show the fixed effects model with lagged dependent variable has three sources of bias. The standard bias, as documented by Nickell (1981), disappears with large N and T. However, the other two, which are caused by the existence of persistent common factors and factors correlated with some regressors, continue even with large T. Therefore, this estimator is unsuitable as a dynamic model. In contrast, Everaert and Groote (2016) also show that all biases disappear with the CCEP estimator under large N and T. Baltagi (2014) concludes the same for the CCEMG estimator as long as the rank condition, described in section 3.3, is met when factors are persistent. Therefore, given that the rank condition is expected to hold, I use the CCEMG estimator with lagged dependent variable in section 5.

4. Data

This section explains in detail the sample selection process as well as SVI and other relevant variables. All variables used throughout the three hypothesis are defined in table 1, including their source.

4.1 Sample construction

The sample used in this thesis is based on all constituents of the S&P 500 index at any moment in time between 2004 and 2016. The list of companies is obtained from Compustat. There are a total of 792, as measured by each company's unique GVKEY identifier. 17 of them have left and joined the index again, which leads to general attrition in the sample (Baltagi, 2009). However, this is not an issue because I am including all data prior to addition and after deletion. This also helps "*minimize survivorship bias and the impact of index addition and deletion*" (Da et al., 2010 p. 8). Doing so is the same as imposing a fixed panel, where firms in the sample do not change. This avoids another concern, regarding the rotating nature of the index, where "*the fraction of households* [or firms] *that drops from the sample is replaced by an equal number of new households* [or new firms]" (Baltagi, 2009 p. 191).

However, the attrition bias identified in section 3.1 still remains, because there is no more data when a company is delisted. The only solution considered here is implicit to the regression models. Attrition is not a concern when caused by the time or firm effects the estimators are designed to control for (Cameron and Trivedi, 2005). In principle, this would be the case with the CCEMG estimator when, for example, a company goes bankrupt or is taken-over following a recession.

Choosing the S&P 500 index, rather than the Russel 3000 as in Da et al. (2011), has several advantages. It includes the largest and most important companies in the US for market participants. They should be "*widely followed by investors, media and analysts and therefore relevant information such as changes in fundamentals should be rapidly incorporated in prices"* (Latoeiro et al., 2013 p. 8). Even if search volume is expected to be high, there isn't necessarily lack of heterogeneity in investor attention. It can work as an advantage because it "*allows for variation in investor information demand, while holding relatively constant differences in the information environment, which is high for all S&P 500 firms*" (Drake et al., 2012 p. 10). The last advantage is that analysing less than 3x the number of companies considerably facilitates the data cleaning task. For example, there is significant manual work involved when linking company identifiers from different databases.

Table 1. Variables Definition

Variables	Description and Source
Measuring Search Behaviour	
SVI	Scaled measure of weekly aggregate search frequency based on company ticker | Google Trends
Abn SVI (ASVI)	Current week SVI minus the median SVI over the previous 8 weeks | Google Trends
Other variables measuring investor attention	
Abn Ret	Weekly stock return in excess of Fama French 3 factor model return | CRSP and Kenneth French Data Library
Log Absolute Abn Ret	The log of the absolute value of the variable Abn Ret | CRSP and Kenneth French Data Library
Log Abn Turnover	The log of current week turnover divided by the average turnover of the previous 52 weeks, where turnover equals weekly trading volume over shares outstanding | CRSP
Log (1 + Bid-Ask Spread)	The log of one plus the closing ask price minus the closing bid divided by their average | CRSP
Log Size	The log of price multiplied by shares outstanding | CRSP
Analyst Coverage	The number of analysts following each company as measured by the latest number of EPS forecasts before each quarterly earnings report | IBES
Log (1 + Advertising/Sales)	The log of one plus advertising expenses over total sales based on the previous fiscal year end | Compustat
Earnings Announcement	Dummy variable equal to one when earnings are announced and equal to zero otherwise | Compustat and IBES
Number of Announcements	The number of S&P 500 constituents announcing quarterly earnings on the same week | Compustat and IBES
Variables specifically related with retail investor attention	
Log Idiosyncratic Volatility (IVOL)	The log of the standard deviation of the residuals after regressing weekly excess returns on the Fama French 3 factor model | CRSP and Kenneth French Data Library
Idiosyncratic Skewness	Skewness of the residuals after regressing weekly excess returns on the raw and squared values of the excess market return | CRSP and Kenneth French Data Library
Log 1/P	The log of the inverse of the stock price | CRSP
Log (1 + Absolute Earnings Surprise)	The log of one plus the absolute difference between actual EPS and the latest median EPS forecast, scaled by the stock price one week before the earnings announcement | IBES and CRSP
Consumer Industry	Dummy variable equal to one for companies in either the consumer discretionary or consumer staples sectors based on two digit GICS and equal to zero otherwise | Compustat

However, the S&P 500 faces one potential drawback. Past evidence suggests that investor attention is stronger among small stocks (e.g., Da et al., 2011; Bank et al., 2011). This is intuitive because it should be easier to see surges in investor interest for companies whose search volume is typically low. Nevertheless, I find three reasons against that being a limitation here.

First, small stocks may lead to econometric problems. Most observations would be concentrated at low levels because each data point is based on the highest of the series. In addition, Google truncates observations to zero when search is very low (below 1). Therefore, small stocks have low variation for most of the sample, which can cause SVI to be highly persistent or even non-stationary. This time dependence has not been questioned, which could invalidate previous findings.

Second, there is always at least one week equal to 100 because it corresponds to the date with the highest search interest. When there is, overall, little search volume for a stock, higher values of SVI may not traduce into substantial interest growth in absolute terms. This is what ultimately matters because a few retail investors cannot influence financial markets (e.g., Barber et al., 2009).

The last reason is that the predictions of Barber and Odean (2008), which form the basis of hypothesis 3, are "*as strong for large capitalization stocks as for small stocks*" (p. 805). Therefore, whether company size matters should be instead considered an (interesting) empirical issue to be tested.

4.2 Google Search Volume

Data on Google Trends starts in January 2004. See section 1 of appendix A for a detailed explanation of how SVI is computed. Currently, we are only able to download the entire series from 2004 to 2016 with monthly frequency. This cannot be overcome by overlapping different series (see section 2 of appendix A). In fact, it would be better not to do so because the CCEMG estimator requires the firm dimension to be greater than the time dimension, as discussed in section 3.3.

I follow the tradition in the literature and use weekly data, where each series has a maximum of 5 years. Since it is of great interest to assess the results in different time periods, weekly SVI is obtained in three non-overlapping samples of 226 weeks each: from January 2004 to April 2008, May 2008 to August 2012, and September 2012 to December 2016. Note that the first sample only has four weeks less than Da et al. (2011), which is good for comparison purposes. As a result, this is the sample period used throughout this thesis, unless stated otherwise.

One critical decision in data collection is how to identify investor interest. Using the company's name is problematic because it can be used for consumption and employment purposes (e.g., Walmart), the name can have different meanings (e.g., Amazon) or there are several ways to refer to the same company (e.g., Kraft Heinz or just Heinz). A potentially better term, following most of the literature, is a company's ticker symbol for three reasons. First, the ticker is unique to each stock, making the choice less subjective. Second, an investor can easily obtain it from financial news (Ding and Hou, 2015). Then, its character combination is often random enough that only someone interested in financial information would type it (e.g., XRX for Xerox). A total of 907 tickers are obtained from CRSP, which is greater than the number of companies (792) because some change ticker over time. To avoid double counting, only the ticker of the main share class is considered. A computer program is used to automatically type-in the tickers into Google Trends and download the data.

However, there are tickers with generic meanings such as 'PH' and 'R'. As in Da et al. (2011), the hypothesis are tested with the whole sample and after excluding 'noisy' tickers. This is done by removing those with 1 and 2 characters, which are most likely capturing non-stock related searches (Cziraki et al., 2017). As discussed in section 5, the main findings do not change after eliminating 95 companies. This provides evidence against the arguments of Vlastakis and Markellos (2012), among others, that tickers may be a bad choice as stock identifiers in Google Trends.

Another important choice to be made is whether we want to capture the search behaviour of all Google users worldwide or just US users. Following Bank et al. (2011), I argue it is better to focus on the US. This considerably reduces the 'noise' inherent to using the ticker as a company identifier. There are so many languages in the world and abbreviations used that SVI would become meaningless for most search terms. Moreover, given the well documented home bias effect in the US (e.g., Coval and Moskowitz, 1999), I expect that enough search interest exists in domestic companies to make SVI a viable proxy.

Following Da et al. (2011), the main variable of interest in this thesis is Abnormal SVI (ASVI). It is defined as current week SVI minus the median SVI over the previous 8 weeks. The idea is that investor attention is best measured by looking at surges in attention as well as when search decreases are assigned a negative value (Latoeiro et al., 2013). For robustness purposes, the main results are assessed by changing the median length to 4, 13, and 26 weeks.

4.3 Other major variables

All market data in terms of share price, trading volume, shares outstanding and closing bid and ask prices are obtained from CRSP. Two adjustments are needed. To compute first day returns, which are missing for some companies, I decided to use the open price. Second, I removed a few observations that have zero volume and negative price because they correspond to a suspended trading day.

Following Barber and Odean (2008), abnormal turnover equals log of current week turnover divided by the average turnover of the previous 52 weeks, where turnover is defined as weekly stock trading volume over shares outstanding.

The bid-ask spread, similarly defined by Eleswarapu (1997) and Hameed et al. (2010), is the log of one plus the closing ask price minus the closing bid price divided by their average. Following those two papers, I average the daily spreads each week to obtain the weekly spread. This data also requires adjustments. Bid and ask prices are replaced by their previous day values when either one is missing and when the bid price is greater than the ask price (which should not be possible).

Idiosyncratic volatility and idiosyncratic skewness are based on Kumar (2009). The former is calculated as the log of the residuals' standard deviation after regressing weekly excess returns on the Fama French 3 factor model. The latter is the skewness of the residuals after regressing weekly excess returns on the raw and squared values of the excess market return. Both variables are computed using a 6 month rolling window and the factors are obtained from Kenneth French Data Library (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/).

The IBES database is used to obtain the number of analysts following each company and a relative measure of earnings surprise. The latter is defined as the log of one plus the absolute difference between actual EPS and the latest median analyst EPS forecast, scaled by the stock price one week before the earnings announcement (Hirshleifer et al., 2008). Since IBES does not produce any output when there are no analysts, Compustat is used to obtain the announcement dates of earnings reports. However, when IBES produces results, the two databases do not match. I follow the procedure in Dellavigna and Pollet (2009) and assume the earlier date between the two databases is the correct one.

Compustat is also used to obtain accounting data (annual advertising expenses and sales) and two digit sectors based on the global industry classification standard (GICS). In order to keep sample size, I follow several papers (e.g., Da et al., 2011; Ding and Hou, 2015) and assume zero advertising expenses when they are missing.

4.4 Summary statistics and correlations

Table 2 reports several descriptive statistics. In a first step, they are computed in the time series of each stock with at least 52 weeks of data and then averaged across all stocks. The exception is the statistic ‘STDEV (Mean)’, where the second step is the standard deviation of each variable’s mean across all firms. It evaluates whether the mean statistic of each variable differs considerably across firms.

We can see there is an average search interest of 45, which is in line with the claim in section 4.1 that these are widely followed companies. In addition, the ‘STDEV (Mean)’ of SVI is 19, indicating that people are not equally interested about all companies. Despite the relatively high average interest, we can still find meaningful variations over time, similar in magnitude to Cziraki et al. (2017). The standard deviation of ASVI is 11.4, which tells us there should be enough dispersion for SVI to work well with large stocks.

As in Drake et al. (2012), it is clear we are in the presence of large stocks. Analyst coverage is on average high, with around 14 analysts following each company, but the attention environment is once again not equal for every firm. The average market cap is $17.9bn, but there are large differences across firms due to outliers such as Exxon Mobil with $500bn in mid-2004. The lack of sample variation in the bid-ask spread, where half of the sample has weekly spreads lower than 0.07%, may cause this variable to be insignificant. We can also see the effect of the earnings season, with an average of 53 companies reporting their quarterly results on the same week.

Many variables have high enough average skewness and excess kurtosis to make a log transformation useful. In the case of idiosyncratic skewness and abnormal returns, this is not applied because these sample statistics would actually increase. Following Da et al. (2011), when a variable can take the value of zero, such as the bid-ask spread, I sum one before apply the log.

