NORWEGIAN SCHOOL OF ECONOMICS NHH

Google search as a measure investor attention: Its influence on Stocks and IPOs in U.S.

Master thesis within the profile of International Business

Bergen, Autumn 2014

Tallal Nawaz

Thesis Supervisor: Tyler J Hull

This page is intentionally left blank

Abstract:

The purpose of this thesis is to investigate new method of retail investor attention by using Google Search Volume Index (SVI) in determining its effect on stocks trading activity, volatility and extend the analysis by incorporating first-day IPO returns. I perform three unique analyses to gauge the relevance of Google search Index as a measure of investor attention. Firstly I examine S&P-500 index overall. Secondly, I measure the effects of SVI on week ahead volume traded on all the stocks listed on Dow30 index. Finally I extend my analysis to include IPOs by examining the effect of SVI on first-day IPO returns. The data used is taken from the start of the year in 2006 till the end of the year in 2013 except for the IPO analysis. The results obtained in this study indicate that Google search query is a reliable proxy to measure attention of retail investors in the short-run. An increase in search queries is associated with a rise in trading activity, volatility and higher first-day IPO returns.

Preface

All the data and the results can be made available upon request at tallal.nawaz@gmail.com.

Acknowledgement

I would take this opportunity to thank a lot of people who helped me in filtering the data sets as well as in guiding me to use the most suitable modeling techniques. Some of the most important names I would like to thank are Professor Thomas Dimpfl, Tyler J Hull, Muhammad Bilal and Hassan Mazahr Gooreja who were kind enough to reply to my queries in the writing process and guiding me in using different statistical measurement tools. I would also like to thank youtube and Coursera for the massive educational training regarding econometric modeling courses, taught in an easy to understand way and also helped me in using Stata, without which I think I would not be able to get any empirical result as output on this paper. I also would like to thank some senior members of Stata user group on Facebook who guided me.

Table of Contents

1. Introduction: 5
1.10 Information and Attention: 5
1.20 Google search as a measure of attention: 6
1.3 Google search and IPO: 8
1.40 Research Focus and findings: 9
1.5 Structure of the paper: 10
2. Literature review: 11
3.0 Sample and Data explanation 13
3.10 Google Search Volume Index: 13
3.20 Capital market data 14
3.30 IPO data 15
4.0 Influence of SVI on S&P-500 16
4.10 Research methodology 16
4.20 Realized volatility: 17
4.30 Descriptive statistics of S&P-500: 18
4.40 Vector Auto Regressive Model – (VAR) 20
4.50 Dynamics of Google Searches and Volatility 21
5.0 Dow Jones Industrial average 28
5.10 Dynamics of Searches and Volume 28
5.20 Google search volume and trading activity 29
6.0 Initial public offering and Google searches 32
6.10 Dynamics of Searches in an IPO 32
6.20 Results 33
6.30 Short conclusion and limitations: 37
7 Discussion of Findings and Concluding Remarks 0
References 1
Appendix 5

1. Introduction:

1.10 Information and Attention:

"What information consumes is rather obvious: it consumes the attention of its recipients. Hence, a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Herbert Simons, 1971 (Herbert 1971)

The first stock market in the world was in Rome, the Roman publicani was established in the second century BC. Bids were placed for the public contract such as tax collection and temple building. Making an informed decision required knowledge of fundamentals and for that purpose Publiciani's enlisted a large group of couriers who traveled throughout the Roman Empire to collect information from locals about their need for public services. Inorder to meet the local demands of public goods a contract with the required information would be drawn up by the local government and that contract was then sent off through the couriers. The publicani's then engaged in bidding process where the one with the lowest cost in completing the public work would be awarded the public works contract (publican 2014. n.d.). Two thousand years later the process of information aggregation has changed significantly but the fundamental importance of information in pricing of the equities and assets remains the same. The process of information delivery has evolved from physical delivery of messages from scribes to digital medium evolving from telegraph to telephone and finally the internet. In modern age of information the use of internet has become vital in businesses, for information sharing and searches. In the process of sharing information access to such information requires information to be searchable. For this purpose significant resources were spent on designing a perfect search engine which was able to scrawl the web and index searches based on the relevancy and volume of traffic (for a detailed description of how a search engine works see Appendix 1 Figure 1).

1.20 Google search as a measure of attention:

Today Google and its sister websites such as Yahoo, Bing and Ask dominate the web search market and account for more than 95% of searches [1] (Welle 2014) . The tool of information search today is the web query engine. The intuition of using search queries to collect information is simple. Search queries reflect the intention of those who search for information. Therefore, when aggregate search volume for a particular product is high, demand for that product is likely to be high (Da, Engelberg and Gao, In search of attention. 2011). Google's Chief Economist Hal Varian along with Choi, 2012 in their paper have proved that search queries have the potential to describe short term interest in a variety of economic activities in real time. They support their claim by providing evidence that search data can predict unemployment claims, automobile sales, travel planning and consumer confidence (Varian and Choi 2012). Traditional asset pricing models assume that prices reflect all available information and that all new information is instantaneously incorporated into asset's prices. These assumptions entail that investors allocate sufficient attention to the asset. In reality however, attention is a scarce cognitive resource (Kahneman 1973), and investors have to be selective in their information processing.

There have been several studies linking search queries in assessing their predictive power. In recent times the importance of search queries has been linked to the attention bias of individual investors. Some proxies for investor attention are, trading volume (Barber and Odean 2008), extreme returns (Barber and Odean 2008), (Mingelgrin, Gervais and Kaniel 2001), stocks in the news (Wall Street Journal) and headlines (Yuan 2008), (Tetlock 2007), advertising expense (Chemmanur and Yan 2009) , and the daily difference in bid-ask spreads (Seasholes and Wu 2007). These variables make a critical assumption that if there is a sudden surge in stock's return or if the traded volume was extreme or if there was news headline, it would only be possible if investors have paid attention to the stock. However, the returns can be driven by

[1] The data pertaining to search index is representative of the rest of the world excluding China. Search engines such as Google, Bing, Yahoo, AOL and Ask are not represented in China. Chinese market is dominated by domestic and government controlled search engines such as Alibaba. For the rest of the world the total searches in 2014 are Google had 1.1bilion estimated monthly visitors, Bing 300m, Yahoo! Search 250m, Ask 145m and AOL Search 70m (Welle 2014).

factors that are unrelated to attention, an article in the media or news headline does not guarantee attention unless investor actually reads it. This is especially true in the modern information age where a wealth of information has created a poverty of attention (Da, Engelberg and Gao, In search of attention. 2011).

In this paper I will introduce a new proxy of measuring investor's attention using aggregate search frequency and then establish the association between Google search volume index as measure of attention and its influence on financial markets. There are several reasons for utilizing Google search volume index as a measure of attention. First, Google is the largest source for search queries worldwide. Indeed, as of March 2014, Google accounted for 70 % of all search queries performed in the United States (Comscore 2014). Thus, the search volume reported by Google is thus likely to be representative of the search behavior of the general population. Secondly, search is a revealed attention measure it directly measures attention, if someone searches for a stock on Google that person is directly paying attention to that stock. Therefore, Google search volume index as a measure of attention is an unambiguous and direct measure of attention. Thirdly, there is strong empirical evidence that has demonstrated the predictive power of Google searches and how it can be used to forecast anything from tourism, automobile sales to influenza (Ginsberg, et al. 2009) .

Google's Search Volume Index (SVI) of search terms is public via the product Google Trends[2]. To compare the search data, results in Google Trends are normalized. The term normalized means that sets of search data are divided by a common variable, like total searches, to cancel out the variable's effect on the data (Google 2014). Panel A of Figure 1-1 plots the weekly SVI of two search terms "Costco" and "Kmart" from January 2004 to April 2014. SVI appears to capture attention well. The SVI for both Costco and Kmart seems to increase during the Christmas holiday season and spikes at some days before the Christmas date consistent with the notion that spending spree starts just before the holidays (November and December) and peaks near the holiday date. The trend is consistent with all the major retailers. This example is

[2] Through Google trends data of information of query terms can be obtained from the Google search Index http://www.google.com/trends. The google search algorithm has been modified countless times by Google Inc to incorporate optimizing the google search Index. Therefore the relative nature of the SVI (Search Volume Index has also changed. The major change regarding the storage of google search queries was done in 2007. In their note google describes the change as a percentage of Google web searches to determine how many searches have been done for the terms compared to the total number of Google searches done during that time. Source: Google Trends (www.google.com/trends).

to illustrate the real time implication of search queries (as a measure of attention) on prices and markets (both retail and financial).

1.3 0 Google search and IPO:

This paper will further explore attention by investors on IPO's as discussed by (Loughran & Ritter, 1995, 2002), among many others. Attention on IPOs induces abnormal returns in the short run followed by long-run return reversal. Moreover before the IPO the attention proxy, in terms of Google search volume index (which from now onwards is going to be referred to as SVI only) shows high search queries for companies' name. Search queries of IPO's are important because the search volume exists prior to the IPO while other trading-based measures, such as do not. Search volume index offers the opportunity to study the effect of Google search volume index in predicting the future price of the stock.

Figure 1-2 shows the SVI plot of Facebook searches. The first peak at the end of the year 2011 was when there were rumors in the media about a potential IPO of Facebook. The second peak at the beginning of the year 2012 highlights the attention time frame when Facebook officially released the news about going public to the media, the third and the highest spike mark the time frame when Facebook unveiled its stocks for trading in open markets. The price of the stock increased to $45 a share from $38 in the short run as soon as the IPO-stock was eligible to trade on the market (Google Corporation 2012).

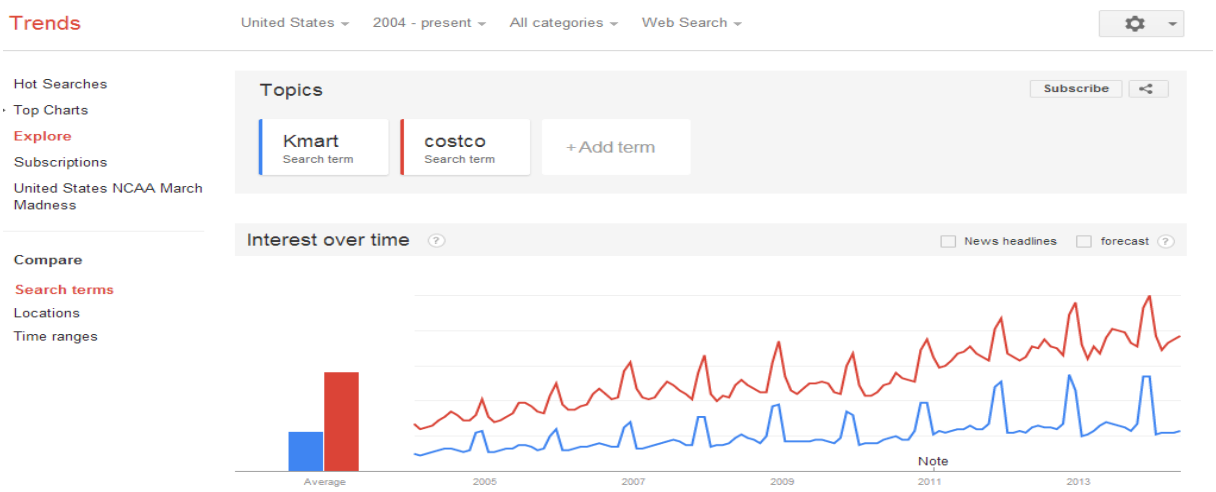

Figure 1-1 Graph shows Google search trend for Kmart and Costco and peak searches made during November and December holiday season.

