UPPSALA
UNIVERSITET

Measuring Information Diffusion about Stocks on the Web

Alexander Sundin, Martin Pettersson, Viktor Edward

Project in Computational Science: Report

February 20, 2015

PROJECT REPORT

Department of Information Technology

Abstract

This project investigates the information diffusion between online stock discussion boards. All the messages written about 500 stocks on ten discussion boards over a three-year period have been collected. The messages have been analysed and the information diffusion between them estimated. From the analysis a positive correlation between the amount of messages posted and the volume traded could be found. Also stocks with large market capitalization had lower information network centralization. The daily tone of the discussions of the stocks was compared to the return, but no correlation could be found.

Contents

1 Introduction

1.1 Background

One of the main drivers for peoples and institutions actions on financial markets is information. How the information is distributed and incorporated into financial markets and asset prices is thus an interesting question. A novel approach to this problem is to use network theory to try to model the information distribution and compare measures of the models with market data.

This project is a part of an ongoing research initiative exploring relationships between stock returns, diffusion of information and investor networks in the market which has received attention both in academia and among practitioners (e.g. in CFA Digest). Research done by Ozsoylev and Walden (2011) [1] and Ozsoylev et al. (2014) [2] provide theoretical models for slow information diffusion through information networks. The models suggest that some information is neither completely public nor completely private. This view, that there is a "semipublic" channel through which information reaches asset prices in addition to the traditional channels, is quite different than the standard approach to asset pricing with heterogeneous information, since the standard approach assumes completely private signals. Anecdotal evidence suggests that the semipublic channel is important.

The theory is consistent with several stylized facts in real-world markets:

1. Most large price movements are not related to public signals.
2. Price volatility is highly time-varying.
3. Trading volume is high and heterogeneous — in a predictable way.
4. Portfolio investments are heterogeneous — in a predictable way

The theory is explored in Ozsoylev et al. (2014) [2], where account level data of trades on the Istanbul Stock Exchange are used to infer the market's information network. From the empirical investor network a number of observations are made:

- Central agents within the network earn higher profits than less central agents.
- Central agents within the network tend to act earlier on information events than their less central neighbors.
- The structure of the network suggests that there is a decentralized information diffusion. This decentralized diffusion is better described by diffusion through localized channels than through mainstream media. Social networks can be one example of such localized channels.

This project looks at the subject from a different angle, starting from some of the available information instead of inferring the market's information network from trades. More specifically, we estimates the information diffusion among market participants in a subset of social networks on the internet, namely internet discussion boards. By crawling these discussion boards and scraping relevant discussions one can estimate how the information tends to spread from one board to another and then check if this process relates to stock data. This is an exciting new area of research, which potentially can deliver academically important insights about what determines stock valuations, as well as insights that are relevant to practitioners about whether such data can be used to for example generate abnormal returns.

A student at Berkeley, Jun-Mok Kim, has carried out initial work, writing a web-crawler that collected data for 20 stocks on 3 discussion boards (Yahoo! Finance, InvestorsHub and The Lion), providing a proof of concept of the approach. Specifically, Jun-Mok collected data about how many times specific words (from a pre-specified list of about 500 words) were mentioned within a specific time period (measured hourly) for a specific stock in a specific discussion board, over a 9-month periods

1.2 Scope

The project builds upon Jun-Mok's initial work but extends it in a number of ways. Firstly, the data gathering was extended to cover more stocks on an increased number of discussion boards over a longer period of time. Secondly, several methods were developed and tested to analyze the data and estimate information diffusion between the discussion boards. Finally, an investigation was conducted to see if the developed estimation methods could be motivated to be related to actual stock data.

The increase in amount of gathered data was accomplished by increasing the three dimensions that mainly limit the amount of data. Namely these are the number of stocks investigated, the number of discussion boards covered and lastly the time period covered. To increase the number of discussion boards covered, this project started out with a list of 26 possible candidates that were evaluated with respect to how much data they comprise as well as an estimate of the difficulty to implement a crawler and scraper for the candidate board was performed. The time period was chosen such that there is a relevant amount of data during the time period from the selected discussion boards and stocks.