Table 3 presents a correlation matrix following the method of table 2 and Da et al. (2011). Overall, the results are aligned with expectations. There is practically no correlation between ASVI and the other variables, meaning that Google search is capturing an unrelated event. The exceptions are absolute abnormal returns and abnormal turnover, where the correlations are 4.7% and 12.3%, respectively. If ASVI is to capture investor attention, it should be positively correlated with returns and trading volume (e.g., Preis et al., 2010; Vlastakis and Markellos 2012).

Table 2. Summary Statistics

This table shows descriptive statistics for all non-dummy variables defined in table 1 but without any log transformation. They are first calculated for each firm with at least 52 weeks of data and then averaged across all firms. 'STDEV (Mean)' stands for the standard deviation of each variable's mean statistic across all firms. 'STDEV' stands for the average of each variable's standard deviation across all firms. The sample includes 728 companies with weekly data from January 2004 to April 2008, but the data is unbalanced.

	Mean	STDEV (Mean)	STDEV	Min	Median	Max	Skew	Excess Kurtosis
SVI	45.16	19.02	13.02	20.48	43.79	98.12	0.74	1.11
ASVI	1.02	1.10	11.36	-27.44	0.08	50.45	0.67	1.66
Abn Ret	-0.01%	0.18%	3.79%	-13.99%	-0.11%	15.06%	0.15	2.03
Absolute Abn Ret	2.75%	1.00%	2.59%	0.02%	2.08%	17.05%	2.10	10.82
Abn Turnover	1.09	0.08	0.53	0.34	0.97	4.42	2.40	15.09
Bid-Ask Spread	0.09%	0.06%	0.06%	0.02%	0.07%	0.48%	2.62	15.22
Size ($ bn)	17.90	34.46	3.65	11.35	17.67	25.78	0.19	2.62
Analyst Coverage	13.91	6.72	2.09	9.76	14.01	17.33	-0.20	0.75
Advertising / Sales	1.29%	2.68%	0.21%	0.94%	1.29%	1.61%	0.00	4.75
Number of Announcements	53.48	0.56	67.23	0.01	20.11	253.51	1.55	1.24
IVOL	3.56%	1.31%	0.94%	2.04%	3.41%	6.07%	0.68	3.47
Idiosyncratic Skewness	0.14	0.35	0.98	-2.07	0.14	2.27	0.01	2.54
1/P	0.04	0.04	0.01	0.02	0.03	0.08	0.50	3.44
Abs Earnings Surprise	0.39%	1.49%	0.61%	0.01%	0.20%	2.64%	1.34	5.18

However, the correlations are still relatively low, which is in light with the argument of Da et al. (2011) that returns and turnover are the outcome of several economic factors (e.g., changes in company growth prospects). Interestingly, there is a significant positive correlation of 26.4% between absolute abnormal returns and abnormal turnover, which is not far from the 31.1% reported by Da et al. (2011).

Furthermore, there is a negative correlation of -11.2% between size and the bid-ask spread, which is in line with the idea that smaller stocks are more illiquid (e.g., Amihud, 2002). Given the lack of small stocks in this sample, this is good news about the usefulness of the bid-ask spread as a liquidity proxy. We can also

Table 3. Correlation Matrix

This table presents contemporaneous correlations for all non-dummy variables defined in table 1. They are first calculated for each firm with at least 52 weeks of data and then averaged across all firms. The sample includes 728 companies with weekly data from January 2004 to April 2008, but the data is unbalanced.

		1	2	3	4	5	6	7	8	9	10	11	12	13
1	ASVI	-												
2	Abn Ret	0.010	-											
3	Log Absolute Abn Ret	0.047	0.023	-										
4	Log Abn Turnover	0.123	0.033	0.264	-									
5	Log (1 + Bid-Ask Spread)	0.015	0.008	0.048	0.073	-								
6	Log Size	0.001	0.024	-0.030	-0.006	-0.112	-							
7	Analyst Coverage	0.005	0.003	-0.006	-0.004	-0.039	0.166	-						
8	Log (1 + Advertising/Sales)	-0.002	0.001	-0.006	-0.001	-0.006	-0.007	-0.009	-					
9	Number of Announcements	0.044	0.016	0.080	0.186	0.039	-0.021	-0.005	-0.003	-				
10	Log IVOL	0.010	0.016	0.184	0.077	0.088	-0.098	-0.029	-0.002	-0.003	-			
11	Idiosyncratic Skewness	-0.001	0.027	0.001	-0.007	-0.006	0.063	-0.039	0.039	-0.001	0.038	-		
12	Log 1/P	-0.004	-0.029	0.047	-0.009	0.146	-0.766	-0.093	-0.013	0.017	0.156	-0.077	-	
13	Log (1 + Absolute Earnings Surprise)	-0.001	0.017	0.020	0.019	0.048	-0.086	-0.015	-0.001	0.002	0.079	0.016	0.176	-

see that smaller stocks have higher idiosyncratic volatility, similar to what Kumar (2009) finds. In addition, size has a correlation of 16.6% with analyst coverage, which is intuitive because it is well documented that larger firms have more analyst forecasts (e.g., Chordia et al., 2007). As a last remark, given that size is the product of shares outstanding with price, its high negative correlation with 1/P is expected.

4.5 Univariate tests

The first univariate test considered here is the cross-sectional dependence (CD) test of Pesaran (2015). Its null hypothesis is that a variable has weak CD whereas the alternative is strong CD. The results are shown in table 4 and the null hypothesis is always rejected. Obviously, given the large sample, with approx. 150,000 observations, it is easy to reject the null even with a very small magnitude (Baltagi, 2009). Nevertheless, several variables report average cross-sectional correlations above 9%, indicating it might be prudent to control for cross-sectional dependence in the regressions. In section 5, this test is also applied to the residuals of each regression given the arguments of Thompson (2010) discussed in section 3.2. Note that the variable number of announcements has a nearly perfect cross-correlation of 99% because, by definition, the total number of earnings announcements at a given week is the same for all firms. Therefore, I expect that this variable cannot be statistically significant with the CCEMG estimator.

The second univariate test is non-stationarity. In the context of hypothesis 3, it implies that "*the time-series average of the cross-sectional coefficients as in Fama and MacBeth (1973) may not converge to the population estimates*" (Chordia et al., 2007 p. 718). More generally, it gives rise to a spurious relation between variables, leading to unreliable inference (Granger and Newbold, 1974). Under large N and T, Phillips and Moon (1999) show there is no problem as long as weak CD holds, which is not the case here. Therefore, I use the cross-sectionally augmented Dickey-Fuller (CADF) test of Pesaran (2007), which allows strong CD. The null hypothesis is that a given variable is non-stationary for all firms while the alternative is stationarity for at least one firm. Following Chordia et al. (2007), the potential candidates are 1/P, size, and analyst coverage, but also advertising over sales due to its much lower (annual) frequency.

In table 4, as suspected, the null of non-stationarity is not rejected for size, advertising over sales and 1/P. To make them stationary, I follow Chordia et al. (2007) and take the residual after regressing each variable on a linear and quadratic

Table 4. Univariate Tests

This table shows the results of two univariate tests using all variables defined in table 1 (except for consumer industry). The test on cross-sectional dependence (CD) is the one of Pesaran (2015), where the null hypothesis is weak CD. The numbers represent the average cross-sectional correlations of each variable. Testing for unit roots is done with the CADF test of Pesaran (2007), where the null hypothesis is non-stationarity for all firms. This test is conducted with an intercept and two lags but also a linear time trend for four variables: log size, analyst coverage, log of one plus advertising over sales, and log 1/P. The numbers on the third column represent t-statistics of the CADF test, whose critical values at the 1% significance level are -3.84 (no trend) and -4.31 (with trend). The sample includes 739 companies with unbalanced weekly data from January 2004 to April 2008. *, **, *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

	Cross-Sectional Correlations (CD Test)	Unit Root or CADF Test
ASVI	1%***	-122.53***
Abn Ret	1%***	-127.87***
Log Absolute Abn Ret	3%***	-127.65***
Log Abn Turnover	25%***	-103.70***
Log (1 + Bid-Ask spread)	9%***	-98.613***
Log Size	24%***	4.42
Analyst Coverage	3%***	-5.23***
Log (1 + Advertising/Sales)	0%***	71.65
Earnings Announcement	13%***	-130.97***
Number of Announcements	99%***	-122.95***
Log IVOL	12%***	-10.42***
Idiosyncratic Skewness	1%***	-18.27***
Log 1/P	13%***	5.48
Log (1 + Absolute Earnings Surprise)	0%***	-32.83***

time trends as well as 51 week dummies. Their residuals now reject the null at the 1% significance level, with t-statistics below -10 (not reported). Furthermore, although analyst coverage rejects the null at the 1% significance level, the t- statistic as defined by Pesaran (2007) is -5.23, which is very close to the underlying critical value of -4.31. Westerlund and Reese (2016) argue that "*the results should be overwhelmingly against the null* (…)" (p. 18) to be confident there are no unit roots. Therefore, as in Da et al. (2011), I sum one before apply the log transformation to analyst coverage and the resulting t-statistic now becomes -10.2 (not reported).

5. Results and Analysis

This section explains in detail each hypothesis and presents all the associated empirical results, including tests that guide model specification.

5.1 Hypothesis 1

H1: SVI is similar to existing proxies of investor attention.

This is the very first hypothesis because SVI is not very useful if it can already be explained by attention proxies in the literature. The hypothesis is rejected by Da et al. (2011) with weekly data and by Drake et al. (2012) with daily frequency. Even though not all regressors are the same, the two papers share the old methodology of Google Trends (before merging with Insights) and pooled OLS estimation with week dummies and robust standard errors clustered by firm.

This hypothesis is tested through a series of panel regressions that consider different estimators: pooled OLS, fixed effects, and CCEMG. The purpose of the last two is to show that researchers are ignoring important characteristics of the data. First, size and advertising over sales don't become stationarity with a log transformation. Second, if the Hausman test is rejected, it is not enough to cluster the standard errors because firm fixed effects have to be removed. Third, besides the regressors, the residuals may also be cross-sectionally correlated, which is only partially corrected with week dummies. Finally, I suspect that ASVI is a persistent variable and its lagged values are correlated with some regressors.

In all three estimators, ASVI is regressed on 7 proxies of attention, with size being used as a control variable: a) absolute abnormal returns, abnormal turnover, analyst coverage, and advertising as a percentage of sales are based on Da et al. (2011); b) the bid-ask spread, earnings announcement date, and the number of earnings announcements are from Drake et al. (2012).

The results are reported in table 5, with five regressions from January 2004 to April 2008. The first regression follows the method of Da et al. (2011) and Drake et al. (2012). The results are consistent with those two papers in terms of low explanatory power (R^2 around 2%) as well as regarding coefficient sign and the level of statistical significance. The exceptions are with size and analyst coverage, which are now statistically significant at the 1% level, and with the bid-ask spread that has an unexpected negative sign. However, this sign becomes positive in the second regression, as variables are adjusted for non-stationarity.

Table 5. ASVI as a function of existing Attention Proxies

This table presents five regressions where the dependent variable is ASVI. All variables are defined in table 1. Except for column (1), size, analyst coverage, and advertising over sales are adjusted for non-stationarity as in section 4.5. The intercept is omitted because it is not meaningful in panel models. Pooled OLS is used in the first two regressions and the third is based on the fixed effects model. These three regressions have 51 week dummies and standard errors, shown in parentheses, are clustered by firm (739 firms). The last two columns use the CCEMG estimator, where cross-sectional averages of all variables are included as additional regressors. The robust version of the Hausman test is conducted with bootstrapped standard errors (100 repetitions). Under the null hypothesis, the random effects model is consistent and the t-statistic, which is the same for all regressions, follows a Chi-squared distribution with 59 degrees of freedom. The t-statistic of the CD test from Pesaran (2015) is calculated on the residuals of each regression and it follows a standard normal distribution. Under the null, the error term is weakly cross-sectionally dependent. The sample period is from January 2004 to April 2008 but the data is unbalanced. *, **, *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

	(1)	(2)	(3)	(4)	(5)
Log Absolute Abn Ret	0.327*** (0.068)	0.346*** (0.068)	0.283*** (0.070)	0.033 (0.087)	0.106 (0.148)
Log Abn Turnover	6.598*** (0.365)	6.547*** (0.362)	6.955*** (0.380)	8.792*** (0.523)	8.011*** (0.544)
Log (1 + Bid-Ask Spread)	-0.011 (1.209)	0.493 (1.222)	3.060** (1.410)	2.333 (2.542)	2.422 (3.002)
Log Size	-0.430*** (0.107)	-0.289*** (0.102)	1.226*** (0.332)	3.431** (1.674)	-1.863 (2.275)
Analyst Coverage	0.029*** (0.006)	0.498*** (0.138)	-0.345 (0.330)	1.542 (2.359)	0.583 (1.749)
Log (1 + Advertising/Sales)	0.040 (0.123)	0.052 (0.123)	-0.340 (0.618)	-11.926 (10.668)	20.637 (119.455)
Earnings Announcement	2.255*** (0.201)	2.254*** (0.201)	2.197*** (0.199)	1.685*** (0.231)	1.996*** (0.266)
Number of Announcements	-0.007*** (0.002)	-0.007*** (0.002)	-0.007*** (0.002)	0.151 (0.107)	0.002 (0.068)
ASVI$_{t-1}$					0.169*** (0.010)
ASVI$_{t-2}$					0.007 (0.005)
ASVI$_{t-3}$					-0.013** (0.006)
ASVI$_{t-4}$					-0.069*** (0.007)
Observations	148,937	148,937	148,937	148,937	145,861
Adj. Stationarity	NO	YES	YES	YES	YES
Week fixed effects	YES	YES	YES	NO	NO
CCEMG Averages	NO	NO	NO	YES	YES
Hausman Test			101.95***		
CD Test	35.78***	36.25***	36.77***	-4.90***	-4.65***
R^2	1.91%	1.91%	2.21%	15.11%	26.95%

The third regression allow us to compare the fixed effects estimator with pooled OLS. The Hausman test strongly rejects the null hypothesis at the 1% significance level, meaning that pooled OLS can lead to biased results. The consequence is that size, advertising over sales, and analyst coverage change sign and the latter is no longer statistically significant even at the 10% level. Thus, there are unobserved firm characteristics that impact ASVI but are also correlated with those variables.