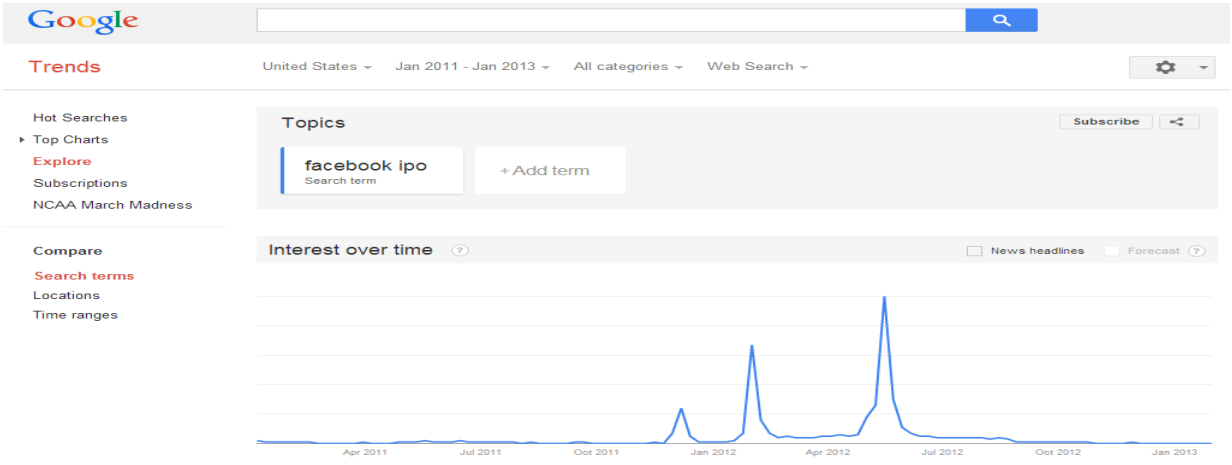

Figure 1-2: Search trend for the term "Facebook IPO"

1.40 Research Focus and findings:

This paper will present three different systems and present empirical estimates to examine if SVI does have an influence on the financial markets and vice versa that is if the rapid changes in capital markets will affect future Google searches. In the first section of the analysis I have used S&P-500 stock index to determine if there is a relationship between Google searches and market volatility and vice versa. The results indicate that future Google searches are in fact determined by previous week's market volatility and current week's market volatility is influenced by the

previous week's SVI, the influence is only valid for searches with a lag of 1 week. In the second chapter of the analysis I have taken the largest firms on the NASDAQ index (the DOW-30 index) and for each stock I have performed a regression estimate to determine if the current weeks traded volume of stock was determined by previous weeks Google searches, I have used lagged volume, Dow-30 market return and realized volatility as control variables. The findings in this section also reinforce the hypothesis that Google searches do have an influence on the stock market. In the third chapter of the analysis I test weather Google searches do in fact have an effect on the first-day IPO returns controlling for all other IPO related factors. Here too I find that Google searches do have a very significant influence on IPOs first day return. However, media seems to have an overall much higher influence on the first-day IPO returns. This study contributes to the finance literature by demonstrating that there are significant linkages between Google searches as an attention tool and its influence on the financial markets. The study is unique since it incorporates different relevant econometric models and data sets to prove the hypothesis.

1.5 Structure of the paper:

The thesis is divided into seven chapters. Chapter 1 provides a brief introduction to the research topic and background as well as the importance of the research topic. Chapter 2 will present a literature review of existing research on the topic those academic papers that have used Google Search Volume as a measure of attention by investors. Chapter 3 will describe the data to be used and describe the method employed to obtain and refine the data for the empirical study in the latter chapters. Chapter 4, 5 and 6 are important chapter because much of the hypothesis testing is done in these chapters. In the chapter 4, I have examined the relationship between the volatility of S&P-500 index and SVI using the VAR estimate. Chapter 5 further examines the relationship between individual SVI for all stocks enlisted on the DOW-30 index. Chapter 6 examines the influence of searches on the first-day IPO returns. Finally chapter 7 presents the discussion and a conclusion of the study.

2. Literature review:

While attention measures such as news have been around for a very long time. Measurement of attention using Google's search volume index is a relatively new topic. Here I give an overview of some of the most important literature regarding the importance of internet searches and its effect on the financial markets.

Da et al., (2011) is perhaps the most cited paper on the use Google search volume index (SVI) to derive a proxy for retail investor attention and study its relationship to existing measures of attention. In their paper, they analyze Russell 3000 stock tickers weekly SVI values and determine SVI's influence with respect to volatility in stock prices. They incorporate Dash 5 reports as basis to determine the buying and selling behavior of retail investors. Their results indicate that retail investors are mostly buyers of attention grabbing stocks. They also determine that SVI is able to capture investors' attention more effectively than existing measures of attention, particularly in the case of less sophisticated retail investors. They further provide evidence that SVI influences first-day IPO returns (Da, Engelberg and Gao 2011). Following a different approach, Da et al, 2012 shows that changes in SVI for the, "products of a firm can predict unexpected announcements related to the revenues of the firm along with the consensus analyst forecast". They conclude that SVI contains significantly important information and market does not fully incorporate into its earnings expectations. More recently, Da et al, 2014 construct a Financial and Economic Attitudes Revealed by Search (FEARS) index by "aggregating daily SVI for keywords related to household financial and economic concerns". Their index is significant enough to forecast "daily realized volatilities of ETFs even after accounting for the effect of variables such as the VIX index, volume and turnover, as well as an alternative sentiment measure" (Da, Engelberg and Gao 2014).

Rubin and Rubin (2010) use the frequency at which firm-related information is edited on Wikipedia, the online free encyclopedia, as an instrument to gauge the degree to which people are engaged with the processing of firm-related information. They provide evidence that, "firms whose information is processed by the population more frequently are associated with lower forecast errors by analysts, smaller forecast dispersions and significant changes in bid-ask spreads" (Rubin and Rubin 2010).

Bank et al. (2011) employ a data set of the German stock market and they test Amihud and Mendelson (2006) hypothesis. They find that the rise in trading activity and stock liquidity is associated with abnormal changes in search queries. They attribute the improved liquidity in the Xetra index to a decrease in asymmetric information costs and conclude that search queries predominantly measures attention from uninformed retail investors. Moreover, they find evidence that an increase in search queries is related to temporarily higher future returns. This effect seems particularly robust for firms with low market capitalization. (Bank, Larch and Peter 2011).

Smith (2012) examines whether changes in the number of Google Internet searches for particular keywords can predict volatility in the foreign currency market. He finds that data on Google searches for the keywords, "financial crisis", "recession" and "economic crisis", have the potential to forecast future consequences. The Google search volume index for the keyword economic crisis and financial crisis is, "significantly related to the week ahead volatility for seven currencies. The Google search volume index for the keyword recession is also significantly related to week ahead volatility for five currencies" (Smith 2012).

Vlastakis and Markellos (2012) study a sample of 30 of the largest stocks traded in the NYSE and NASDAQ. Their study finds that demand for information (Google search volume index) and supply of information is significantly associated with historical and implied measures of volatility and to the volume of stocks traded, even after controlling for market returns and information supply. Moreover, they confirm that demand for information increases significantly during periods of higher price volatilities (Vlastakis and Markellos 2012).

Joseph et al. (2011) analyses search for tickers in S&P500 from the period 2005-2008 and construct a sentiment factors on Google search queries, similar to Dao et al, 2011, they prove the ability of online ticker searches to predict future stock returns and abnormal trading volumes. Particularly, they claim that Google searches on ticker serve as a valid measurement tool for retail investors. Based on prior examination of retail investor sentiment, they expect online search queries to forecast stock volatility and trading volume. In a sample of S&P 500 firms over the period 2005–2008, they, "find that, over a weekly horizon, online search intensity reliably predicts abnormal stock returns and trading volumes" (Joseph, Wintoki and Zhang 2011).

Antweiler and Frank (2004) examine financial internet blogs and message boards and find that higher activity in the financial blogs and message boards have a statistically significant influence on stock returns, a clear sign that a considerable number of investors use online source for information (Antweiler and Frank 2004).

Recent work by Mondria et al. (2012) uses measures of aggregate search frequency from AOL (American on Line) search engine as direct measures of attention to study home bias and Wu and Mondria (2011) uses Google search index to construct a measure of asymmetric attention and focus on asset pricing implications of asset allocation theories (Mondria and Wu 2012).

3.0 Sample and Data explanation

3.10 Google Search Volume Index:

Data collection process is divided into three parts. In the first part, weekly SVI data related to Google search queries are compiled for the term "S&P-500". In the second part, weekly SVI's of the twenty four of the largest stocks of DOW-30 index are assembled. In the third part, weekly SVI's of the 229 stock IPO's in the year 2014 are assembled. Those firms whose SVI's had a lot of null values were removed from the sample. Inorder to understand Google search index it is important to understand how this index is calculated as per Bank et al, 2011, "Google search volume index for a specific keyword provided by Google Trends is not given in absolute terms, but as a value relative to the total number of searches on Google in the corresponding time interval. For each search term, this relative value is then normalized so that the search volume always varies between 0 (i.e., a period in which search volume does not meet a designated threshold) and 100 (i.e., a period in which the highest relative volume was observed)" (Bank, Larch and Peter 2011).