In this project the estimation of information diffusion was limited to methods using dictionaries. These dictionaries were used to detect keywords that are relevant in this context. After keyword extraction,s two groups of methods to estimate information diffusion were tested. The first group are methods that match keywords or keyword sets between messages in different ways. The second group are methods that estimate the tone of messages as positive or negative and then matches tone across messages. From the information diffusion estimates a network can be inferred. These networks estimate how information moves between the different discussion boards.

Finally the obtained network models were compared to actual stock data to see if any interesting conclusions could be made. One hypothesis for this project was that smaller stocks have more centralized networks. In other words, information about smaller stocks is less spread over the entire population than information about big stocks. Another hypothesis was that the tone of the discussions is related to the market returns. Moreover, centralized networks have better correlation between the tone of the discussion and stock returns. Yet another hypothesis that was tested is if there is any correlation between the number of messages posted and trading volume. If true, this hypotheses would suggest that the posted messages are related to actual trades.

2 Data gathering

As mentioned in the introduction one of the goals of this project was to extend the amount of data gathered from Jun-Mok's earlier work. This was done by increasing the three dimensions: number of stocks investigated, number of discussion boards scraped and finally the time period. A record of how the increases were made as well as some aspects of the implementation of the crawler and scraper is provided in the following subsections.

2.1 Stocks

The number of stocks investigated was increased from 20 in Jun-Mok's analysis to 500. The stocks were chosen such that 270 of them are considered as big stocks (with between 917455 and 37290600 shares outstanding) and 230 considered as small stocks (with between 89509 and 288933 shares outstanding). Furthermore the stocks were picked from New York Stock Exchange, NASDAQ and American Stock Exchange which all operate in Eastern Standard Time.

2.2 Data sources

In order to decide upon the additional discussion boards a study was carried out by looking at the 26 candidates. These boards got graded after two main properties and ranked accordingly.

The first property was the activity of the board. Having more discussion boards in the analysis is useless if the newly included ones don't contain any information. The second property was the structure of the board, how difficult will it be to write a crawler and a scraper for the specific board. Since 500 stocks were to be investigated it is not feasible to add the addresses manually. It had to be possible to automatize the finding of discussions for each stock and the messages had to be tagged in some way to make them accessible. The second property is also important for future research. If someone wishes to look at a new set of stocks it should only require a change of an input parameter.

Based upon empirical studies of these two properties the following ten discussion boards were chosen.

- Yahoo! Finance
- Investors Hub
- The Lion
- Raging Bull
- Stockhouse
- The Motley Fool
- Investors Village
- Investor Discussion Board
- HotStockMarket
- Silicon Investor

2.3 Implementation details

All of the ten discussion boards had a different structure which led to the need for a unique crawler for each one. And since it is not feasible to add the addresses manually automated search process was needed. In some cases, like Yahoo! Finance, it was easy to find the relevant sub-board since one only had to change the ticker name in the web page url. However for most cases it was trickier and required the use of the board's search functions. A board's search function will most often not direct the crawler to the right sub-board but to a list of the search results. The search outcome must then be scanned to find the right sub-board for a stock.

The crawlers were written in Python using the libraries urllib and Beautiful Soup. In short, urllib acts as a high-level interface for fetching data across the World Wide Web. The library abstracts away all communication details and simply returns the resource that the supplied URL points to. The Beautiful Soup library parses html content and returns a tree structure that is easy to navigate and search to find specific elements.

Several problems were encountered during the implementation and testing of the crawlers. Some of these problems were:

- Not all the boards use the same time zone. Some use the users local time zone while others use the servers local time zone. This had to be carefully examined so that all gathered data was transformed to the same time zone to be able to accurately examine the information diffusion.
- The crawling is time consuming because of the large number of requests sent. Even though the response time and latency for each request is relatively fast the aggregated time of all request becomes long. Care had to be taken to minimize the amount of requests through using the shortest crawling path to get the data needed.

- Some web pages are coded in a way that is not well-suited for our purposes. The HTML tags are not necessarily named with ids or classes consistently or even named at all, this makes it complicated to find the relevant data on the web pages.
- Unusually long response times and network connection exceptions needed to be handled carefully so that the crawlers were robust against such problems. This was needed since the crawlers had to run for several days without supervision in some cases.

The collected data was stored in 5000 csv files, each one representing a particular stock and discussion combination from our list of 500 stocks and ten discussion boards.