Nevertheless, the first three regressions share the same problem. The CD test of Pesaran (2015) indicates that the regression residuals are strongly cross-sectionally dependent. To mitigate this concern, the last two columns apply the CCEMG estimator of Pesaran (2006). When cross-sectional dependence is taken into account, absolute abnormal returns, the bid-ask spread, and number of announcements become statistically insignificant. These are variables with high average cross-sectional correlations, as shown in section 4.5, although the insignificance of the latter was already expected. Consequently, at least for large stocks, the results of Da et al. (2011) change, where absolute abnormal returns no longer have a strong positive relation with ASVI[6].

Moreover, the R^2 of the fourth regression increases to 15%, but this does not mean that existing proxies have started to explain much more variation in ASVI. The model was augmented with 9 new regressors, which are the cross-sectional averages of ASVI and of the regressors, and some of these unreported variables are highly statistically significant[7].

The last column suggests that ASVI is persistent up to its fourth lag, where only the second isn’t statistically significant at the 5% level. High search interest in one week is followed by high interest in the following week, which then partially reverses in the third and fourth weeks. While the first lag persistency is in line with Da et al. (2011), the partial reversal is a new contribution to the literature. However, this is not surprising given the way ASVI is constructed. It is positive when an important event occurs because current week SVI increases substantially, but then it becomes negative in the next few weeks as SVI goes back to its regular levels.

Furthermore, there are two major implications when including lagged ASVI. The coefficient of absolute abnormal returns increases substantially more than its

[6] A similar conclusion applies to the absolute raw returns used in Drake et al. (2012).

[7] Table 5 (and all others) do not report adjusted R^2 because the difference with the regular statistic is rather small given the large sample being used.

standard error while size becomes statistically insignificant. The former supports the conjecture in section 3.4 that returns are correlated with lagged ASVI. The insignificance of size (and of analyst coverage) is expected since there are only large stocks (Drake et al., 2012). However, not even Da et al. (2011) with smaller stocks finds consistent significance.

The robustness of these results is examined in table C.1 of appendix C. The exclusion of ‘noisy’ tickers does not have a meaningful impact. Abnormal turnover and earnings announcement are positively related with ASVI and statistically significant at the 1% level regardless of how ASVI is computed, the regression models used and the time period considered. There is a similar consistency with the first and fourth lags of ASVI. In contrast, the relation of ASVI with number of announcements using pooled OLS is not robust over time. In different time periods, it is also important to use more robust regression models. Finally, the explanatory power of existing attention proxies is not as small as previously thought, given that the R^2 of the fixed effects regression is almost 9% in the 2012-2016 sample.

5.2 Hypothesis 2

H2: SVI does not capture the search behaviour of retail investors.

Institutional investors use sophisticated data providers like Bloomberg. This contrasts with retail investors, who do not have the financial means or a large enough portfolio to justify the purchase (Da et al., 2011; Ben-Rephael et al., 2017). Therefore, individuals most likely resort to a web search engine like Google.

Using the last regression of table 5, I further include five variables that have been positively associated with retail investor activity[8]. Based on Kumar (2009), I include idiosyncratic volatility, idiosyncratic skewness, and 1/P. The former two try to capture the undiversified nature of retail investors’ portfolios. The latter is based on the premise that these investors have a preference for ‘penny’ stocks. Furthermore, a scaled measure of absolute earnings surprise is included because Hirshleifer et al. (2008) argue that retail investors are net buyers following positive and negative surprises due to increased attention. Finally, based on Lou (2014), I

[8] As opposed to Drake et al. (2012), among others, I do not include a measure of ownership such as institutional holdings over shares outstanding. The reason is it has severe shortcomings as pointed out by Barber et al. (2009). For example, there is considerable overestimation in the record of holdings, with 4% of total observations having ratios above 100%. Thus, winsorizing at the usual levels (99%) to remove outliers would not significantly alleviate the problem.

consider whether search volume is higher for companies in the consumer sector (as defined in table 1) with non-zero advertising expenses.

Note that researchers usually resort to proprietary databases when studying retail activity because it is difficult to identify from aggregate market data (e.g., Barber and Odean, 2008; Barber et al., 2009). Alternatively, Da et al. (2011) use Dash-5 monthly reports. These are reports that US market centres are obliged to file with the SEC since 2001. The researchers find a strong relation between SVI and small orders across market centres, but especially in the old Madoff centre that tended to have more retail orders executed (compared to NYSE and Archipelago). Given that it is very time consuming to obtain this data, I do not follow their method.

Table 6 reports the results where each regression has one of the new retail activity variables. The sign of all coefficients is positive, which is aligned with expectations. However, as anticipated, the statistical significance is in general weak. The exception is idiosyncratic volatility, which is a persistent variable that becomes statistically significant at the 1% level when its first lag is included. Given that idiosyncratic volatility and its lag have opposing but very similar coefficient magnitudes, the relevant driver might be its one period change. This is confirmed in a sixth regression where current week idiosyncratic volatility loses its statistical significance and positive impact on ASVI to its one period change.

To test the robustness of these results, table C.2 of appendix C presents seven regressions with all retail activity variables put together. Only one variable is robust to how ASVI is computed and different time periods, although idiosyncratic skewness is a ‘2nd best variable’. There is strong evidence that weeks with an increase in idiosyncratic volatility are associated with a surge in people’s interest. The effect is more pronounced when ‘noisy’ tickers are removed. This means that when firm-specific risk is perceived to increase, people demand more company information, which is in line with the arguments of Arbel (1985) and Leuz and Verrecchia (2000) described in section 2.2. Given that this risk is unsystematic, the results are in favour of the idea that SVI is capturing individual investor activity. It is also worth noting that abnormal turnover and earnings announcement are still consistently strongly related with ASVI across all regressions.

Table 6. ASVI and Retail Investor Attention

This table presents six regressions where the dependent variable is ASVI. All variables are defined in table 1. Size, analyst coverage, and advertising over sales are adjusted for non-stationarity as in section 4.5. The CCEMG estimator is used, where cross-sectional averages of all variables are included as additional regressors. The intercept is omitted because it is not meaningful in panel models. Standard errors are shown in parentheses. The sample period is from January 2004 to April 2008 but the data is unbalanced. *, **, *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Log Absolute Abn Ret	0.189* (0.100)	0.039 (0.140)	0.041 (0.145)	0.076 (0.145)	-0.020 (0.154)	-0.024 (0.152)
Log Abn Turnover	7.767*** (0.582)	8.293*** (0.526)	8.145*** (0.546)	8.255*** (0.540)	7.178*** (0.590)	7.124*** (0.600)
Log (1 + Bid-Ask Spread)	-4.102 (5.256)	5.356* (2.950)	3.548 (2.908)	2.547 (2.992)	-0.720 (3.657)	-0.452 (3.642)
Log Size	7.359 (5.903)	6.544 (6.527)	-0.712 (2.220)	0.540 (2.351)	6.109 (4.011)	5.328 (4.266)
Analyst Coverage	0.940 (1.988)	-0.071 (2.457)	-1.000 (1.007)	1.343 (1.825)	-0.971 (3.109)	-0.567 (3.159)
Log (1+ Advertising/Sales)	47.550 (30.048)	-46.115 (41.332)	-42.525 (29.816)	-46.423 (115.538)	-79.529 (129.193)	-79.526 (129.193)
Earnings Announcement	1.866*** (0.209)	2.110*** (0.263)	2.095*** (0.268)	1.957*** (0.265)	1.968*** (0.265)	2.001*** (0.270)
Number of Announcements	0.087 (0.103)	0.096 (0.105)	0.003 (0.059)	0.008 (0.060)	0.089 (0.068)	0.089 (0.068)
Idiosyncratic Skewness	0.183 (0.236)					
Log 1/P		68.178 (67.159)				
Log (1 + Absolute Earnings Surprise)			98.599 (98.084)			
Consumer Sector x Advertising/Sales				3.735 (4.037)		
Log IVOL					7.811*** (1.962)	-0.640 (1.955)
Log IVOL$_{t-1}$					-8.680*** (2.278)	
Change Log IVOL						8.986*** (2.274)
Observations	145,861	145,861	145,861	145,861	145,861	145,861
CCEMG Averages	YES	YES	YES	YES	YES	YES
ASVI$_{t-1,\dots,t-4}$	YES	YES	YES	YES	YES	YES
R^2	27.92%	27.97%	27.91%	27.50%	28.83%	28.73%

5.3 *Hypothesis 3*

H3: SVI does not predict positive returns in the short-term for large stocks, which reverse within a year.

Barber and Oden (2008) argue that individual investors have many investment opportunities, but their selling options are constrained to the stocks they own since short selling is difficult. Thus, these investors are net buyers of stocks that catch their attention, irrespective of company size and whether those attention shocks result from positive or negative information. The subsequent price increase is short-lived, given the uninformed nature of individual investors. Da et al. (2011) confirm this argument called positive price pressure. They report that high search volume predicts abnormally high returns in the next two weeks, which start reversing after the 4th week but without having strong statistical significance. The effect is more pronounced among stocks traded by retail investors, but the results only hold for the smaller half of the Russell 3000 index. Cziraki et al. (2017) also find there is no predictability for large stocks when using the S&P 500 with monthly frequency.

To test hypothesis 3, I run two sets of regressions. First, I closely follow Da et al. (2011) by estimating five Fama-MacBeth (1973) regressions. Future Fama-French 3 factor abnormal returns[9] (in basis points) are regressed on current week: ASVI, size, change in idiosyncratic volatility, absolute abnormal returns, abnormal turnover, bid-ask spread (as in Cziraki et al., 2017), number of analysts, and advertising as a percentage of sales. In the first four regressions, the dependent variable is over the next 1 to 4 weeks while the last examines long-run abnormal returns from weeks 5 to 52. ASVI is also interacted with size and with the change in idiosyncratic volatility. The goal is to assess whether ASVI's impact is stronger, respectively, among smaller stocks and those with higher retail activity. The independent variables are cross-sectionally demeaned and also standardized. The standard errors are adjusted using the Newey-West (1987) formula with 8 lags.

However, the approach of Da et al. (2011) may ignore important correlations between omitted variables and the regressors. Therefore, the results are compared in a second set of regressions with the same variables but under the CCEMG estimator, which might change previous conclusions.

[9] Da et al. (2011) and Cziraki et al. (2017) use instead DGTW abnormal returns. I do not follow their method for two reasons. First, these returns require the computation of book to market, which is problematic because 3.5% of observations have a negative ratio. Then, their findings should be robust to the way abnormal returns are computed, which only Cziraki et al. (2017) verifies.

The results are shown in table 7. When Fama-MacBeth (1973) regressions are used, ASVI and size have a negative relation with future abnormal returns while abnormal turnover and the bid-ask spread are positively associated. This confirms the main findings of Da et al. (2011) and Cziraki et al. (2017). I even obtain stronger results for ASVI. Its negative impact is statistically significant at the 1% level in the first week and greatly reverses during weeks 5 to 52. Taken at face value, this would lead us to strongly support Da et al. (2011) in that the predictions of Barber and Odean (2008) are not valid for large stocks. In fact, they appear to be reversed, leading to negative price pressure, which is counter-intuitive.