Da et al. (2011) suggests that it is preferable to use the ticker symbols of stocks instead of the firm name based on three reasons. Firstly, searches may be made for reasons other than investments. Secondly, the name of the company can be spelled in a variety of ways. Thirdly,

Google Trends input does not allow searches for "alpha-arithmetic terms", which would deter the use of names for companies such as 3M. Even though I agree with Da et al. (2011) that the search queries for the company name does include some irrelevant elements, by people searching, for example, for Jobs, product reviews, prices of product etc. However, I assume that this component is purely deterministic (based on seasonality or a time trend) or there is a random noise. It can be the case that the name of the company is spelt in a variety of ways. Inorder to eliminate this bias I adopt the following keyword procedure. For the company name inserted in Google trend I check for the variations of spellings of names and check which keyword has the highest search volume relative to the others. Additionally, In order to avoid problems with noise arising from the fact that some of the search queries can have a general meaning, I survey the background in which these keywords are used in searches (this is provided by Google trends to categorize searches done for financial, company related information, product support or other purposes). Those keywords are excluded that are found to be associated for searches other than company information or financial information. However, I exclude terms that identify a firm's legal form (e.g. "LLP", "INC", "LTD" etc.). Google Trends allowed me to filter the search volume so that only queries submitted within USA are obtained. Google searches in the rest of the world are excluded so as to measure only changes in searches specific to US for US listed companies.

3.20 Capital market data

The capital market data is extracted from Wharton Research Database (WRDS) through CRSP. All the stock ticker symbols are entered in text file and their stock data extracted. The data include each stock's ticker symbols, name, daily return, closing and opening prices ask and bid prices and the number of stocks traded on a daily basis. The daily data is converted into a weekly data. The daily volume data is aggregated to form a weekly volume index. For data related to S&P-500 index and the DOW-30 stock firms the time length spans from January 2006 to December 2013. However, for the IPO analysis stock data of 229 firms was obtained for the year 2014. The data for IPO is significant as per Dao et al, 2011 where the results of only 185 stocks were estimated. In this study I used a shortlisted version of 42 IPO stocks.

3.30 IPO data

This paper utilizes the IPOSCOOP database for basic information related to all the IPOs that took place in the year 2014. I collect all IPOs of common stocks completed between January 20014 and November 2014 in the United States from IPOScoop.com. However those IPO's which were scaled back or those who had missed values were removed. In the year 2014 there were a total of 229 IPOs. However after shortlisting on the basis that the IPOs must have significant amount of search query data (SVI) I am only left with 42 unique IPO firms along with their first day IPO returns, offering price, first day closing price, first day returns and SCOOP ratings. IPOSCOOP ratings are important and I am going to use them further in the analysis. SCOOP ratings are constructed by surveying the Wall Street investment professionals and their value of significance to the IPO. The range of the rating is from as low as 1 to as high as 5. The daily Google search volume index is constructed manually from Google trends data for USA. It has been made sure that the time range of both the components are same type (i.e. weekly) and in the same range. IPOScoop.com LLC is an independent research firm predicting IPOs' opening-day performances (IPOSCOOP n.d.). Further this analysis is to incorporate the influence of Media for which weekly data of articles, headlines, news items of a firm are collected from the Factiva database. Factiva aggregates content from all available source, and provides information related to dissemination of information (Jones 2014). Pre-IPO assets are extracted from compustat in WRDS (Database 2014)against all the IPOs of the year 2014.

4.0 Influence of SVI on S&P-500

4.10 Research methodology

In this section I will analyze the S&P-500 index analyzing if large stock movements do capture retail investors attention. The volatility of the S&P-500 index has an effect on the searches of the Index. Or it can be the case that the searches of "S&P-500" have an effect on the volatility of the index. In this section I test that all the variables are in fact jointly determined. That is there is no dependent or independent variable and the independent variable has the same explanatory role as the dependent variable (this will become clear in the next section – VAR). This is the first of its kind of study on S&P-500. This is important because S&P-500 index is a better indicator to measure the overall market movement. It contains the largest 500 companies and its volume is representative of 90% of stocks traded. Most US investors keep a close watch on S&P-500 (Google Corporation 2012). This effect is observable as in the figures below. The graph compared simultaneously depicts a strong co-movement between the volatility of the S&P-500 index and Google search queries for its name. If volatility is low, searches related to the S&P-500 are generally at their overall average level, but in turbulent times search queries surge to levels which exceed their average by manifold. For example, when the volatility of the S&P-500 spiked at an almost record high of over 120, the number of searches for the index rose at par with the realized volatility.

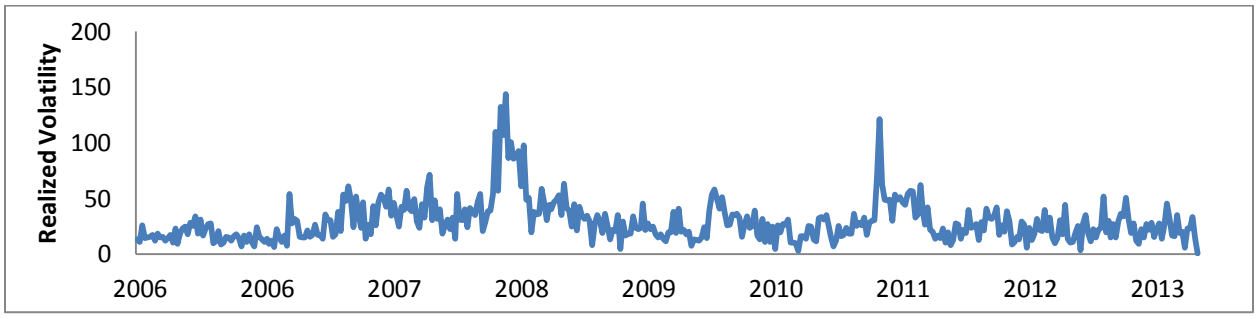

Figure 4-1: Realized volatility of S&P 500

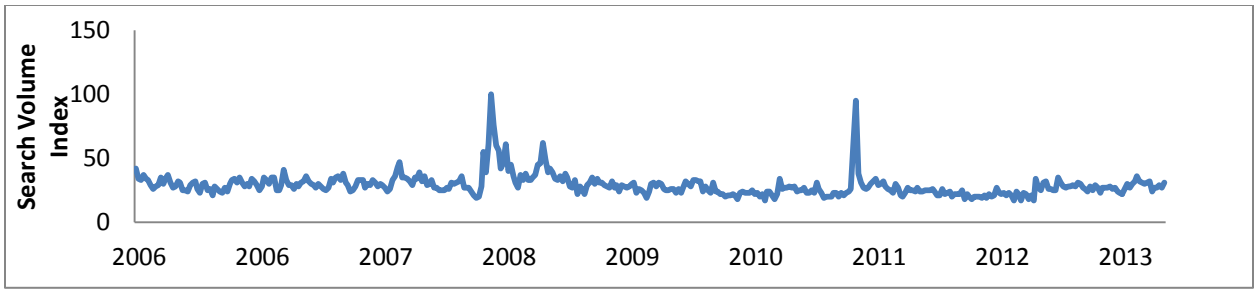

Figure 4-2: Frequency of searches of S&P-500 in Google trends

4.20 Realized volatility:

One of the most accurate and popular measures of historical volatility in the literature is realized volatility, (Andersen, Bollerslev and Diebold 2007), (Andersen, Bollerslev, and Diebold 2003) and (Bank, Larch and Peter 2011). To compute the realized volatility, I first compute the daily returns using end of the day stock prices ($p_{i,t}$) as follows:

$$Return_i = \frac{P_{i,t}}{P_{i,t-1}} - 1 \qquad (1)$$

My analysis focuses on the S&P-500 index from July 1, 2006 to December 31, 2013. Intra-day index prices are obtained from Yahoo finance database (Yahoo 2014). I construct a time series of daily realized volatility RV_i as introduced by (Andersen, Bollerslev and Diebold 2007):

$$RV_i = \sqrt[2]{\sum_{j}^{n} r_{i,j}^2} \qquad (2)$$

I then compute these price changes over a week by using a technique developed by (Corsi 2005). Where he takes a mean of the average realized volatility in a given week for all the changes in prices in order to convert the daily realized volatility into a weekly Realized Volatility (RV). Where n=5, assuming that the stock market operates five days a week.

4.30 Descriptive statistics of S&P-500:

Descriptive statistics of S&P-500 index realized volatility, search volume and stocks traded volume are presented in Table 1. As is evident from the kurtosis and skewness measures, the realized volatility time series is heavily skewed and far from being normally distributed. I therefore resort to the log of realized volatility as, amongst other variables, as suggested by (Andersen, Bollerslev and Diebold 2007), (Bank, Larch and Peter 2011) and (Andersen, Bollerslev, and Diebold 2003). The correlation coefficient between the realized volatility, search volume index and volume of traded stocks is quite significant as shown in Table 1 panel B. I find a correlation coefficient between the search terms SVI and realized volatility to be around 65.7%. The correlation coefficient between the search terms SVI and volume traded to be about 45% and 34% for the data in logarithms.

Table 1: This table provides descriptive statistics of search volume index (SVI), realized volatility (RV) and trading volume (VOL) of the S&P-500 index between January 1, 2006 and December 31, 2013. Panel A holds summary statistics while the panel B presents correlation estimates.

Panel A: Summary Statistics

	Mean	Standard Deviation	Kurtosis	Skewness	Minimum	Maximum
SVI	28.9	8.6	22.3	3.6	17.0	100.0
logSVI	1.4	0.1	4.5	1.3	1.2	2.0
Vol	19205.7	6545.7	1.2	0.8	2293.9	46641.1
Log vol	10.3	0.2	2.5	-0.6	9.4	10.7
RV	28.8	18.8	8.2	2.3	0.3	143.3
LogRV	2.8	0.6	4.9	-0.8	-1.0	4.3

Panel B: Correlations

	SVI	logSVI	Vol	LogVol	RV	LogRV
SVI	1					
logSVI	0.9588	1				
Vol	0.4389	0.4312	1			
LogVol	0.3405	0.3372	0.9602	1		
RV	0.6573	0.5937	0.5667	0.524		
LogRV	0.4355	0.4142	0.5354	0.5674	0.8721	1

An important concern when measuring investors' attention to the S&P-500 index is that it is known under different names and acronyms. I have used the search term "S&P-500" to measure retail investors' attention to the index. I have found that this name is most commonly used search term when retail investors are interested in the S&P-500 index.

Table 2: This table shows the top ten search terms that are most correlated with the search term "S&P-500" during the sample period January 2006 to December 2013. The third column shows the search volume scaled relative to the average search volume of "S&P-500", which is standardized to 100. For example, 37.04% in row 1 means that the search volume of "s&p" relative to the search volume of "S&P-500" is 37.04%.

Top searches for S&P 500		
Rank	Search term	Relative volume
1	s&p	37.04%
2	s&p 500	25.93%
3	s&p index	7.41%
4	s p 500	5.56%
5	s&p 500 index	5.56%
6	s and p	3.70%
7	s & p	3.70%
8	futures	3.70%
9	dow	3.70%
10	etf	1.85%

Table 2 compares the level of search volume relative to the search term "S&P-500". The search term "s&p" amounts to 37% of the total searches that were searched containing the term S&P. The ticker "S&P-500" is also a popular search term and its queries out of the total volume of searches made was around 25.93%. Other search terms such as s&p index, "s&p 500 index", "s and p", "s & p futures" and "dow" have a search volume that are less, i.e. their volume in terms of total searches is below five percent as compared to "S&P-500".