However, as suspected, the null hypothesis of the Hausman test is strongly rejected and, to a smaller extent, the same applies to the CD test. This indicates that the previous method may not be the best one. When the CCEMG estimator is used, the main conclusions significantly change. Even though the statistical significance is not strong and the magnitude is small, a one standard deviation increase in current week ASVI predicts higher abnormal returns next week by 2.2 bps. This positive effect lasts four weeks, and then reverses within one year. When ASVI is interacted with company size, the resulting coefficient is negative. Thus, the effect is stronger among ‘smaller’ stocks, which is the only similar prediction as Da et al. (2011). In addition, there is weak evidence that the bid-ask spread and abnormal turnover are positively related with future abnormal returns. The latter casts doubt on the existence of a volume-return premium as first argued by Gervais et al. (2001) and confirmed in Da et al. (2011).

A comprehensive robustness analysis on the CCEMG estimator is shown in table C.3 of appendix C. An alternative to the method of Da et al. (2011) is still necessary because the null of the Hausman test is always rejected as well as the null of the CD test, even when the latter is based on the residuals of a more robust estimator. Across different time periods and window lengths to compute ASVI, the results of Barber and Odean (2008) consistently hold for large stocks. ASVI predicts higher abnormal returns one week ahead in the future and lower returns from weeks 5 to 52. When ‘noisy’ tickers are removed, these two effects are stronger. The first week results become statistically significant at the 1% level when the one period lag of ASVI is included, but this is not robust across time periods (not reported). ASVI is not always stronger among smaller stocks and for those with supposedly higher retail investor activity. This latter result is consistent with the idea that it is difficult to identify retail activity from aggregate market data.

Table 7. ASVI and Future Abnormal Returns

This table reports the results of two panels with five regressions each. The dependent variable is the Fama French 3 factor abnormal return (in basis points) during the next 4 weeks and during weeks 5 to 52. All variables are defined in table 1. Size, analyst coverage, and advertising over sales are adjusted for non-stationarity as in section 4.5. The independent variables are standardized so that their impact can be interpreted as a one standard deviation change. The regressions on panel A are based on the Fama-MacBeth (1973) procedure. All variables are also cross-sectionally demeaned and standard errors are adjusted using the Newey-West (1987) formula with 8 lags. The regressions on panel B are based on the CCEMG estimator, where cross-sectional averages of all variables are included as additional regressors. The intercept is omitted because it is not meaningful in panel models. For both panels, the robust version of the Hausman test is conducted with bootstrapped standard errors (100 repetitions). Under the null, the random effects model is consistent and the t-statistic follows a Chi-squared distribution with 59 degrees of freedom. The t-statistic of the CD test from Pesaran (2015) is calculated on the residuals of each regression and it follows a standard normal distribution. Under the null, the error term is weakly cross-sectionally dependent. Standard errors are shown in parentheses. The sample period is from January 2004 to April 2008 but the data is unbalanced. *, **, *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

	Week 1	Week 2	Week 3	Week 4	Week 5-52
	(1)	(2)	(3)	(4)	(5)
Panel A. Fama-MacBeth (1973) cross-sectional regressions					
ASVI	-2.751*** (1.043)	-1.853 (1.213)	-1.034 (1.247)	-1.470 (1.109)	34.954*** (10.768)
Log Size x ASVI	0.080 (1.718)	-1.711 (1.361)	0.404 (1.091)	-0.408 (1.098)	6.890 (9.155)
Log Size	-10.959*** (2.494)	-9.865*** (2.401)	-8.680*** (2.407)	-8.419*** (2.308)	8.588 (26.075)
Change Log IVOL x ASVI	-0.273 (1.600)	0.075 (1.775)	-2.135 (1.303)	2.613* (1.359)	-17.779 (13.208)
Change Log IVOL	-0.979 (1.784)	1.408 (1.640)	0.179 (2.054)	-0.054 (1.780)	36.799*** (11.067)
Log Absolute Abn Ret	-4.392*** (1.496)	-5.868*** (1.757)	-1.078 (1.505)	-1.922 (1.517)	-107.260*** (21.466)
Log Abn Turnover	1.832 (1.660)	0.595 (1.721)	-1.223 (1.543)	0.105 (1.571)	36.492 (22.393)
Log (1 + Bid-Ask Spread)	2.108 (2.260)	2.321 (2.219)	3.294** (1.528)	2.149 (1.598)	1.111 (17.544)
Analyst Coverage	5.513** (2.660)	4.911* (2.760)	4.447 (2.781)	4.241* (2.568)	29.679 (42.265)
Log (1 + Advertising/Sales)	-0.188 (1.725)	-0.540 (1.723)	-0.445 (1.708)	-0.317 (1.696)	-48.912** (23.160)
Observations	148,198	147,459	146,720	145,981	145,242
Hausman Test	154.32***	254.69***	173.35***	222.15***	111.32***
CD Test	-8.06***	-8.15***	-2.85***	-3.07***	-9.35***
R^2	2.62%	2.49%	2.32%	2.39%	2.49%

(continued)

Table 7. - *Continued*

	Week 1 (1)	Week 2 (2)	Week 3 (3)	Week 4 (4)	Week 5-52 (5)
Panel B. CCEMG estimator of Pesaran (2006)					
ASVI	2.218* (1.254)	0.222 (0.657)	2.159 (2.905)	2.202 (2.364)	-2.967 (2.052)
Log Size x ASVI	-2.714* (1.541)	-0.272 (0.623)	-0.828 (2.686)	-1.900 (1.757)	2.631 (1.716)
Log Size	-12.612*** (1.366)	-7.961*** (0.561)	-8.506*** (0.573)	-6.737*** (0.439)	-0.461 (1.884)
Change Log IVOL x ASVI	-0.072 (0.107)	0.202 (0.214)	-0.232 (0.296)	0.246 (0.223)	0.276 (0.279)
Change Log IVOL	0.245 (0.256)	-0.056 (0.040)	0.042 (0.038)	-0.118 (0.185)	-0.022 (0.077)
Log Absolute Abn Ret	0.014 (0.038)	0.010 (0.029)	0.024 (0.057)	-0.034 (0.042)	-0.071 (0.059)
Log Abn Turnover	-0.099 (0.134)	0.076 (0.102)	-0.094 (0.146)	0.049 (0.054)	-0.254** (0.129)
Log (1 + Bid-Ask Spread)	0.002 (0.027)	-0.010 (0.037)	0.046 (0.056)	0.042 (0.037)	-0.061 (0.122)
Analyst Coverage	0.316 (0.310)	-3.265 (3.283)	0.423 (0.285)	0.419* (0.235)	1.415 (1.525)
Log (1 + Advertising/Sales)	-20.893 (19.924)	1.637 (31.746)	-24.038 (15.154)	-23.127 (14.694)	-199.046 (149.708)
Observations	148,198	147,459	146,720	145,981	145,242
CCEMG Averages	YES	YES	YES	YES	YES
Hausman Test	154.32***	254.69***	173.35***	222.15***	111.32***
CD Test	0.04	1.50	-0.40	0.03	2.78***
R^2	17.35%	15.88%	15.15%	14.17%	16.60%

6. Conclusion and Future Research

We know since Merton (1987) that investor attention matters. The literature has tried to confirm past predictions and reach new conclusions. However, researchers cannot escape from the most critical task: how to best measure people's attention? With the Internet, a new class of investor attention proxies has emerged. Da et al. (2011) popularized the use of Google search volume as a demand proxy.

This thesis assesses in detail the robustness of several findings in the SVI literature to recent time periods and to more robust panel estimators. If Google search volume did not capture company stock search, I would not find a consistent relation with dates of earnings announcements and abnormal turnover. In contrast, all other proxies fail. This includes absolute abnormal returns, which Da et al.

(2011) shows to be strongly positively associated with ASVI. When considering all (indirect) proxies together in recent periods, I confirm there is still substantial variation in ASVI left unexplained, but not as little as previously thought.

Furthermore, I provide consistent evidence that more robust panel models to time and firm effects can significantly change previous findings. The best example is in hypothesis 3, where the results under the CCEMG estimator of Pesaran (2006) are the opposite of those with Fama-MacBeth (1973) regressions. I show that an increase in ASVI for large stocks predicts higher abnormal returns one week ahead, which reverse within one year. As a result, I conclude it is very important to examine the structure of the data in order to choose the most appropriate estimators.

Future research

This thesis invites researchers to reconsider the robustness of their predictions. For example, Ding et al. (2015) find with only 6 years of data and pooled OLS that weekly SVI improves the liquidity and enlarges the shareholder base of S&P 500 stocks. However, table 4 shows that the bid-ask spread has a non-negligible cross-sectional correlation of 9%. This is also relevant for the variable number of shareholders because their sample only has large companies. Thus, it is important to see if the results hold with the CCEMG estimator and in different time periods.

Moreover, this thesis confirms that SVI captures people’s attention to earnings announcements. I further propose testing the impact of index additions and deletions. Chen et al. (2004) find that the stock price response following changes to the S&P 500 index is asymmetric. Added firms experience a permanent price impact whereas excluded firms only face a temporary decline. Their explanation is based on changes in investor awareness. Investors become aware of a stock when it is added to the index but they do not simply forget about those that are deleted. Therefore, SVI is a great candidate to test this hypothesis, which had only been done by those researchers with (relatively weak) market proxies.

It is important to keep in mind that SVI is still far from being a perfect proxy. Regardless of how search terms are defined, we cannot be sure of people’s final intentions (Latoeiro et al., 2013). Having a financial database like Yahoo! Finance would help, since Lawrence et al. (2016) suggests it is superior to Google Trends. Nevertheless, narrowing the analysis to retail investors seems too restrictive. Over the last decades, institutional investors have increasingly become the largest stock owners (Stambaugh, 2014). Ben-Rephael et al. (2017) are the first to study their search behaviour and find that it considerably differs from retail activity, namely in how quickly they react to news. Their study could be easily extended to test all kind of attention theories and predictions mentioned throughout this thesis.

A related topic is the impact of increasing algorithmic trading by sophisticated investors (McGowan, 2010). The question has always been asked in terms of human's attention but computers do not suffer from the same limitations. Algorithms have contributed to improved liquidity (e.g., Hendershott et al., 2011) but no paper directly links them with the investor attention literature. For instance, Dellavigna and Pollet (2009) find that the post-earnings announcement drift is more pronounced on Friday announcements. They argue investors are more distracted at the end of week, but this cannot occur with algorithms. Therefore, is the drift lower for stocks more heavily traded by computers? A database on algorithmic activity such as NYSE ProTrac would shed light on that and related questions.

Finally, paying attention does not directly affect financial markets, but rather indicates when investors start processing new information (Latoeiro et al., 2013). This is an area still to be explored because of the difficulty in tracking people's effort (Mondria et al., 2010). We would need more comprehensive datasets than the ones currently available to obtain, for example, a measure of the time people spend reading news articles. I leave that, and related issues, to future research.

References

Ahn, Seung C., Young H. Lee, and Peter Schmidt. "Panel data models with multiple time-varying individual effects." *Journal of Econometrics* 174.1 (2013): 1-14.

Amihud, Yakov. "Illiquidity and stock returns: cross-section and time-series effects." *Journal of financial markets* 5.1 (2002): 31-56.

Andrei, Daniel, and Michael Hasler. "Investor attention and stock market volatility." *The Review of Financial Studies* 28.1 (2014): 33-72.

Antweiler, Werner, and Murray Z. Frank. "Is all that talk just noise? The information content of internet stock message boards." *The Journal of Finance* 59.3 (2004): 1259-1294.

Arbel, Avner. "Generic stocks: An old product in a new package." *The Journal of Portfolio Management* 11.4 (1985): 4-13.

Askitas, Nikolaos, and Klaus F. Zimmermann. "Google econometrics and unemployment forecasting." *Applied Economics Quarterly* 55.2 (2009): 107-120.

Bai, Jushan. "Panel data models with interactive fixed effects." *Econometrica* 77.4 (2009): 1229-1279.

Bai, Jushan, and Serena Ng. "Large dimensional factor analysis." *Foundations and Trends® in Econometrics* 3.2 (2008): 89-163.

Baltagi, Badi H. *Econometric Analysis of Panel Data*. 3rd ed. Wiltshire: John Wiley & Sons, 2009

Baltagi, Badi H. *The Oxford handbook of panel data*. New York: Oxford University Press, 2014.

Bank, Matthias, Martin Larch, and Georg Peter. "Google search volume and its influence on liquidity and returns of German stocks." *Financial markets and portfolio management* 25.3 (2011): 239-264.

Barber, Brad M., and Terrance Odean. "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors." *The Review of Financial Studies* 21.2 (2008): 785-818.

Barber, Brad M., Terrance Odean, and Ning Zhu. "Do retail trades move markets?." *The Review of Financial Studies* 22.1 (2009): 151-186.

Beatty, Sharon E., and Scott M. Smith. "External search effort: An investigation across several product categories." *Journal of consumer research* 14.1 (1987): 83-95.

Ben-Rephael, Azi, Zhi Da, and Ryan D. Israelsen. "It Depends on Where You Search: Institutional Investor Attention and Underreaction to News." *The Review of Financial Studies* 30.9 (2017): 3009-3047.

Berry, Thomas D., and Keith M. Howe. "Public information arrival." *The Journal of Finance* 49.4 (1994): 1331-1346.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." *Journal of computational science* 2.1 (2011): 1-8.

Cameron, A. Colin, and Pravin K. Trivedi. *Microeconometrics: methods and applications*. New York: Cambridge university press, 2005.

Campello, Murillo, Antonio Galvao, and Ted Juhl. "Policy heterogeneity in empirical corporate finance." Working paper (2013).

Chemmanur, Thomas, and An Yan. "Product market advertising and new equity issues." *Journal of Financial Economics* 92.1 (2009): 40-65.

Chen, Hailiang, et al. "Wisdom of crowds: The value of stock opinions transmitted through social media." *The Review of Financial Studies* 27.5 (2014): 1367-1403.

Chen, Honghui, Gregory Noronha, and Vijay Singal. "The price response to S&P 500 index additions and deletions: Evidence of asymmetry and a new explanation." *The Journal of Finance* 59.4 (2004): 1901-1930.

Choi, Hyunyoung, and Hal Varian. "Predicting initial claims for unemployment benefits." *Google Inc* (2009): 1-5.

Chordia, Tarun, Sahn-Wook Huh, and Avanidhar Subrahmanyam. "The cross-section of expected trading activity." *The Review of Financial Studies* 20.3 (2007): 709-740.

Chudik, Alexander, and M. Hashem Pesaran. "Large panel data models with cross-sectional dependence: a survey." Working Paper (2013).

Corwin, Shane A., and Jay F. Coughenour. "Limited attention and the allocation of effort in securities trading." *The Journal of Finance* 63.6 (2008): 3031-3067.

Coval, Joshua D., and Tobias J. Moskowitz. "Home bias at home: Local equity preference in domestic portfolios." *The Journal of Finance* 54.6 (1999): 2045-2073.

Cziraki, Peter and Mondria, Jordi and Wu, Thomas, "Asymmetric Attention and Stock Returns". Working Paper (2017)

Da, Zhi, Joseph Engelberg, and Pengjie Gao. "Internet search and momentum." Working Paper (2010).

Da, Zhi, Joseph Engelberg, and Pengjie Gao. "In search of attention." *The Journal of Finance* 66.5 (2011): 1461-1499.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. "The sum of all FEARS investor sentiment and asset prices." *The Review of Financial Studies* 28.1 (2014): 1-32.

Das, Sanjiv R., and Mike Y. Chen. "Yahoo! for Amazon: Sentiment extraction from small talk on the web." *Management science* 53.9 (2007): 1375-1388.

DellaVigna, Stefano, and Joshua M. Pollet. "Investor inattention and Friday earnings announcements." *The Journal of Finance* 64.2 (2009): 709-749.

Ding, Rong, and Wenxuan Hou. "Retail investor attention and stock liquidity." *Journal of International Financial Markets, Institutions and Money* 37 (2015): 12-26.

Drake, Michael S., Darren T. Roulstone, and Jacob R. Thornock. "Investor information demand: Evidence from Google searches around earnings announcements." *Journal of Accounting Research* 50.4 (2012): 1001-1040.

Drake, Michael S., Darren T. Roulstone, and Jacob R. Thornock. "The determinants and consequences of information acquisition via EDGAR." *Contemporary Accounting Research* 32.3 (2015): 1128-1161.

Eleswarapu, Venkat R. "Cost of transacting and expected returns in the Nasdaq market." *The Journal of Finance* 52.5 (1997): 2113-2127.

Everaert, Gerdie, and Tom De Groote. "Common correlated effects estimation of dynamic panels with cross-sectional dependence." *Econometric Reviews* 35.3 (2016): 428-463.

Fama, Eugene F., and James D. MacBeth. "Risk, return, and equilibrium: Empirical tests." *Journal of political economy* 81.3 (1973): 607-636.

Fang, Lily, and Joel Peress. "Media coverage and the cross-section of stock returns." *The Journal of Finance* 64.5 (2009): 2023-2052.

French, Kenneth R., and Richard Roll. "Stock return variances: The arrival of information and the reaction of traders." *Journal of financial economics* 17.1 (1986): 5-26.

Frieder, Laura, and Avanidhar Subrahmanyam. "Brand perceptions and the market for common stock." *Journal of financial and Quantitative Analysis* 40.1 (2005): 57-85.

Gervais, Simon, Ron Kaniel, and Dan H. Mingelgrin. "The high-volume return premium." *The Journal of Finance* 56.3 (2001): 877-919.

Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* 457 (2009): 1012-1014.

Gow, Ian D., Gaizka Ormazabal, and Daniel J. Taylor. "Correcting for cross-sectional and time-series dependence in accounting research." *The Accounting Review* 85.2 (2010): 483-512.

Graevenitz, Georg, et al. "Does Online Search Predict Sales? Evidence from Big Data for Car Markets in Germany and the UK." Working Paper (2016).

Granger, Clive WJ, and Paul Newbold. "Spurious regressions in econometrics." *Journal of econometrics* 2.2 (1974): 111-120.

Grossman, Sanford J., and Joseph E. Stiglitz. "On the impossibility of informationally efficient markets." *The American economic review* 70.3 (1980): 393-408.

Grullon, Gustavo, George Kanatas, and James P. Weston. "Advertising, breadth of ownership, and liquidity." *The Review of Financial Studies* 17.2 (2004): 439-461.

Gwilym, Owain Ap, et al. "In Search of Concepts: The Effects of Speculative Demand on Stock Returns." *European Financial Management* 22.3 (2016): 427-449.

Hameed, Allaudeen, Wenjin Kang, and Shivesh Viswanathan. "Stock market declines and liquidity." *The Journal of Finance* 65.1 (2010): 257-293.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld. "Does algorithmic trading improve liquidity?" *The Journal of Finance* 66.1 (2011): 1-33.

Hirshleifer, David, and Siew Hong Teoh. "Limited attention, information disclosure, and financial reporting." *Journal of accounting and economics* 36.1 (2003): 337-386.

Hirshleifer, David, et al. "Do individual investors cause post-earnings announcement drift? Direct evidence from personal trades." *The Accounting Review* 83.6 (2008): 1521-1550.

Hirshleifer, David, Sonya Seongyeon Lim, and Siew Hong Teoh. "Driven to distraction: Extraneous events and underreaction to earnings news." *The Journal of Finance* 64.5 (2009): 2289-2325.

Hjalmarsson, Erik. "Predicting global stock returns." *Journal of Financial and Quantitative Analysis* 45.1 (2010): 49-80.

Hou, Kewei, Wei Xiong, and Lin Peng. "A tale of two anomalies: The implications of investor attention for price and earnings momentum." Working Paper (2009).

Hsiao, Cheng. *Analysis of panel data*. 3rd ed. New York: Cambridge university press, 2014.

Huberman, Gur, and Tomer Regev. "Contagious speculation and a cure for cancer: A nonevent that made stock prices soar." *The Journal of Finance* 56.1 (2001): 387-396.

Huberman, Gur. "Familiarity breeds investment." *The Review of Financial Studies* 14.3 (2001): 659-680.

Joseph, Kissan, M. Babajide Wintoki, and Zelin Zhang. "Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search." *International Journal of Forecasting* 27.4 (2011): 1116-1127.

Kadlec, Gregory B., and John J. McConnell. "The effect of market segmentation and illiquidity on asset prices: Evidence from exchange listings." *The Journal of Finance* 49.2 (1994): 611-636.

Kahneman, Daniel. *Attention and effort*. Vol. 1063. New Jersey: Prentice-Hall, 1973.

Karabiyik, Hande, Simon Reese, and Joakim Westerlund. "On the role of the rank condition in CCE estimation of factor-augmented panel regressions." *Journal of Econometrics* 197.1 (2017): 60-64.

Kihlstrom, Richard. "A general theory of demand for information about product quality." *Journal of Economic Theory* 8.4 (1974): 413-439.

Kumar, Alok. "Who gambles in the stock market?" *The Journal of Finance* 64.4 (2009): 1889-1933.

Latoeiro, Pedro, Sofia B. Ramos, and Helena Veiga. "Predictability of stock market activity using Google search queries." Working Paper (2013).

Lawrence, Alastair, et al. "Yahoo Finance search and earnings announcements." Working Paper (2016).

Leuz, Christian, and Robert E. Verrecchia. "The economic consequences of increased disclosure." *Journal of accounting research* 38 (2000): 91-124.

Li, Xinyuan. “Nowcasting with Big Data: is Google useful in Presence of other Information?” Working Paper (2016)

Lintner, John. "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets." *The review of economics and statistics* 47.1 (1965): 13-37.

Lou, Dong. "Attracting investor attention through advertising." *The Review of Financial Studies* 27.6 (2014): 1797-1829.

McGowan, Michael J. "The rise of computerized high frequency trading: use and controversy." *Duke L. & Tech. Rev.* 16 (2010)

Merton, Robert C. "A simple model of capital market equilibrium with incomplete information." *The Journal of Finance* 42.3 (1987): 483-510.

Newey, Whitney K., and Kenneth D. West. "Hypothesis testing with efficient method of moments estimation." *International Economic Review 28.3* (1987): 777-787.

Nickell, Stephen. "Biases in dynamic models with fixed effects." *Econometrica: Journal of the Econometric Society* 49.6 (1981): 1417-1426.

Nieuwerburgh, Stijn, and Laura Veldkamp. "Information acquisition and under-diversification." *The Review of Economic Studies* 77.2 (2010): 779-805.

Mitchell, Mark L., and J. Harold Mulherin. "The impact of public information on the stock market." *The Journal of Finance* 49.3 (1994): 923-950.

Mondria, Jordi. "Portfolio choice, attention allocation, and price comovement." *Journal of Economic Theory* 145.5 (2010): 1837-1864.

Mondria, Jordi, Thomas Wu, and Yi Zhang. "The determinants of international investment and attention allocation: Using internet search query data." *Journal of International Economics* 82.1 (2010): 85-95.

Peng, Lin. "Learning with information capacity constraints." *Journal of Financial and Quantitative Analysis* 40.2 (2005): 307-329.

Peng, Lin, and Wei Xiong. "Investor attention, overconfidence and category learning." *Journal of Financial Economics* 80.3 (2006): 563-602.

Pesaran, M. Hashem. "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica* 74.4 (2006): 967-1012.

Pesaran, M. Hashem. "A simple panel unit root test in the presence of cross‐section dependence." *Journal of Applied Econometrics* 22.2 (2007): 265-312.

Pesaran, M. Hashem. "Testing weak cross-sectional dependence in large panels." *Econometric Reviews* 34.6-10 (2015): 1089-1117.

Petersen, Mitchell A. "Estimating standard errors in finance panel data sets: Comparing approaches." *The Review of Financial Studies* 22.1 (2009): 435-480.

Phillips, Peter CB, and Hyungsik R. Moon. "Linear regression limit theory for nonstationary panel data." *Econometrica* 67.5 (1999): 1057-1111.

Preis, Tobias, Daniel Reith, and H. Eugene Stanley. "Complex dynamics of our economic life on different scales: insights from search engine query data." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 368.1933 (2010): 5707-5719.

Reese, Simon, and Joakim Westerlund. "Estimation of factor-augmented panel regressions with weakly influential factors." *Econometric Reviews* (2015): 1-65.

Reese, Simon, and Joakim Westerlund. "Panicca: Panic on Cross‐Section Averages." *Journal of Applied Econometrics* 31.6 (2016): 961-981.

Rubin, Amir, and Eran Rubin. "Informed investors and the internet." *Journal of Business Finance & Accounting* 37.7‐8 (2010): 841-865.

Ryan, Paul, and Richard J. Taffler. "Are Economically Significant Stock Returns and Trading Volumes Driven by Firm‐specific News Releases?" *Journal of Business Finance & Accounting* 31.1‐2 (2004): 49-82.

Sarafidis, Vasilis, and Tom Wansbeek. "Cross-sectional dependence in panel data analysis." *Econometric Reviews* 31.5 (2012): 483-531.