4.40 Vector Auto Regressive Model – (VAR)

This section will explain the methodology that is devised to conduct the study and describe what VAR is so that the analysis could be understood form a theoretical perspective. There can be a case that all the variables are in fact jointly determined. That is there is no dependent or independent variable, an independent variable has the same explanatory role as the dependent variable. My study involves S&P-500 Indices measuring weather Google searches are explained by volatility in S&P-500 index, and at same time testing if volatility can be explained by Google searches. The model further incorporates traded volume of the S&P-500 stocks. Before proceeding with the analysis and the result it is pertinent to describe the instrument, VAR (Vector Auto Regressive Model).

An important characteristic of the multivariate class of models investigated in this chapter is that each variable in the system is expressed as a function of its own lags as well as the lags of all of the other variables in the system. VAR is quite accurately defined by Hurn et al, 2012 as follows:

"*This model is known as a vector auto regression (VAR), model that is characterized by the feature that every equation has the same set of explanatory variables. This feature of VAR has several advantages. First, estimation is straightforward, being simply the application of ordinary least squares applied to each equation one at a time. Second, the model provides the basis of performing causality tests which can be used to quantify the value of information in determining financial variables. Christopher A, Sims was the first one to use technique that he developed using United States data on the nominal interest rate, money, prices and output in 1980. He suggested that to start with it was useful to treat all variables as determined by the system of equations. The model will have an equation for each of the variables under consideration. The most distinguishing feature of the system of equations is each equation will have exactly the same the set of explanatory variables. This kind of model is known as a vector autoregressive model (VAR)*". (Hurn, et al. 2012). An example of the bivariate VAR (p) model is:

$$y_{1,t} = c_{10} + \sum_{j=1}^{i} \beta_{11,j}\, y_{1,t-j} + \sum_{j=1}^{i} \beta_{12,j} y_{2,t-j} + \epsilon_{1,t} \qquad (3)$$

$$y_{1,t} = c_{10} + \sum_{j=1}^{i} \beta_{11,j}\, y_{1,t-j} + \sum_{j=1}^{i} \beta_{12,j} y_{2,t-j} + \epsilon_{1,t} \quad (4)$$

"*Where $y_{1,t}$ and $y_{2,t}$ are the dependent variables, p is the lag length which is the same for all equations and $\epsilon_{1,t}$ and $\epsilon_{1,t}$ are disturbance terms. The estimation of VAR is performed by simply applying ordinary least squares (OLS) to each equation one at a time. Despite the model being a system of equations, ordinary least squares applied to each equation is appropriate because the set of explanatory variables is the same in each equation*" (Hurn, et al. 2012, pp 74-88).

4.50 Dynamics of Google Searches and Volatility

In this paper I study in detail the dynamics of S&P-500 searches and its volatility. I find that Google searches today cause volatility next week. I use (Dimpfl and Jank 2012) paper where they show Dow Jones trading activity influences internet searches. Using their approach I will examine the dynamics of S&P-500 between realized volatility, search queries and trading volume by estimating three vector autoregressive (VAR) models.

Let y_t be a vector which contains the variables of interest, then the VAR (p) reads as follows:

$$y_t = c + \sum_{j=1}^{i} \beta y_{t-j} + \epsilon_t \quad (5)$$

Where c is a vector of constants and ϵ_t is a vector of white noise innovations. Equation above is fundamental to our analysis. Using Eq. 5 as reference I construct three VAR models. In the first model, I investigate the dynamics between realized volatility and search queries, i.e.$y_t = (\log RV_t \;\; \log SVI_t)$. In Model 2, I will investigate the dynamics between search queries and trading volume is i.e. $y_t = (\log VOL_t \;\; \log SVI_t)$. Lastly, Model 3 is going to test the joint model of realized volatility, search queries and trading volume i.e. $y_t = (\log RV_t \;\; \log VOL_t \;\; \log SVI_t)$.

In simple terms the VAR (3) for Model 1 would evaluate an OLS regression estimate where the RV is the dependent variable (equation 6) and in the second estimation SVI becomes the dependent variable (equation 7):

$$RV_{k,t} = c + \beta_1 RV_{k,t-1} + \beta_2 RV_{k,t-2} + \beta_3 RV_{k,t-3} + \beta_4 SVI_{k,t-1} + \beta_5 SVI_{k,t-2} + \beta_6 SVI_{k,t-3} + \epsilon_{k,t} \quad (6)$$

$$SVI_{k,t} = c + \beta_1 RV_{k,t-1} + \beta_2 RV_{k,t-2} + \beta_3 RV_{k,t-3} + \beta_4 SVI_{k,t-1} + \beta_5 SVI_{k,t-2} + \beta_6 SVI_{k,t-3} + \epsilon_{k,t} \quad (7)$$

Before I proceed with VAR it is important to mention the specification of lag structure p in VAR. If the lag length is too short important parts of the dynamics are excluded from the model. "*If the lag structure is too long then there are redundant lags which can reduce the precision of the parameter estimates. Moreover, in choosing a lag structure in a VAR, care needs to be exercised as degrees of freedom can quickly diminish for even moderate lag lengths. An important practical consideration in estimating the parameters of VAR (p) model is the optimal choice of lag order. A common data-driven way of selecting the lag order is to use information criteria. The three most commonly used information criteria for selecting time series model are the Akaike information criterion (AIC), the Hannan information criterion (HQIC) and the Schwarz information criterion (SBIC)*" (Hurn, et al. 2012). Therefore before proceeding in estimation of the above VAR model AIC test is conducted to determine the most suitable lag that would be required to estimate the VAR estimate model. AIC is the most relevant test to test for the lag structure in this model.

Table 3: The AIC value in the table shows that lag 3 is the best to estimate this VAR model. The other two measurements depict the ideal lag value to be 2

VAR: LAG Order Selection test (Pre-estimation for lags)								
	Criteria's				Observations	=	415	
LAGS	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	342.86	0.00	- 1.65	- 1.64	-1.63			
1	563.41	441.10	4.00	0.00	0.00	-2.69	-2.67	-2.63
2	586.56	46.31	4.00	0.00	0.00	-2.79	-2.74688*	-2.68809*
3	592.22	11.303*	4.00	0.02	.00021*	**-2.79331***	-2.74	-2.66
4	594.78	5.13	4.00	0.27	0.00	-2.79	-2.72	-2.61

The AIC value in the table above shows that lag required to best estimate this VAR model is at lag 3. Also note that there is not a huge difference in lags of the model. The other two measurements depict the ideal lag value to be 2. The difference between the AIC and the other two is not that significant. I am using AIC test to check for the lag as it is the most used measure in modern econometrics. Lower the AIC number better would be the VAR estimate. Now that we have determined what the lag value of our estimate is going to be we are going to estimate the three VAR models that were discussed above. The VAR models are constructed with a lag value of 3 and tested for significance as shown in the Table 4.

Table 4: This table displays the estimation results of a Vector Autoregressive model, VAR (3), for log realized volatility (log-RV), log search volume index (log-SVI) and log trading volume (log-VOL) for the SP-500 index. P-values are given in parentheses. Significances at the 1% level are highlighted in bold.

VAR estimation

	Model 1		Model 2		Model 3		
	(1)	(2)	(1)	(2)	(1)	(2)	(3)
VARIABLES	LogRV	logSVI	logSVI	logVol	LogRV	logSVI	logVol
$\log RV_{t-1}$	**0.265*** **	**0.0702** **			**0.227** **	0.0074	0.0020
	(0.000)	(0.0067)			(0.00)	(0.669)	(0.939)
$\log RV_{t-2}$	**0.280** **	**0.438*** **			**0.255** **	0.0037	0.0375
	(0.000)	(0.00055)			(0.000)	(0.827)	(0.147)
$\log RV_{t-3}$	**0.154*** **	0.0035*			0.119	0.0024	-0.0182
	(0.0056)	(0.835)			(0.0378)	(0.887)	(0.486)
$\log SVI_{t-1}$	**0.1940*** **	**0.690*** **	**0.766*** **	0.0602	**0.1470***	**0.760*** **	0.0530
	(0.0005)	(0.000)	(0.000)	(0.468)	(0.431)	(0.000)	(0.534)
$\log SVI_{t-2}$	0.17600	0.01630	-0.0081	0.1570	0.2310	-0.0098	0.1310
	(0.399)	(0.794)	(0.904)	(0.128)	(0.317)	(0.887)	(0.212)
$\log SVI_{t-3}$	-0.14800	0.0904*	0.0617	**-0.210*** **	-0.1840	0.0587	**-0.194** **
	(0.39)	(0.0795)	(0.246)	(0.010)	(0.315)	(0.281)	(0.0201)
$\log VOL_{t-1}$			**-0.139*** **	**0.602*** **	0.1710	**-0.144** **	**0.603** **
			(0.000)	(0.000)	(0.173)	(0.000)	(0.000)
$\log VOL_{t-2}$			0.0730	-0.0375	-0.0206	0.0712	-0.0599
			(0.0963)	(0.578)	(0.892)	(0.114)	(0.386)
$\log VOL_{t-3}$			0.0453	**0.276*** **	0.1330	0.0412	**0.282** **
			(0.223)	(0.000)	(0.306)	(0.284)	(0.000)
Constant	-0.0137	**0.290** **	**0.347** **	**0.668** **	**-1.080** **	**0.395** **	**0.740** **
	(0.938)	(0.000)	(0.001)	(0.000)	(0.0125)	(0.002)	(0.000)
Observations	415	415	415	415	415	415	415

In Model 1 I find significant coefficient for the autoregressive parameters for the realized volatility for all the lags of realized volatility but search volume index shows a significant autoregressive term for only lag 1. In model 1 part 1 the all the variables that have a p-value of

1% are highlighted in bold. The results show that current volatility to some extent is positively influenced by the previous search queries. The coefficient of 0.1940 is economically quite significant to suggest that there is a relationship between the searches and volatility. And in the same, model 1 part 2, the results show that past volatility does influences present search queries to some extent for lags 1 and 3. So my first model of VAR is able to predict that SVI is influenced by previous week's volatility. It also shows that current week's volatility is influenced by previous week's searches. This effect is concentrated to the first lag which is significant on the 1% level only while a lag of order 2 is not found to be significant. This shows that after an initial shock in volatility, search intensity rises in the subsequent week.