Seasholes, Mark S., and Guojun Wu. "Predictable behavior, profits, and attention." *Journal of Empirical Finance* 14.5 (2007): 590-610.

Siganos, Antonios. "Google attention and target price run ups." *International Review of Financial Analysis* 29 (2013): 219-226.

Siganos, Antonios, Evangelos Vagenas-Nanos, and Patrick Verwijmeren. "Facebook's daily sentiment and international stock markets." *Journal of Economic Behavior & Organization* 107 (2014): 730-743.

Sharpe, William F. "Capital asset prices: A theory of market equilibrium under conditions of risk." *The Journal of Finance* 19.3 (1964): 425-442.

Shi, Rongsheng, et al. "Does attention affect individual investors' investment return?" *China Finance Review International* 2.2 (2012): 143-162.

Stambaugh, Robert F. "Presidential address: Investment noise and trends." *The Journal of Finance* 69.4 (2014): 1415-1453.

Tetlock, Paul C. "Giving content to investor sentiment: The role of media in the stock market." *The Journal of Finance* 62.3 (2007): 1139-1168.

Thompson, Robert B., Chris Olsen, and J. Richard Dietrich. "Attributes of news about firms: An analysis of firm-specific news reported in the Wall Street Journal Index." *Journal of Accounting Research* 25.2 (1987): 245-274.

Thompson, Samuel B. "Simple formulas for standard errors that cluster by both firm and time." *Journal of financial Economics* 99.1 (2011): 1-10.

Vlastakis, Nikolaos, and Raphael N. Markellos. "Information demand and stock market volatility." *Journal of Banking & Finance* 36.6 (2012): 1808-1821.

Vozlyublennaia, Nadia. "Investor attention, index performance, and return predictability." *Journal of Banking & Finance* 41 (2014): 17-35.

Westerlund, Joakim, and Jean-Pierre Urbain. "On the estimation and inference in factor-augmented panel regressions with correlated loadings." *Economics Letters* 119.3 (2013): 247-250.

Westerlund, Joakim, and Jean-Pierre Urbain. "Cross-sectional averages versus principal components." *Journal of Econometrics* 185.2 (2015): 372-377.

Westerlund, Joakim, Hande Karabiyik, and Paresh Narayan. "Testing for Predictability in panels with General Predictors." *Journal of Applied Econometrics* 32.3 (2017): 554-574.

Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: The MIT Press, 2009

Yuan, Yu. "Market-wide attention, trading, and stock returns." *Journal of Financial Economics* 116.3 (2015): 548-564.

Appendix A – Google Search Volume

1. Computation of SVI – step by step

There are two steps involved[10]. Let us define the search term as 'VRSK', the country as the US and the period to be between January 2015 and December 2016. First, each week (from Sunday to Saturday) between 2015 and 2016, Google divides the absolute number of searches on 'VRSK' by the sum of all searches that week in the US. For example, assuming 10 million Google searches in the first week of 2015 in the US, 10,000 searches were on 'VRSK', which indicates an interest ratio of 0.1%. In the second week of 2015, 15,000 searches out of 20 million were on 'VRSK', indicating an interest of 0.075%. Therefore, 'VRSK' was more popular in the first week of 2015, even though the second week had a higher absolute number of searches. In a final step, Google determines which week between 2015 and 2016 has the highest interest ratio. If we assume that's the first week of 2015, then 0.1% becomes 100%, and all other weeks are scaled as a proportion. Therefore, the second week of 2015 is assigned the value: (0.075/0.1)*100 = 75%.

2. Overlapping different time series – an example

Consider real values[11] taken on September 28th 2017 for two series on VRSK: May 2008 to September 2012 and January 2012 to December 2015. In the former, SVI is 6 in the week of April 25th and in the week of May 1st. In the latter, SVI is 51 and 38, respectively. We can see there is an inconsistency because the value on April 25th cannot be simultaneously equal and larger than the value next week. This means that the relation between two observations on the same series is not constant in different series (6 = 6 *but* 51 > 38). Therefore, the data points we choose as a scaling factor when merging two series will influence how the new observations are related to each other, which invalidates this method[12]. This problem is clear on appendix B of Li (2016). The researcher shows a value of 64 being re-scaled to 35 and then a value of 72 becomes 32.9. This is problematic because the relation between observations has changed, which would incorrectly influence ASVI.

[10] Source: Google Trends - https://support.google.com/trends/answer/4365533?hl=en

[11] It is impossible to fully replicate past studies because SVI is based on a random sample of Google searches. However, Da et al. (2011) find that correlations are above 97%, indicating this issue has no implications. In addition, repeated searches are eliminated, which avoids double counting.

[12] The difference is not always that large but, in that example, it is clear that the discrepancy is not due to rounding or due to the random sampling method described in the previous footnote.

Appendix B – Stata User-Written Commands

In what follows, I list the Stata user-written commands that were applied in three tests and two regressions presented in this thesis. Note that I did not change or adapt any of their code and some require other (user-written) commands, which I do not mention. They can all be obtained through Stata, except for the last one that is found here: https://sites.google.com/site/judsoncaskey/data

Tests

- **pescadf:** computes the cross-sectionally augmented Dickey-Fuller (CADF) test of an earlier version of Pesaran (2007). Author: Piotr Lewandowski, Warsaw School of Economics, Institute for Structural Research. Last update: 08-October-2007
- **rhausman:** computes the robust version of the Hausman test using a bootstrap procedure. Author: Boris Kaiser, University of Bern. Last update: 07-November-2015
- **xtcdf:** computes the cross-sectional dependence (CD) test of Pesaran (2015). Author: Jesse Wursten, KU Leuven. Last update: 31-July-2017

Regressions

- **xtcce:** implements the common correlated effects mean group (CCEMG) estimator of Pesaran (2006). Author: Timothy Neal, University of New South Wales. Last update: 18-October-2016
- **xtfmbJ:** implements the Fama and MacBeth (1973) two step procedure. Authors: Daniel Hoechle, University of Basel, and Judson Caskey, University of Texas. Last update: 02-May-2007

Appendix C – Robustness Analysis

Table C.1 Hypothesis 1: Robustness Analysis

Each panel presents four regressions where the dependent variable is ASVI. All variables are defined in table 1. Size, analyst coverage, and advertising over sales are adjusted for non-stationarity as in section 4.5. The intercept is omitted because it has no clear interpretation in panel models. Columns (1) to (4) of panel A are the same as column (5) of table 5 with one difference each. In the first three regressions, ASVI is calculated as current week SVI minus the median of the last 4, 13 and 26 weeks, respectively. The fourth regression only considers companies whose ticker symbol has three or four characters, which removes 95 companies out of 739. Panels B and C repeat the analysis of table 5 from May 2008 to August 2012 and September 2012 to December 2016, respectively. Standard errors are shown in parentheses. The data is unbalanced in all samples. *, **, *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

	(1)	(2)	(3)	(4)
Panel A. Changing ASVI window length and excluding noisy tickers				
Log Absolute Abn Ret	0.275* (0.147)	-0.027 (0.138)	0.132 (0.101)	0.061 (0.154)
Log Abn Turnover	7.327*** (0.758)	9.177*** (0.649)	8.656*** (0.557)	9.012*** (0.558)
Log (1 + Bid-Ask Spread)	3.887 (3.132)	5.357 (3.614)	2.090 (3.431)	1.871 (2.980)
Log Size	-0.080 (1.736)	-2.312 (5.458)	9.332** (3.767)	2.144 (1.507)
Analyst Coverage	0.816 (1.724)	0.960 (2.426)	59.319 (54.914)	1.119 (1.658)
Log (1 + Advertising/Sales)	59.395 (97.382)	34.287 (172.729)	81.952 (194.310)	49.894 (181.220)
Earnings Announcement	1.903*** (0.233)	2.001*** (0.199)	1.797*** (0.190)	2.327*** (0.297)
Number of Announcements	-0.106 (0.081)	0.002 (0.122)	0.027 (0.035)	-0.003 (0.015)
$ASVI_{t-1}$	0.064*** (0.009)	0.184*** (0.010)	0.178*** (0.011)	0.145*** (0.009)
$ASVI_{t-2}$	-0.107*** (0.004)	0.028*** (0.008)	0.031*** (0.008)	0.004 (0.005)
$ASVI_{t-3}$	-0.133*** (0.005)	0.002 (0.007)	0.015*** (0.005)	-0.019*** (0.005)
$ASVI_{t-4}$	-0.170*** (0.004)	-0.031*** (0.008)	-0.011* (0.006)	-0.065*** (0.006)
Observations	149,576	142,023	132,073	127,173
CCEMG Averages	YES	YES	YES	YES
Hausman Test	100.62***	106.78***	177.14***	123.44***
CD Test	-4.70***	-4.53***	-4.18***	-5.71***
R^2	20.75%	32.10%	44.30%	29.08%

(continued)

Table C.1 - *Continued*

	(1)	(2)	(3)	(4)
Panel B. May 2008 to August 2012				
Log Absolute Abn Ret	0.121** (0.053)	0.100* (0.054)	-0.021 (0.077)	-0.026 (0.099)
Log Abn Turnover	6.054*** (0.378)	6.271*** (0.402)	7.968*** (0.783)	6.631*** (0.482)
Log (1 + Bid-Ask Spread)	1.842* (0.943)	1.821 (1.181)	3.387 (3.865)	2.603 (4.196)
Log Size	0.047 (0.068)	0.159 (0.283)	8.809 (7.277)	1.577 (2.408)
Analyst Coverage	0.149 (0.152)	0.722** (0.359)	-3.064 (5.548)	9.569 (7.197)
Log (1 + Advertising/Sales)	0.183** (0.092)	-0.573 (0.410)	41.993 (75.335)	2.583 (62.604)
Earnings Announcement	2.964*** (0.241)	2.934*** (0.239)	2.531*** (0.231)	2.278*** (0.357)
Number of Announcements	-0.006*** (0.001)	-0.006*** (0.001)	-0.029 (0.086)	-0.019 (0.047)
$ASVI_{t-1}$				0.270*** (0.014)
$ASVI_{t-2}$				-0.001 (0.010)
$ASVI_{t-3}$				-0.001 (0.011)
$ASVI_{t-4}$				-0.069*** (0.007)
Observations	139,009	139,009	139,009	136,195
Week fixed effects	YES	YES	NO	NO
CCEMG Averages	NO	NO	YES	YES
Hausman Test		149.30***		
CD Test	75.82***	75.62***	-4.35***	-1.96**
R^2	3.88%	4.02%	22.96%	37.72%
Panel C. September 2012 to December 2016				
Log Absolute Abn Ret	0.480*** (0.059)	0.416*** (0.055)	0.323*** (0.066)	0.405*** (0.056)
Log Abn Turnover	9.649*** (0.510)	10.222*** (0.514)	10.671*** (0.593)	9.442*** (0.510)
Log (1 + Bid-Ask Spread)	6.596*** (2.085)	1.433 (8.318)	25.637 (24.628)	59.590* (33.527)
Log Size	0.183** (0.089)	1.383*** (0.466)	8.349*** (2.229)	0.230 (8.428)
Analyst Coverage	0.307** (0.155)	-0.565 (0.497)	1.972 (2.184)	5.522 (5.170)

(continued)

Table C.1 - *Continued*

	(1)	(2)	(3)	(4)
Log (1 + Advertising/Sales)	0.462*** (0.119)	0.317 (0.564)	11.926 (113.740)	-28.316 (60.379)
Earnings Announcement	4.252*** (0.316)	4.173*** (0.312)	3.273*** (0.290)	3.998*** (0.325)
Number of Announcements	0.004* (0.002)	0.003* (0.002)	0.094 (0.075)	-0.007 (0.038)
$ASVI_{t-1}$				0.320*** (0.010)
$ASVI_{t-2}$				0.011* (0.006)
$ASVI_{t-3}$				0.002 (0.004)
$ASVI_{t-4}$				-0.061*** (0.006)
Observations	131,964	131,964	131,964	129,303
Week fixed effects	YES	YES	NO	NO
CCEMG Averages	NO	NO	YES	YES
Hausman Test		157.18***		
CD Test	73.91***	72.46***	-3.21***	0.88
R^2	8.45%	8.83%	31.81%	47.39%

Table C.2 Hypothesis 2: Robustness Analysis

This table reports seven regressions where the dependent variable is ASVI. All variables are defined in table 1. Size, analyst coverage, and advertising over sales are adjusted for non-stationarity as in section 4.5. In the first four regressions, ASVI is calculated as current week SVI minus the median of the last 4, 8, 13, and 26 weeks, respectively. The regression on column (5) only considers companies whose ticker symbol has three or four characters, which removes 95 companies out of 739. All regressions are based on the CCEMG estimator, where cross-sectional averages of all variables are included as additional regressors. The intercept is omitted because it has no clear interpretation in panel models. The sample period in columns (1) to (5) is from January 2004 to April 2008 while for column (6) it is from May 2008 to August 2012 and for column (7) it is from September 2012 to December 2016. Standard errors are shown in parentheses. The data is unbalanced in all samples. *, **, *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log Absolute Abn Ret	0.130 (0.262)	-0.050 (0.105)	-0.158 (0.132)	-0.364 (0.237)	-0.161 (0.180)	0.049 (0.100)	0.125* (0.076)
Log Abn Turnover	7.287*** (0.648)	7.872*** (0.662)	9.166*** (0.770)	7.584*** (0.964)	9.129*** (0.577)	6.438*** (0.660)	9.640*** (0.551)
Log (1 + Bid-Ask Spread)	-0.077 (6.805)	3.787 (9.987)	1.767 (5.427)	6.020* (3.213)	-1.561 (4.687)	-10.221 (14.599)	-14.781 (39.712)
Log Size	6.256 (16.013)	78.957 (48.727)	69.992 (74.581)	32.776* (16.902)	34.447 (32.296)	59.426 (44.827)	-85.021 (117.805)
Analyst Coverage	0.506 (2.890)	12.224 (21.285)	-4.256 (5.321)	2.284 (4.293)	-0.028 (3.162)	-0.875 (3.584)	-5.440 (9.371)
Log (1+ Advertising/Sales)	-30.199 (47.947)	25.347 (68.168)	-67.949 (65.803)	-24.250 (31.089)	12.218 (42.605)	-16.282 (25.070)	26.600 (18.420)
Earnings Announcement	2.366*** (0.424)	1.473*** (0.223)	1.614*** (0.323)	1.543*** (0.276)	2.221*** (0.371)	2.548*** (0.256)	3.604*** (0.271)
Number of Announcements	0.199** (0.096)	0.191 (0.173)	0.008 (0.042)	0.002 (0.033)	0.015 (0.027)	0.024 (0.034)	-0.073 (0.119)
Idiosyncratic Skewness	0.566* (0.300)	1.203* (0.688)	0.744*** (0.255)	-0.107 (0.567)	0.688** (0.279)	-0.003 (0.160)	0.287 (0.215)

(continued)

Table C.2 - *Continued*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log 1/P	-2.104 (15.926)	91.956* (54.464)	61.116 (76.045)	18.182 (15.548)	28.843 (32.703)	52.350 (40.865)	-87.544 (113.986)
Log (1 + Absolute Earnings Surprise)	0.702 (0.884)	-0.381 (1.409)	-1.019 (0.968)	0.518 (1.537)	-0.728 (0.658)	4.520 (5.315)	-1.501 (9.882)
Consumer Sector x Advertising/Sales	14.003 (11.852)	8.083 (7.486)	12.073 (8.460)	18.161* (10.241)	3.976 (3.026)	-10.810 (11.452)	-6.078 (4.728)
Log IVOL	-3.197 (2.729)	-4.401 (5.791)	-2.095 (2.313)	4.427 (4.054)	-0.754 (2.216)	0.754 (1.690)	-1.949 (3.976)
Change Log IVOL	11.315*** (2.628)	10.410** (4.736)	5.691** (2.674)	9.409*** (2.998)	8.852*** (2.699)	8.306*** (1.908)	10.446*** (1.658)
Observations	149,576	145,861	142,023	132,073	127,173	136,195	129,303
CCEMG Averages	YES	YES	YES	YES	YES	YES	YES
$ASVI_{t-1,\ldots,t-4}$	YES	YES	YES	YES	YES	YES	YES
R^2	26.03%	32.38%	37.88%	45.00%	34.41%	42.88%	52.33%

Table C.3 Hypothesis 3: Robustness Analysis

Each panel presents five regressions where the dependent variable is the future Fama French 3 factor abnormal return (in basis points). All variables are defined in table 1. Size, analyst coverage, and advertising over sales are adjusted for non-stationarity as in section 4.5. The CCEMG estimator is used, where cross-sectional averages of all variables are included as additional regressors. The intercept is omitted because it has no clear interpretation in panel models. The regressions on panel A are based on next week's abnormal return. In the first four regressions, ASVI is calculated as current week SVI minus the median of the last 4, 8, 13 and 26 weeks, respectively. ASVI in the last regression is based on the previous 13 weeks and only considers companies whose ticker symbol has three or four characters, which removes 95 companies out of 739. Panels B, C, and D repeat the analysis of table 7 by, respectively, excluding noisy tickers and considering the period from May 2008 to August 2012 and from September 2012 to December 2016. Standard errors are shown in parentheses. The data is unbalanced in all samples. *, **, *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

	Week 1	Week 1	Week 1	Week 1	Week 1
	(1)	(2)	(3)	(4)	(5)
Panel A. Changing ASVI window length					
ASVI	0.956	1.394***	-7.687	3.936	0.913*
	(0.842)	(0.481)	(8.823)	(2.507)	(0.488)
Log Size x ASVI	-0.910	-1.186**	7.489	-4.005*	-0.730
	(0.657)	(0.554)	(8.447)	(2.133)	(0.473)
Log Size	-10.83***	-11.41***	-10.417***	-12.127***	-10.499***
	(0.499)	(0.500)	(1.058)	(0.520)	(0.393)
Change Log	0.088	-0.041	-0.176	0.149	0.036
IVOL x ASVI	(0.065)	(0.086)	(0.270)	(0.094)	(0.054)
Change Log	-0.063**	-0.015	-0.018	-0.004	-0.027
IVOL	(0.025)	(0.028)	(0.046)	(0.034)	(0.023)
Log Absolute Abn	-0.011	-0.014	-0.168	-0.031	-0.011
Ret	(0.020)	(0.017)	(0.142)	(0.036)	(0.018)
Log Abn	0.000	0.001	-0.060	0.041	0.003
Turnover	(0.032)	(0.031)	(0.052)	(0.035)	(0.029)
Log (1 + Bid-Ask	-0.034	-0.014	0.177	-0.080	0.009
Spread)	(0.027)	(0.034)	(0.197)	(0.073)	(0.031)
Analyst Coverage	0.235	0.189	1.094	38.136	0.562**
	(0.302)	(0.299)	(0.927)	(37.066)	(0.285)
Log (1 +	8.536	4.512	-18.669	12.793	-28.488
Advertising/Sales)	(13.752)	(13.262)	(16.371)	(57.028)	(35.674)
$ASVI_{t-1}$		0.001			
		(0.003)			
Observations	151,280	147,429	144,354	134,380	125,867
CCEMG Averages	YES	YES	YES	YES	YES
Hausman Test	410.69***	145.70***	205.10***	277.96***	288.93***
CD Test	0.825	0.729	-0.985	0.788	0.514
R^2	19.12%	23.00%	31.40%	28.83%	22.14%

(continued)

Table C.3 - *Continued*

	Week 1 (1)	Week 2 (2)	Week 3 (3)	Week 4 (4)	Week 5-52 (5)
Panel B. Excluding noisy tickers					
ASVI	1.457** (0.671)	-0.190 (0.374)	-1.466 (1.429)	-0.431 (0.389)	-3.233** (1.557)
Log Size x ASVI	-1.241** (0.570)	0.037 (0.451)	2.893 (2.740)	0.410 (0.372)	2.418** (1.218)
Log Size	-11.35*** (1.141)	-8.348*** (0.335)	-8.583*** (1.043)	-6.119*** (0.303)	-1.646 (1.988)
Change Log IVOL x ASVI	0.060 (0.098)	0.031 (0.113)	-0.372 (0.269)	0.058 (0.065)	0.127 (0.181)
Change Log IVOL	0.112 (0.126)	0.003 (0.024)	0.033 (0.052)	0.082 (0.070)	-0.030 (0.074)
Log Absolute Abn Ret	0.019 (0.054)	-0.030* (0.017)	-0.017 (0.042)	-0.048 (0.031)	-0.099 (0.066)
Log Abn Turnover	-0.107 (0.119)	0.018 (0.046)	-0.007 (0.035)	0.004 (0.023)	-0.299** (0.142)
Log (1 + Bid-Ask Spread)	-0.002 (0.028)	0.007 (0.034)	-0.002 (0.065)	0.047 (0.032)	0.012 (0.142)
Analyst Coverage	0.517* (0.275)	0.330 (0.265)	0.836** (0.336)	0.374** (0.178)	2.072 (1.775)
Log (1 + Advertising/Sales)	-20.748 (18.486)	-10.603 (32.390)	-4.742 (26.868)	-14.352 (21.381)	-145.291 (164.697)
Observations	129,201	128,561	127,921	127,281	126,641
CCEMG Averages	YES	YES	YES	YES	YES
Hausman Test	154.76***	536.61***	431.40***	723.59***	111.70***
CD Test	-1.314	0.538	-0.983	-0.879	1.827**
R^2	29.42%	25.18%	24.52%	23.68%	29.22%
Panel C. May 2008 to August 2012					
ASVI	1.197 (0.827)	-1.762 (3.130)	2.872 (2.727)	-2.603 (3.224)	-1.512 (1.607)
Log Size x ASVI	-0.380 (0.779)	0.641 (4.152)	-3.898 (3.484)	0.883 (2.468)	-0.103 (1.756)
Log Size	-17.80*** (1.116)	-10.636** (4.717)	-12.301** (4.939)	-9.318*** (0.959)	4.155* (2.222)
Change Log IVOL x ASVI	0.038 (0.155)	-0.321 (0.283)	-0.114 (0.199)	0.850 (0.837)	0.239 (0.351)
Change Log IVOL	0.072 (0.091)	-0.067 (0.193)	0.187 (0.193)	-1.108 (0.865)	-0.108 (0.217)
Log Absolute Abn Ret	-0.045 (0.028)	0.011 (0.071)	0.077 (0.145)	0.410 (0.401)	-0.112 (0.196)
Log Abn Turnover	-0.005 (0.060)	-0.10911	0.224 (0.261)	0.339 (0.453)	-0.522 (0.439)

(continued)

Table C.3 - *Continued*

	Week 1 (1)	Week 2 (2)	Week 3 (3)	Week 4 (4)	Week 5-52 (5)
Log (1 + Bid-Ask Spread)	0.778 (0.558)	-0.361 (0.834)	-0.627* (0.354)	-0.277 (0.432)	1.008** (0.493)
Analyst Coverage	-0.105 (0.329)	1.413 (2.042)	-2.336 (2.565)	3.477 (3.978)	0.262 (2.712)
Log (1 + Advertising/Sales)	-24.129 (130.082)	-59.069 (100.930)	29.522 (19.882)	-69.656 (56.368)	77.388 (159.750)
Observations	138,324	137,641	136,959	136,277	135,561
CCEMG Averages	YES	YES	YES	YES	YES
Hausman Test	525.14***	184.75***	330.99***	440.37***	108.04***
CD Test	6.24***	5.32***	4.52***	5.34***	4.80***
R^2	30.78%	24.29%	23.03%	22.27%	29.25%
Panel D. September 2012 to December 2016					
ASVI	0.293 (0.353)	0.397 (0.314)	0.859* (0.504)	-0.384 (0.500)	-6.182* (3.418)
Log Size x ASVI	-0.324 (0.398)	-0.408 (0.351)	-0.231 (0.604)	-0.022 (0.608)	5.023 (3.528)
Log Size	-12.59*** (0.478)	-9.557*** (0.541)	-8.277*** (0.595)	-7.402*** (0.388)	3.263 (2.180)
Change Log IVOL x ASVI	-0.036 (0.051)	0.104 (0.095)	0.096* (0.052)	-0.032 (0.065)	-0.152 (0.253)
Change Log IVOL	-0.020 (0.021)	-0.007 (0.025)	-0.065* (0.037)	-0.028 (0.022)	0.002 (0.056)
Log Absolute Abn Ret	-0.027** (0.013)	-0.007 (0.017)	0.007 (0.017)	0.005 (0.015)	0.035 (0.053)
Log Abn Turnover	0.003 (0.020)	0.009 (0.017)	0.006 (0.021)	0.028 (0.018)	-0.713*** (0.126)
Log (1 + Bid-Ask Spread)	0.792*** (0.282)	0.224 (0.487)	0.445** (0.196)	0.660* (0.342)	0.194 (0.830)
Analyst Coverage	0.190 (0.455)	0.179 (0.653)	0.660 (0.584)	-0.244 (0.225)	-1.463 (1.749)
Log (1 + Advertising/Sales)	28.231* (16.232)	37.415 (23.025)	40.812*** (12.842)	27.216* (14.166)	-10.780 (57.457)
Observations	131,291	130,655	130,019	129,383	128,747
CCEMG Averages	YES	YES	YES	YES	YES
Hausman Test	184.40***	211.77***	198.18***	272.95***	418.71***
CD Test	7.57***	9.30***	7.61***	7.54***	8.16***
R^2	22.40%	17.79%	16.95%	16.92%	29.62%

Appendix D – Preliminary Thesis

(Remainder of page intentionally left blank)

João Pancada - 1002939

Preliminary Master Thesis

Online search intensity and stock returns: is there a new proxy for investor attention?