In model 2, I test the hypothesis that an increase in the number of searches should be followed by a rising levels of trading activity, i.e. the index trading volume should increase or decrease in the subsequent week after the search. I find that volume has a significant predictive power for search queries; particularly the first lag volume is quite statistically significant with a coefficient of negative 0.139. This suggests that a change in volume reduces the search volume index by a factor of 0.139. In model 2 part 2 search queries of the previous week with a lag 3 enter significantly in the volume equation with a coefficient of negative 0.210, and is statistically significant which reinforces the model 2 part 1 that once information has been sought via Google, trading volume changes, in my case it decreases. In Model 3 I test for the significance of all the variables in relation to each other. For model 3 part 1(logRV) there is only one statistically significant lag 1 variable of SVI at a p-value of 10% that is able to explain the variation in current volatility. In model 3 part 2 (LogSVI) the variable logvol with a lag of 1 is statistically significant reinforcing the previous model 2 results that current week searches are reduced by a change in the volume of S&P-500 stocks in the previous week.

One of the approaches is to track the effects of shocks through the model on the dependent variables is to construct an Impulse response function. In this way the full dynamics of the system are displayed and how the variables interact with each other over time. In performing impulse response analysis a natural candidate to represent a shock is the disturbance term ϵ_t of the VAR as it represents that part of the dependent variables that is not predicted from past information. The approach in impulse response analysis is to transform into another disturbance term. Formally the transformed residuals are referred to as shocks (Hurn, et al. 2012). This is

confusing to explain but as I will explain IRF using the VAR (3) Model 1 as an example the relevance of IRF is going to become clearer.

Figure 0-1 provides the impulse response functions of the Model 1. IRF is used to define the economically meaningful definitions that volatility can affect search queries immediately, but search queries do not affect volatility. The intuition behind this as per (Dimpfl and Jank 2012) is that there is a fundamental volatility shock that in turn triggers retail investor's attention and, thus, searches queries. Search queries, on the other hand, would not arise without a preceding event in the S&P-500 index.

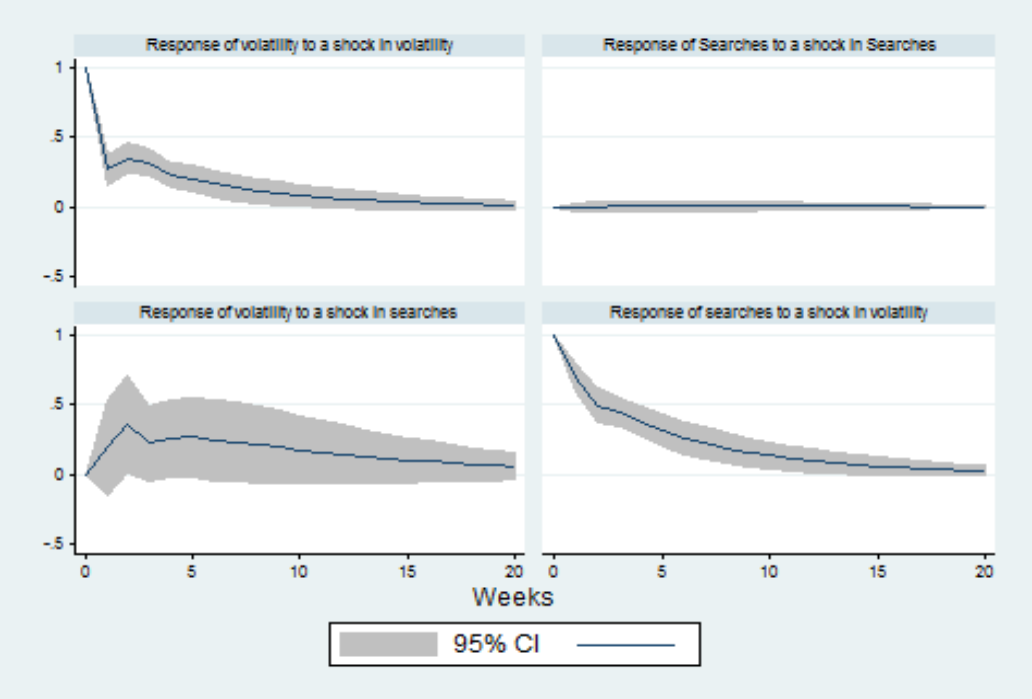

Figure 0-1: The graph shows the impulse response function of VAR (3) of realized volatility and SVI. Shaded areas indicate 95% confidence bounds.

The top right and the top left graphs in Figure 0-1 present the response of logRV and logSVI, respectively, to their own shocks. The top left graph is a slowly decaying function, these shocks are highly persistent. This shows that previous changes in volatility have a sharper influence closer to the current volatility. As the week lag increase previous week's volatility has less of an effect on the current volatility. The top right graph is a straight line this is because previous weeks shock in Google searches has no effect on the current searches. The bottom right figure

holds the impulse response of searches to a shock in realized volatility. After a volatility shock, attention slowly decays in a smooth exponential pattern. The bottom left figure presents the response of volatility to a shock in searches. After a volatility shock, attention is higher for small amount of time after which the IRF slowly decays.

Short Conclusion:

In this section I have established that there is a strong co-movement of S&P-500 realized volatility and SVI for its name. A heightened number of searches today are followed by an increase in volatility tomorrow. This research also establishes that Google searches have a more significant influence on the trading volume of a stock with a one week lag.

5.0 Dow Jones Industrial average

5.10 Dynamics of Searches and Volume

In this section, I will take a different approach to analyze the effect of Google search volume index on the financial markets. In this section I will use the measure of stocks traded volume as a measurement of Google search volume index. I will take a subset of stocks of S&P-500 comprising of Dow Jones industrial average index (Dow-30). Since Dow Jones is the most watched and oldest index in US, it's appropriate to assume that Dow represent the retail investor sentiment. The changes in stocks traded on Dow Jones accurately measure the attention of the retail investor (Da, Engelberg and Gao 2011). I start by investigating the relationship between SVI and trading activity. For the proxy of trading activity, I use the number of shares traded in a week. For measurement of attention I use SVI. SVI related data for individual stocks in Dow Jones is collected from the Google trends on a weekly basis. This relationship provides an insight into the sufficiency of Google search volume as a proxy for investor attention, and what impact it has on the traded stock volume (Vlastakis and Markellos 2012). Data pertaining to each stock is individually downloaded from CRSP in the Wharton Research Database. I get SVI's for each individual company from the Google trends.

I perform a regression analysis focusing on all the stocks that are enlisted on Dow Jones Industrial Average index from the year 2006 till 2013. Dow jones index was revised four times since 2006. My analysis includes only the original 24 stocks out of 30 that were in the index prior to year 2006. The new additions to the Dow are removed in order to have a coherent data of SVI and stock volume. In addition to measuring volumes and SVI effect I also incorporate other control variables that can be significant. I am well aware of possible influence of the variations in the market return and volatility, which are usually found to have a strong significant variable in a model describing stock movements (Vlastakis and Markellos 2012). Therefore it is pertinent to use these control variables. These variables include realized volatility, lagged volume and the market effect. The realized volatility is constructed in a time series of weekly realized volatility RV_i as introduced by Andersen et al 2001and converted into weekly realized volatilities.

5.20 Google search volume and trading activity

The research design is based on the combination of study by (Chordia, Roll and Subrahmanyam 2001), (Chordia,, Sahn-Wook and Subralunanyam 2007) and (Vlastakis and Markellos 2012). Chordia et al, 2012 describe that high turn over rate is associated with more frequent smaller trades, which progressively form a larger fraction of trading volume over time. These small trades are predominantly the retail investors who use Google searches to maximize the returns on their portfolios. Unlike sophisticated investors who rely more on the analyst and customized information, retail investor construct their own basket of stock using whatever basic tool that are available.

(Vlastakis and Markellos 2012) Develop a regression model where they show how information demand and information supply can be correlated with the realized volatility, implied volatility and VIX index of S&P-500. I extend their method to analyze the Dow 30 index to recent time. The multivariate regression in equation is used to study the influence of search volume on the traded volume in the week after the search query is submitted. The lagged volume is added to control for the auto regressive pattern of the traded volume and lastly market return is added to control for the stock market effect. The analysis is done in detail too ascertain if the regression coefficient of SVI do have a predictive power on the volume of stocks traded. This is useful due to the significance in correlation found between realized volatility and search volume.

$$Volume_{k,t} = c + \beta_1 SVI_{k,t-1} + \beta_2 RV_{k,t} + \beta_3 MRkt_{k,t} + \beta_4 TV_{k,t-1} + \epsilon_{k,t} \quad (8)$$

The regression is performed on all the twenty four stocks and the results are presented in

. These findings are in agreement with the correlation analysis and the findings of earlier studies that indicate that higher search volume is positively related to the search volume index. The SVI coefficient is statistically as well as economically significant in all but six of the stocks traded, even after controlling for effects of realized volatility, market returns and the first volume auto-lag. The coefficient of SVI on stocks traded of AT&T for instance shows that if the SVI in week before the trade goes up by a factor of 1 than volume traded in next week is going to increase by factor of 0.172, this result is statistically significant at 1% p-value. The R-squared value is around 46%. For all regressions a Breusch Pagan test for hetroskedasticty is applied and robust standard errors employed whenever the null hypothesis of homoscedasticity is rejected. In the analysis only 2 stocks regression I found were homoscedastic and had miss-specified variable problem. Also it can be proved that there is an association between traded stock volume and realized volatility. Also as predicted by (Chordia,, Sahn-Wook and Subralunanyam 2007) there is a strong association between the volumes traded and lag of volume of stocks traded. The lag volume coefficient is both statistically as well as economically significant for all the Dow-30 stocks. The impact of market return appears to be imperfect in most of the cases. The R-squared value of all the regressions seems to indicate that the model is quite accurate in predicting the variance of volume of stocks traded.

Table 5: This table presents the results of OLS regressions between volumes of stocks traded, realized volatility, and Google search volume with a lag of 1 and market return variable. The last two columns present the corresponding values of the R-squared statistic and the total number of observations. A star, double star and triple star denote significance at 10%, 5% and 1% level, respectively.