Hand-in date:
01.03.2017

Campus:
BI Oslo

Examination code and name:
GRA 19502 Master Thesis

Programme:
Master of Science in Financial Economics

“This thesis is a part of the MSc programme at BI Norwegian Business School. The school takes no responsibility for the methods used, results found and conclusions drawn."

Abstract

In this thesis I propose to test the validity of online search frequency data as a proxy for investor attention. Search frequency is given by Google's Search Volume Index (SVI), which measures the regularity with which internet users look for any specific words through Google's search engine. Using data on S&P 500 stocks from 2004 to 2016, I expect to (1) corroborate previous studies on the SVI's power to capture (retail) investor attention; (2) find a strong relationship not only with returns but also with (idiosyncratic) volatility; (3) help explain a famous anomaly - the index effect - as set forth by previous researchers.

Introduction

We are currently living in the information age. The internet, and subsequent innovations, has allowed a myriad of news and data to spread almost instantly across the world. In a globalized economy, it is important not to miss this constant information flow because businesses and countries are more interconnected than ever. Investors are perhaps the ones benefitting the most from being regularly updated since financial markets are very sensitive to new information. In the limit, as traditional asset pricing models usually posit, asset prices reflect all available information, with new one immediately incorporated once it is made public (also known as semi-strong form of market efficiency).

However, there is only so much information one can process each day that investors need to be selective. Consequently, their attention is limited and such models fail to account for this added complexity. This can have profound implications for the efficiency of financial markets because, depending on the case, the time lag will be greater or smaller. For instance, Dellavigna and Pollet (2009) postulate that on Fridays, when investors tend to be more distracted, there is a more pronounced post-earnings announcement drift due to greater initial under reaction. In another study, Chen, Noronha, and Singal (2004) argue that increased investor awareness can help explain the index effect anomaly.

The problem is that investor attention cannot easily be measured. Although proxies abound in the literature[1], they are unsatisfactory because they fall in either one of two categories: (1) supply side of information (e.g. news and advertising expenses); (2) market data (e.g. extreme returns and turnover). None of them measure the changes in intensity with which investors try to collect data or read news, i.e., the demand for information.

The aim of this thesis is to examine the validity of a 'demand' proxy as well as to expand its applicability, at a time when there is much more data available. Da, Engelberg and Gao (2011) were the first to introduce Google's Search Volume Index (SVI) to measure investor attention. SVI is made available to the public via Google Trends (https://www.trends.google.com) since 2004. When a user inputs a search term into Google Trends, the application returns the search volume history

[1] See, for example, Gervais, Kaniel, and Mingelgrin (2001), Barber and Odean (2008), Yuan (2008), and Chemmanur and Yan (2009).

for that term scaled by the time-series maximum (a scalar). As an illustrative example, figure 1 plots the weekly SVI for […].

There are two good reasons to believe *ex-ante* that using aggregate search data from Google is a good proxy. First, most people start searching for online information via a web search engine[2]. Google continues to dominate the US (and world) market, with a market share of 72% in 2015[3]. Therefore, it is expected that Google represents the general search behaviour in the US and other countries. Although the SVI may capture the behaviour of everyone, with the proper selection of keywords one can expect a high probability of most queries coming only from (potential) investors, as explained in the data section.

In addition, search frequency directly reveals how much attention individuals are paying to certain events (or companies) over time.

The contribution of this thesis is twofold: in terms of the data used and extending the study. First, in terms of data, there is now 13 years of weekly observations, rather than the 4.5 years the authors used, which allow for more robust conclusions. I will also use a more comprehensive method to identify companies by keywords in SVI, as described in the data section. In addition, Google has changed the method of determining SVI, from an absolute measure to a relative one. It is, thus, important to assess whether this compromises altogether the results of previous studies or actually has the potential to improve them.

Second, I will explore the potential of the SVI in other areas, namely on its relationship with (idiosyncratic) volatility as well as whether it can help explain the well documented index effect. Please refer to the medothology section for a detailed explanation.

[2] Source:

[3] Source: http://returnonnow.com/internet-marketing-resources/2015-search-engine-market-share-by-country/

Literature Review

1. Investor attention
   a. E.g.: Considered a scarce cognitive resource (Kahneman, 1973),
2. Retail investor behaviour in stock market + Sentiment
3. Long before revealed behaviour proxies for investor attention were available, a full body of literature delved into the issue. Without trying to go into much detail, the literature can be divided, in my opinion, into two distinct categories. […]
4. Search induced proxies were not a novel with the SVI. Before that, researchers had already tried to use other means, such as online blogs and Wikipedia, to gauge revealed search behaviour by internet users
   a. Blogs: Antweiler and Frank (2004)
   b. Wikipedia: Rubin and Rubin, 2010
   c. AOL: Mondria et al. (2012)
5. SVI – who first introduced SVI in general?
   a. […]
6. And then for financial markets?
   a. Mondria and Wu 2012
   b. Joseph, Wintoki and Zhang 2011
   c. Smith 2012
   d. Bank, Larch and Peter 2011
   e. Da et al. 2014
7. (Idiosyncratic) Volatility and returns
8. Index effect: To the best of my knowledge, this hypothesis remains untested
   a. Chen, Honghui, Gregory Noronha, and Vijay Singal. "The price response to S&P 500 index additions and deletions: Evidence of asymmetry and a new explanation."

Methodology

This thesis aims at exploring the following questions:

Relationship between SVI and other proxies

Motivation: This should be the very first thing to present because it is pointless to propose a proxy that can be explained by the other ones that already exist in the literature b/c the results would be the same.

How, why this way and expectations:

- Simple correlation → > 0 but low
    - Correlation between log SVI and ASVI and other measures. First compute for each stock with a minimum of 1 year data and then average across stocks.
- Run a VAR for each stock with at least 2 years of weekly data based on 4 measures with weekly frequency (lagged 1 week).
    - Include both a constant and a time trend in the VAR
    - Average the coefficients across stocks and show P-Value.
    - P-value is based on block bootstrap
    - Regress many variables on ASVI → expect low R^2

Capturing the attention of retail investors?

Motivation: This is important because it is one of the main limitations of the proxy (the other is only having weekly data and a short time span, even if 3x times larger than Da). Professional or institutional investors have access to better and more comprehensive data providers such as Bloomberg or Reuters. Therefore, retail investors, who do not have the financial means to purchase such services or a portfolio big enough to justify it, will most likely resort to Google.

How and why this way: By checking whether search frequency is significantly higher for stocks with lower institutional ownership. Problem: because most US companies have a very high float, maybe I'll not be able to prove that.

Expectation: yes

Investor attention theories

1. <u>Barber and Odean (2008)</u>
    a. Motivation: use SVI to test their attention theory
    b. How and why this way: […]
    c. Expectation: theory should hold. After proving the theory they conclude that an increase in SVI for Russell 3000 stocks predicts higher stock prices in the next 2 weeks and an eventual price reversal within the year.
2. <u>However, in my opinion, two issues arise:</u>

   a. SVI should be firstly related with volume activity and only then, through the relationship between volume and returns documented in many papers (e.g. Gervais, Kaniel, and Mingelgrin, 2001), it is related with returns. This is because a stock with more attention should have as a first effect more trading activity.
   b. The question of causality. What's driving the high SVI? Is it news/rumours about the company or is it high stock returns? The paper about time series momentum says that past returns predict future returns. So, is there an endogenous relation between SVI and returns?
3. Are there any other hypothesis/theories in the literature?

SVI and stock volatility

Motivation: Apart from returns, shouldn't we also see a relationship with volatility of individual stocks? At least, Dimpfl and Jank (2011) suggest a relationship with stock market volatility. In addition, shouldn't it be stronger for IVOL? Because retail investors in general hold undiversified portfolios, idiosyncratic volatility is priced (many papers document this).

Explaining anomalies - Index effect

Motivation: Does the SVI contribute to the explanation of the Index effect? The reason is that, among many hypothesis suggested to explain the Index effect, there is one that relates with investor awareness (Chen, Noronha, and Singal, 2004). Therefore, SVI is a very good candidate to test it.

Expectation: Given that their theory is intuitive, the SVI should not only confirm their results but also provide stronger ones because SVI should be a stronger proxy than the ones the authors use.

How and why this way: check abnormal SVI before addition, at the time of announcement, and afterwards. Then I should compare my results with the paper's ones. Two potential problems: a) clean the data on index deletions, because that paper has to deal with many issues; b) only test from 2004-2016 so the sample is rather small and it is not the same as they have (but at least is the same index).

Plan

1. When are index changes announced? Should have "a surge in public attention" that week.
2. Or maybe 1 or 2 weeks before because of rumors? Since it may be easy to pin point which company will be added or removed.

3. Have to see what happens in the week of actual change -> expect no surge despite high stock and volume activity due to rebalance mainly from index funds?
4. Compare the increase in SVI cross companies – do the companies with more attention have higher returns?
5. We should also see attention being constantly higher than before for additions but not the reverse for deletions

Data

Search frequency is available since January 2004 and with weekly frequency on Google Trends. Thus, weekly SVI data is obtained for each of the companies that constituted the S&P 500 index at any moment in time from January 2004 to January 2017. Examining all stocks that were part of the index during that period guarantees that the results are not influenced by survivorship bias. Data on each company (tweekly returns, volume, institutional ownership, etc.) was obtained through […]

The choice of the S&P 500, rather than the Russel 3000 as in Da, Engelberg and Gao (2011), is due to its much higher visibility in the news and importance to market participants. Since it is one of the most followed indexes in the World[4], new information about its constituents is constantly being produced. This reduces the issue of missing SVI data, which occurs when a given word is searched so little that Google assumes a zero value. In addition, the fact that search volume is, in general, high for all firms does not create a problem of lack of heterogeneity in investor attention. This is because in Trends, as explained in the introduction, each week's value is measured relative to the week with the highest search volume.

The critical decision in data collection is which keywords should be used to identify investor interest in any given company. Simply writing the company's name is problematic because it can be used for consumption purposes (e.g., "Walmart") or the name has different meanings (e.g., "Amazon").

A much better keyword is the company's ticker symbol because it is unique to each stock and often times the combination of letters is so random that only someone interested in the stock would type it (e.g., XRX for Xerox). However, there are in fact tickers with everyday meanings such as "FOX", "PH", and "R", which could contaminate the inference of investor behaviour for the associated companies. To

[4] Source:

avoid subjectivity, the research is conducted with the whole sample, but I aim to confirm the results after excluding some “problematic” tickers.

The issue with using the ticker is that it seems too restrictive. Many (potential) investors do not know the ticker of the company they are interested in and so write instead, e.g., “Apple Yahoo”. The idea is to combine different kinds of searches for the same company, through principal component analysis (PCA) just like in Rao and Srivastava (2013).

In order to obtain from Trends weekly observations for all stocks, a computer program will be used to automatically type in the pre-selected keywords and then download the data.

References

Baker, Malcolm, and Jeffrey Wurgler. "Investor sentiment in the stock market. "The Journal of Economic Perspectives 21.2 (2007): 129-151.

Barber, Brad M., and Terrance Odean. "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. "Review of Financial Studies 21.2 (2008): 785-818.

Chen, Honghui, Gregory Noronha, and Vijay Singal. "The price response to S&P 500 index additions and deletions: Evidence of asymmetry and a new explanation." The Journal of Finance 59.4 (2004): 1901-1930.

Da, Zhi, Joseph Engelberg and Pengjie Gao. 2011. "In search of attention." The Journal of Finance 66 (5): 1461-1499.

Dimpfl, Thomas, and Stephan Jank. "Can internet search queries help to predict stock market volatility?." European Financial Management 22.2 (2016): 171-192.

Gervais, Simon, Ron Kaniel, and Dan H. Mingelgrin. "The high-volume return premium." The Journal of Finance 56.3 (2001): 877-919.

Rao, Tushar, and Saket Srivastava. "Modeling movements in oil, gold, forex and market indices using search volume index and Twitter sentiments." Proceedings of the 5th Annual ACM Web Science Conference. ACM, 2013.