Dependent variable: Log(Volume) of stocks traded							
Stock	**LogSVI_lag1**	**LogRV**	**logVol_lag1**	**mrkt_return**	**Constant**	**Observations**	**R-squared**
AT&T	0.172***	0.251***	0.386***	-1.821***	5.08***	417	0.457
AXP	0.024	0.423***	0.471***	-0.968	4.586***	417	0.799
BA	0.132**	0.404***	0.276***	-0.92	5.772***	417	0.562
CAT	0.274***	0.356***	0.393***	-1.706***	4.633***	417	0.676
Cbank	0.149***	0.214***	0.87***	-1.535	1.188***	417	0.921
DD	0.03	0.366***	0.342***	-1.655***	5.42***	417	0.612
DIS	-0.447***	0.295***	0.292***	-2.469***	6.622***	417	0.592
GE	-0.087	0.366***	0.526***	-2.32***	4.704***	417	0.805
HD	-0.512***	0.325***	0.422***	-1.898***	5.894***	417	0.714
IBM	0.309***	0.331***	0.339***	-1.572***	5.019***	417	0.714
INTC	0.294***	0.287***	0.403***	-0.808	4.967***	417	0.562
JNJ	0.233***	0.241***	0.417***	-2.42***	4.578***	417	0.571
JPM	0.865***	0.328***	0.503***	-1.178*	3.145***	417	0.853
KO	0.08	0.275***	0.473***	-0.558	4.39***	417	0.532
MCD	-0.018	0.315***	0.426***	-1.575**	4.875***	417	0.55
MMM	0.686***	0.34***	0.357***	-1.957***	4.011***	417	0.7
MRK	-0.2***	0.358***	0.347***	-0.395	5.932***	417	0.535
MSFT	0.379***	0.314***	0.308***	-1.686***	5.701***	417	0.601
PFE	0.149***	0.326***	0.491***	-1.289**	4.536***	417	0.653
PG	0.407***	0.261***	0.408***	-1.559**	4.34***	417	0.601
UTX	0.196***	0.285***	0.387***	-2.621***	4.668***	417	0.616
VZ	-0.077	0.245***	0.445***	-1.352*	4.893***	417	0.444
WMT	-0.312***	0.333***	0.551***	-1.365**	4.529***	417	0.767
XOM	0.309***	0.255***	0.471***	-2.091***	4.178***	417	0.72

*** p<0.01, ** p<0.05, * p<0.1

6.0 Initial public offering and Google searches

6.10 Dynamics of Searches in an IPO

In this section I am going to test (Da, Engelberg and Gao 2011) claim that IPO's first day return are influenced by Pre-IPO week Google searches. They conduct their study on 185 IPOs from the year 2004 to 2007. This paper examines tests the Dao et al, 2011 hypothesis, that is if SVI does has an influence on the first day IPO returns. And controlling for other factors the SVI is significantly able to influence the first day IPO returns.

An appropriate setting to observe the effect of retail investors attention on asset prices is a company's IPO. IPOs have two very significant characteristics 1) Overall IPO's on average have higher first day returns (Loughran and Ritter 1995) and 2) In the long-run IPO's exhibit underperformance (Loughran and Jay 2002). When the trading starts IPO stocks that receive more retail attention are expected to experience higher buying pressure from the retail investors. Buying from retail investors induces higher first day stock returns since it is difficult to short sell these stocks. In the long-run as soon as the price pressure induced by attention of retail investors disperses, stock prices eventually go in reverse, resulting in long-run underperformance (Da, Engelberg and Gao 2011).

One of the instruments to measure retail investor's attention prior to the IPO is to collect information related to Google searches for that stock. The abnormally high Google searches prior to the IPO shows the attention of retail investor's. Data related to all the IPOs that took place in 2014 is obtained from IPOSCOOP website. I employ company names provided by IPOSCOOP to search for the stock in Google Trends to obtain the search volume index. Company names are manually entered into Google Trends and CSV file is downloaded and stored. This is done for all the IPO's that took place in the year 2014. For the year 2014 there were a total of 229 IPO's out of which only 42 IPO's were shortlisted for examination in this study. The basic criteria for short listing the IPO's for this study was to remove all those IPOs that were scaled back, those which had "missed" values and those who had insignificant Google searches.

6.20 Results

Observing the SVI's trend for all Stocks IPOs it is confirmed that there are significant changes in Google searches around the time of the IPO. Plotting a graph of mean Google searches volume (SVI) around the IPO date I find that there is a significant upwards trend in the SVI starting 2 to 4 weeks prior to the IPO week followed by a significant jump in SVI during the IPO week, see Figure 6-1 for illustrative purposes. The SVI not only reverts to its pre-IPO level 2 to 4 weeks after the IPO as predicted but also dips down significantly after 7 weeks showing a long run return reversal of the IPO's as per Ritter and Welch 2002.

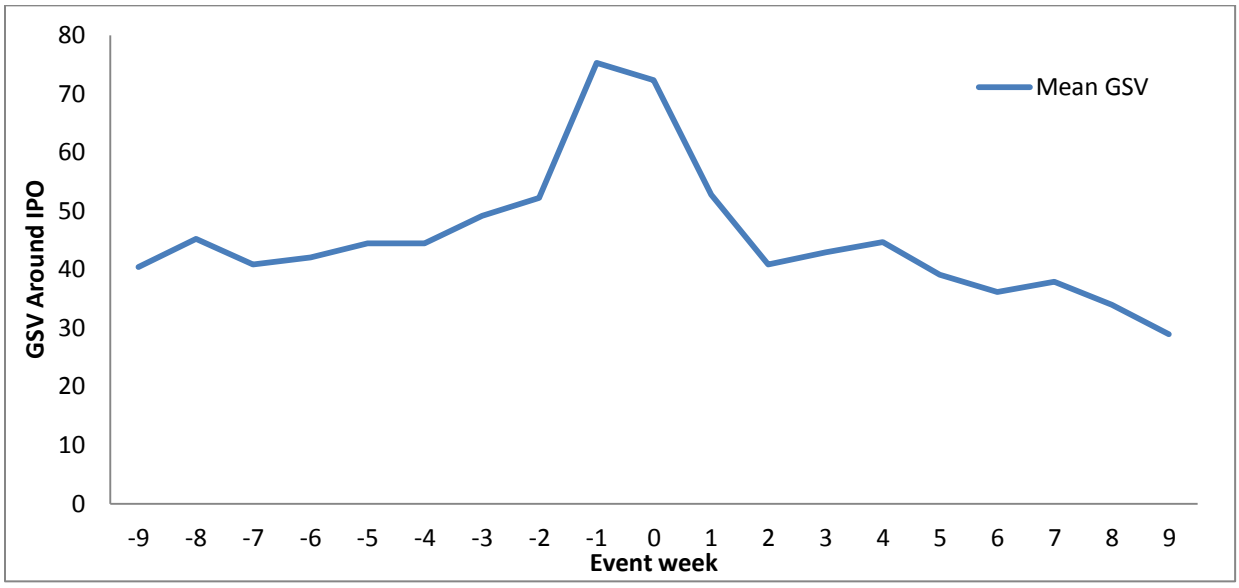

Figure 6-1: Plots the mean of the search volume index around the week of the IPO. Week 0 is the week of the IPO. The sample period is from January 2014 to October 2014. There were 42 IPOs out of a total of 229 with valid SVI in this sample.

Next, I examine the relation between first day IPO returns and IPOs rating (which is a primary attention measuring tool by Wall Street investors) developed by IPOSCOOP. Figure 6-2 summarizes the main results. The results show that IPOs that are highly anticipated by the market (the ones with higher IPOSCOOP ratings) have higher returns than those that have low ratings. The IPOs with low star ratings have first-day average returns of 10.84% while the set of IPOs with high ratings have a much higher first-day average returns of 19.84%. The difference in Low star rating in-between the median and mean first-day returns is about 9.04%. The difference

in High star rating in-between the median and mean first-day returns is 0.2%. This is due to the fact that there are a very few High star IPOs.

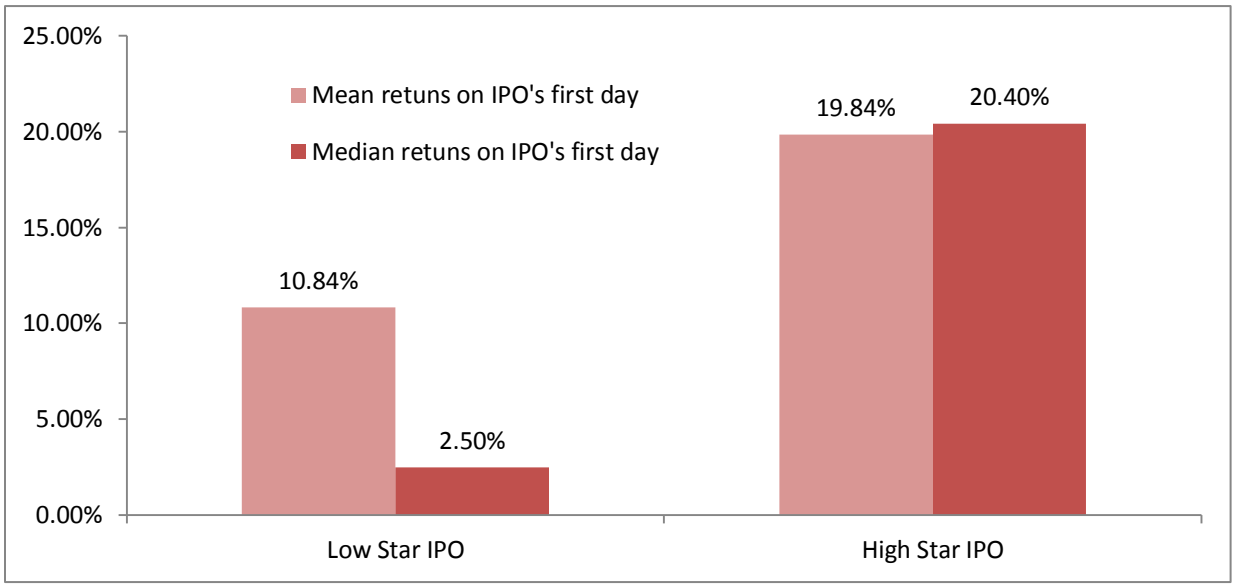

Figure 6-2 Plots Low Star IPO and high star IPOs with their average first-day returns. On a scale of 1-5 High Star IPOs are those IPOs whose rating is equal to or greater than 3 and Low Star IPOs are those whose ratings are below 3. These ratings are defined by IPOSCOOP based on the IPO anticipation of wall-street investors and are available at their website. The graph shows high star IPOs have higher first day returns as compared to the low-star IPO's

I perform a comprehensive analysis using regressions in Table 6. Regressions allow me to control for characteristics such as SVI while also controlling for other variables that can be related to first day IPO returns. As per Cook, Kieschnick and Van Ness 2006 control variable of media, IPO offer, Underwritter ranking and Assets size prior to the IPO are factors that can be related to the first day return of the IPO. All these varibale have been included in a a sequence to ascertain our claim that attention, specifically SVI can be a good predictor of first day returns. In all the regressions, the dependent variable is the IPO's first-day returns, computed as the closing price divided by the offering price (IPOSCOOP) minus one, see the $Return_i$ equation (9).

$$Return_i = \frac{P_{cl}}{P_{Op}} - 1 \tag{9}$$

In addition to Google Search Volume Index, I introduce one other control variable for the Google search volume index shown by Barber and Odean 2008 to have strong predictive power for IPO's first-day returns is Media. Barber and Odean 2008 have shown that those stocks that are mostly in news or media have higher attention. Stocks that are up for an IPO generally get higher Media attention, so media is an important control variable. Media is defined as the number of news articles recorded by Factiva (Jones 2014) on a weekly basis (using company name as the search criterion) before the IPO date. Factiva is a database managed by Dow Jones and provides information of news related to stocks that are in news in some major publication such as Wall street Journal etc. CSV files for data related to news article prior to the IPO are manually downloaded for all 42 IPOs stocks.

Variables that define the characteristics of an IPO are going to act as a control variable in the regression analysis. These variables are defined as below

- Log (offer): Offer is the total value of the company at the time of the IPO, and is defined as the product of offering price and volume of the shares offered.

- Log (assets): The firms total assets prior to the IPO.

- Underwriter ranking: (Manaster and Carter 1990) ranking of lead underwriter, obtained from Jay Ritter's website. If there had been more than one underwriter for the IPO. I have summed up all there rankings in order to gauge the effect of the underwriters reputation on first day return of IPO stocks. The ratings are on a scale from -9 to 9, See (Manaster and Carter 1990) for more detail.

- Star rating: These ratings are on the level of 1-5, 5 being the highest and 1 the lowest. These ratings are surveyed individually by all the Wall Street investors and averaged to get an overall sentiment from the market about the significance of the IPO. These rating are available on the IPOSCOOP website (IPOSCOOP n.d.). I have used a dummy variable instead to differentiate between the high star and the low star rated IPOs, to discount for the effect of missing SVI values of Low star rated IPOs.

Regression 1 in Table 2 confirms that SVI, alone, is statistically very significant with an R-squared value of around 40%. However, the regression coefficient of 0.00947 suggests that a one standard deviation increase in SVI leads to approximately 1.0% higher first day returns. The squared term of SVI in regression 2 does increase the R-squared value but the economic significance of the variable decreases. Regression 3 confirms the predictive power of the news variable; Media, as documented by Liu, Sherman and Yong 2014 and (Barber and Odean 2008). A one standard deviation change in Media coverage increases the first day returns by almost 0.2 %. The regression coefficient on a stand-alone basis is statistically significant at 1% with an R-squared value of around 68% which shows that Media is better able to predict the variance in first-day IPOs stock returns as compared to the SVI on a standalone basis. Higher Media seems to be a better predictor than SVI in terms of a more a higher R-squared but with a low economic significant coefficient.

Regressions 4 to 7 in Table 2 control for other IPO characteristics. In regression 4, I have included IPO offer size as IPO related control variable the regression coefficient on SVI drops significantly to 0.00263, but remains significant at 10% p-value. The variable offer size has a negative coefficient emphasizing that the size of the IPO is not a significant predictor of how high the first day IPO returns are going to be. The R-squared value improves to about 75% proves that media and SVI jointly are strong predictors of the variance in the first day IPO returns. In regression 5, I have included underwriter ranking as an IPO related control variable. The R-squared value remains the same which emphasis that there is no direct or indirect linkage of the underwriter rankings on the first-day returns of IPOs. In regression 6 I have included assets prior to the IPO as a control variable. Estimation in regression 5 proves that there is no influence on the first-day IPO returns on assets prior to the IPO. However, when I include the interaction variable in regression 7 between the SVI and Ratings the results show that this variable is statistically significant at 10% p-value. The R-squared value shows that interation variable of SVI and ratings have a higher R-squared value of around 77%. The control variable for IPO such as offer size, underwriter ranking and assets size have no effect on the first day IPO returns. However the variables Media, SVI and the SVI*ratings have statistically significant effect on the first day IPO returns of the stock.

6.30 Short conclusion and limitations:

The results from the IPO analysis reveal that SVI has a significantly higher forecasting power for first-day IPO return. Additionally this study proves that IPO with higher first day returns under-perform in the long-run. However there are limitations to this study regarding the hypothesis development for this study. In the above analysis I have proved that higher lagged searches are followed by a higher return. However, it could be the case that retail investors expect higher first-day returns and they increase the frequency of searches of the IPO prior to the IPO. So in fact higher *expected* future returns is the reason for higher SVI.

The results can also extended to include the Chemmanaur, 1993 hypothesis about IPO underpricing/overpricing of high value firms prior to the IPO. Chemmanaur uses stocks volume of stocks held by insiders to gauge the effect of future expected return of first-day IPO returns (T. Chemmanur 1993). This study can also be utilized to incorporate the underpricing of stocks prior to IPO by using a different proxy. The proxy that I suggest that can be used to estimate the underpricing or overpricing of the stock can be the "number of revisions of offering price" prior to the IPO. As per (Loughran and Ritter 1995) the higher the number of revisions in the offering price higher will be probability weather the stock is underpriced or overpriced. Using the variable *"number of revisions of offering price"* as a proxy for underpriced or over priced stocks I can forecast the future expected return of first-day IPO returns.

Table 6: This table regresses IPO first-day returns on pre-IPO week search volume (SVI) and other IPO characteristics. The dependent variable is the individual IPO's first-day return. Independent variables are defined in the text above. The sample period of IPOs is from January 2014 to October 2014. Only regular and common stock IPOs traded on NYSE, and NASDAQ with a valid SVI (searched using company names) are retained in the sample. Standard errors are (in parentheses) *, **, and * represent significance at the 10%, 5%, and 1% level, respectively.**

Dependent variable : First-day returns of IPO's

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SVI	**0.00947*** **			0.00263*	0.00288*	0.00305*	0.00142
	(0.00182)			(0.00153)	(0.00159)	(0.00168)	(0.00203)
SVI_2		**0.00010*** **					
		(0.00002)					
Media			**0.00199*** **	**0.00177*** **	**0.00175*** **	**0.00170*** **	**0.00209*** **
			(0.00021)	(0.00025)	(0.00025)	(0.00030)	(0.00040)
Log offer				-0.06642**	-0.07589**	-0.06355	-0.07903
				(0.02808)	(0.03207)	(0.04736)	(0.04803)
Under writer ranking					0.00174	0.00183	0.00218
					(0.00277)	(0.00281)	(0.00279)
log assets						-0.00921	-0.01847
						(0.02574)	(0.02625)
SVI*Ratings							**0.0347***
							(0.00248)
Constant	-0.11931	0.02967	-0.18288***	1.02204*	1.15116*	0.97119	1.24577
	(0.09601)	(0.06217)	(0.06375)	(0.54492)	(0.58657)	(0.77791)	(0.79245)
Observations	42	42	42	42	42	42	42
R-squared	0.40488	0.49914	0.68217	0.75142	0.75404	0.75492	0.76790

7 Discussion of Findings and Concluding Remarks

Existing measures of investor attention such as turnover, extreme returns, news headlines, and advertising expense are indirect proxies Da et al, 2011. In this paper, I proposed a direct measure of investor attention using search frequency in Google (SVI). The thesis decisively confirms that there is a relationship between Google searches and volatility in stock market and also have an influence on first ay returns of an IPO stock. I started by analyzing the S&P-500 index to determine the nature of influence between the SVI, volatility and volume traded. I find that variables volume is more significantly related to SVI. On the basis that SVI effects volume of traded stocks. I construct a sample of the largest stocks traded in Nasdaq using a sample of Dow Jones 30 index. For each individual stock I have shown that SVI has a very significant influence on traded volume of all the stocks of Dow jones 30. I also provide evidence that SVI captures the retail investor's attention, using first-day IPO returns. I find that an increase in volatility in SVI for Dow Jones 30 stocks predicts higher searches in the next 1-2 weeks. This paper proves (Da, Engelberg and Gao 2011) claim that SVI also contributes to a large first-day return and long-run underperformance for a sample of IPO stocks. Beyond testing theories of attention, this paper also illustrates the usefulness of search query data in financial applications. Search volume is a very practical way to reveal and compute the interests of investors that can help in predicting short run fluctuations in the market.

References

Chordia, T., R. Roll, and A Subrahmanyam. "Market liquidity and trading activity." *Journal of Finance*, 2001: 56(2):501-530.

Manaster, Steven, and Richard B Carter. "Initial Public Offerings and Underwriter Reputation." *Journal of Finance*, 1990: 1045-67.

Andersen, T. G., T. Bollerslev,, and F. X. Diebold. "Modeling and Forecasting." *Econometrica*, 2003: 529-629.

Andersen, T. G., T. a Bollerslev, and F. X. Diebold. "Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility, The." *The Review of Economics and Statistics*, 2007: 701-720.

Antweiler, W., and M. Z. Frank. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *The Journal of Finance*, 2004: 1259–1294.

Bank, M., M Larch, and G Peter. "Google search volume and its efect on liquidity and returns of German stocks." *Financial Markets and Portfolio Management*, 2011: 239-264.

Bank, Mathias, Martin Larch, and George Peter. "Google search volume and its inuence on liquidity and returns of German Stocks." *Financial Markets and Protofolio Management*, 2011.

Barber, 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*, 2008: Vol. 21, Issue 2, pp. 785-818.

Chemmanur, Thomas J. "The pricing of initial public offerings: A dynamic model with information production." *Journal of Finance*, 1993: 285-304.

Chemmanur, Thomas, and An Yan. "Advertising, attention, and stock returns." *Working paper, Boston College and Fordham University*, 2009.

Chordia,, T, H Sahn-Wook, and A Subralunanyam. "The cross-section of expected trading activity." *Review of Financial Studies* , 2007: 20(3):709-740.

Comscore. *ComScore Releases February 2014 U.S. Search Engine Rankings.* 2014. https://www.comscore.com/Insights/Press_Releases/2014/3/comScore_Releases_February_2014_U.S._Search_Engine_Rankings.

Cook, , Douglas O., Kieschnick Robert, and A. Van Ness Robert. "On the marketing of IPOs." *Journal of Financial Economics*, 2006: 35–61.

Cook, Douglas O., Robert Kieschnick, and Robert A. Van Ness. "On the marketing of IPOs." *Journal of Financial Economics* 82 (2006): 35–61.

Corsi, Fulvio . "Measuring and Modelling Realized Volatility: from Tick-by-tick to Long Memory." *Submitted for the degree of Ph.D. in Economics*, 2005: 125.

Da, Zhi, Joseph Engelberg, and Peng Gao. "In search of fundamentals. ." *AFA 2012 Chicago Meetings Paper*, 2012.

Da, Zhi, Joseph Engelberg, and Pengjie Gao . "The Sum of All FEARS Investor Sentiment and Asset Prices." *The Review of financial studies*, 2014.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. "In search of attention." *The Journal of Finance* 66, no. 5 (2011): 1461–1499.

Database, Wharton Research. *Database.* 2014.

Dimpfl , Thomas , and Stephan Jank . "Can Internet Search Queries Help to Predict Stock Market Volatility?" *Finance Meeting EUROFIDAI-AFFI Paper*, 2012: 1-34.

Gao, Peng. "The sum of all fears: investor sentiment and asset prices. ." *Working Paper, University of Notre Dame and University of North Carolina at Chapel Hill.*, 2011b.

Ginsberg, Jeremy , Matthew Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. "Detecting influenza epidemics using search engine query data." *Nature* 457 (2009): 1012–1014.

Google. 2014. https://support.google.com/trends/answer/www.google.com/trends.

Google Corporation. *Google Finance.* May 25, 2012. https://www.google.com/finance?cid=296878244325128.

Grullon, G, G Kanatas, and J P Weston. "Advertising, Breadth of Ownership, and Liquidity." *"Review of Financial Studies 17*, 2004: 439-461.

Gujarati , Damodran. *Basic Econometrics.* New York: McGraw Hill Book Co, 2003.

Herbert, Simon A. *Designing Organizations for an Information-Rich World.* Baltimore: John hopkins Press, 1971.

Hurn, Stan , Vance Martin, Peter Phillips, and Jun Yu. *Financial Econometric Modelling.* University of Cambridge, 2012.

IpoScoop. n.d. https://www.iposcoop.com/index.php.

IPOSCOOP. "IPOSCOOP." n.d. https://www.iposcoop.com/index.php (accessed November 26, 2014).

Jones, Dow. *Factiva.* October 2014. www.factiva.com (accessed October 2014).

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*, 2011: 1116–1127.

Kahneman, Daniel. NJ: Prentice-Hall, Englewood Cliffs, 1973.

Liu, Laura, Ann E. Sherman, and Zhang Yong . "The Long-Run Role of the Media: Evidence from Initial Public Offerings." *Management Science*, 2014: 1945-1964 .

Ljungqvist, Alexander, Vikram Nanda, and Raj Singh. "Hot markets, investor sentiment, and IPO pricing." *Journal of Business 79, .* 79 (2006): 1667–1702.

Loughran, Tim, and Jay Ritter. "The new issues puzzle." *Journal of Finance* 50 (1995): 23-51.

Loughran, Tim, and Ritter Jay. "Why don't issuers get upset about leaving money on the table in IPO's." *Review of Financial Studies* 15 (2002): 413–443.

Mingelgrin, Dan H, Simon Gervais, and Ron Kaniel. "The High-Volume Return Premium." *jjournal of finance*, 2001: 876-919.

Mondria, J., and T. Wu. "Asymmetric attention and stock returns." *Working paper, University of Toronto*, 2012.

"publican 2014." *Encyclopædia Britannica Online.* n.d. http://www.britannica.com/EBchecked/topic/482582/publican (accessed March 22, 2014).

Ritter, Jay, and Ivo Welch. "A review of IPO activity, pricing, and allocations." *Journal of Finance 57*, 2002: 1795–1828.

Rubin, Amir, and Eran Rubin. "Informed Investors and the Internet." *Journal of Business Finance & Accounting*, 2010.

Seasholes, Mark S., and Guojun Wu. "Predictable behavior, profits, and attention,." *Journal of Empirical Finance* 14 (2007): 590–610.

Shepardson , Mathew . "How the internet works! Before you market understand your medium." January 21, 2014. http://communityboost.org/how-the-internet-works-before-you-market-understand-your-medium/.

Smith , Geoffrey Peter . "Google internet search activity and volatility prediction in the market for foreign currency." *Finance Research Letters*, 2012.

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

Varian, Hall, and Hyunyoung Choi. "Predicting the Present with Google Trends." *Economic Record*, 2012: 2 – 9.

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

Welle, Deutsche. "Most Popular Search Engines 2014 - Shift Ranking of May 8." May 8, 2014. http://www.dw.de/most-popular-search-engines-2014-shift-ranking-of-may-8/a-17618235.

Yahoo. *Yahoo finance.* October 2014. finance.yahoo.com.

Yuan , Yu. "Attention and trading,." *Working paper University of Iowa*, 2008.

Appendix

0-1: How Google handles a web search query

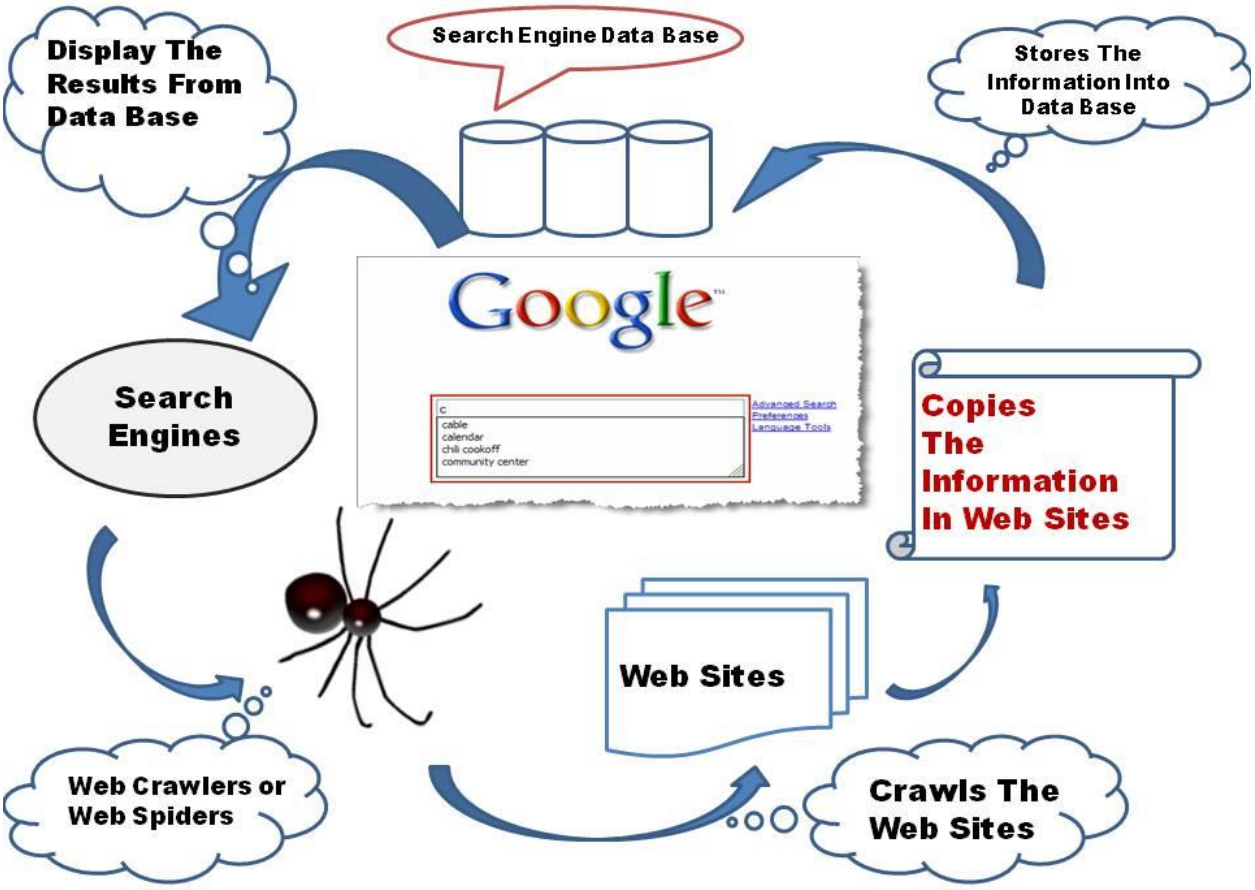

Figure 0-2: This figure shows how a typical web search index such as Google works. “Search engines are collections of programs that perform three basic functions online; indexing, crawling, and ranking. The most common program that runs these most basic indexing and crawling function is an automated robot. These web crawling robots are sometimes called “bots”, “crawlers”, or “spiders”. Indexing is the recording of the information in the form of coding and content on a web page. When a spider is on a webpage it scans the code and collects data that is relevant to its programing. Crawling is the process of mapping and moving from page to page through links. As the spider moves along and scans the web it stores the information in massive hard drives located in servers around the world called a datacenter. With thousands of these spiders moving along the web at amazing speeds the search engines are about to quickly store updated information on the web. After the data is collected and stored, the search engine will have access to this information for future searches. When a user starts a search the searching engine must sort and organize the information according to its importance relative to the keyword used. This process is called ranking” **(Shepardson 2014)**.

0-3: Search terms for the companies enlisted at the Dow Jones

No	Dow Jones Industrial	Ticker	Permno	Google search Term
1	3M Co.	MMM	22592	"3M"
2	American Express Co.	AXP	59176	"American Express"
3	AT&T Inc.	T	66093	"AT&T"
4	Boeing Co.	BA	19561	"Boeing"
5	Caterpillar Inc.	CAT	18542	"Caterpillar Inc."
6	Chevron Corp.	CVX	14541	"Chevron Corporation"
7	Cisco Systems, Inc.	CSCO	76076	"Cisco Systems"
8	Coca-Cola Co.	KO	11308	"The Coca-Cola Company"
9	E.I. DuPont de Nemours & Co.	DD	11703	"DuPont"
10	Exxon Mobil Corp.	XOM	11850	"Exxon Mobil Corp."
11	General Electric Co.	GE	12060	"General Electric Co."
12	Goldman Sachs Group Inc.	GS	86868	"Goldman Sachs"
13	Home Depot Inc.	HD	66181	"home depot"
14	Intel Corp.	INTC	59328	"Intel Corporation"
15	International Business Machines Corp.	IBM	12490	"IBM"
16	Johnson & Johnson	JNJ	22111	"Johnson & Johnson"
17	JPMorgan Chase & Co.	JPM	47896	"JPMorgan Chase"
18	McDonald's Corp.	MCD	43449	"McDonald's"
19	Merck & Co. Inc.	MRK	22752	"Merck & Co."
20	Microsoft Corp.	MSFT	10107	"Microsoft Corporation"
21	NIKE Inc.	NKE	57665	"Nike"
22	Pfizer Inc.	PFE	21936	"pfizer"
23	Procter & Gamble Co.	PG	18163	"Procter & Gamble"
24	The Travelers Companies, Inc.	TRV	49154	"The Travelers Companies"
25	United Technologies Corp.	UTX	17830	"United Technologies Corporation"
26	UnitedHealth Group	UNH	92655	"UnitedHealth Group"
27	Verizon Communications Inc.	VZ	65875	"Verizon Communications"
28	Visa Inc	VZ	65875	"Visa Inc."
29	Wal-Mart Stores Inc.	WMT	55976	"Walmart"
30	Walt Disney Co.	DIS	26403	"The Walt Disney Company"