3 Data analysis

With the goal of finding properties which relate to market data the gathered messages were analysed in a multi-step process. This process led to inferring networks that describe how information diffuses between the different discussion boards. The term information diffusion networks are introduced as a name for these networks. The structure of the information diffusion networks were then compared to market data to try to find relations.

To model an information diffusion network the information diffusion between the different discussion boards was estimated. Estimating the information diffusion is a task divided into two sub-tasks. The first one is classifying the messages. It is unlikely that the exact same message is posted in two boards so some classification is needed to be able to relate messages to each other. Secondly a diffusion needs to be defined. How closely in time should two messages have to be posted and what other restrictions are needed to define a diffusion?

3.1 Message classification

As earlier mentioned, when some topic is discussed in one forum and then later discussed in another, one can't assume that they are discussed with the exact same written messages. Some sort of model to classify and relate messages is needed. Two different models were developed. Both models use dictionaries consisting of positive and negative financial keywords, developed by Loughran and McDonald[3]. These dictionaries have been developed to represent words that are either positive or negative in a financial context and consists of 354 positive and 2329 negative words.

The first model classifies messages by searching through the gathered messages and reducing them to the keywords found in the dictionaries. This method is based on the hypothesis that the dictionaries represent keywords that express the relevant information in the messages. The results of the classification is a set of keywords for each message. As an example the following sentence:

*"... resulting in a **collaboration**. The **collaboration** has so far been a **success**, but I'm sure it will end up as a **failure**."*

gets reduced to the following set of keywords:

$$\{\textbf{collaboration}, \textbf{collaboration}, \textbf{success}, \textbf{failure}\}.$$

In contrast, the second model tries to classify the tone of the messages as either positive or negative. The tone is expressed as a positivity score defined as the number of positive keywords found divided by the total number of keywords found. This implies that a message with a positivity score of 1 only contains positive keywords and that one with a positivity score of zero contains only negative words. Using the same example as above "collaboration" and "success" are found in the positive dictionary while "failure" is found in the negative dictionary. So this message would have a positivity score of $^3/_4$. In addition to the scoring two thresholds are introduced to classify a message as positive if the positivity score is above the positive threshold or as negative if the positivity score is below the negative threshold. These two thresholds act as tuning parameters for the model.

Both these methods may seem to be naive, or too easily classifying messages similarly. But considering the fact that for most stocks there are not more than a few messages posted per day and that messages later are compared with messages posted on the same context (e.g. about the same stock) this decreases the risk of classifying two unrelated messages equally.

3.2 Information diffusion

Defining the concept information diffusion consists of two components. The first one beeing defining when a diffusion happens and the second one defining when two messages are considered as related. The following definition has been used to determine when a diffusion of information occurs.

Definition 3.1 (Diffusion)**.** If an event **X** occurs at time t in **Forum A**, then in **Forum B** the same event **X** occurs in the time period $[t_{i+1}, t_i + \Delta t]$ without happening in the time period $[t_{i+1} - \Delta t, t_i]$ in **Forum B** diffusion has ocurred from **Forum A** to **Forum B**. Δt is a parameter representing the maximum time window during which a diffusion happens and thus has to be chosen differently depending on the dynamics of the studied context. In this project hourly aggregates of messages are continued so the smallest time step is one hour.

An event is defined differently depending on which of the two classification models that are used. For the keyword set model each keyword during an hour define an event. In the positivity score model an event is an hour classified as either negative or positive. The hourly aggregate was chosen to further reduce the risk of classifying unrelated discussions as related.

3.3 Information diffusion networks

From the diffusion events defined in previous subsections the information diffusion networks were generated. A network was generated for each stock and each node represents a discussion board. Each such node in the network has a one-way connection to every other node with a weight consisting of the amount of outgoing diffusions to that node. The degree centrality can be calculated to represent how central that node is in the given network.

Definition 3.2 (Discussion Board Degree Centrality)**.** The Discussion Board Degree Centrality p is defined as

$$p = N + 1 \tag{1}$$

where N is the amount of outgoing diffusions from the discussion board to all other discussion boards.

From each node having their degree centrality defined, one can reason about how centralized a network is; this property is called centralization and is defined as the following.

Definition 3.3 (Weighted Network Centralization)**.** The network Centralization C given a stocks Diffusion Information Network is defined as

$$C = \frac{\sum_{i=1}^{N} p_* - p_i}{\sum_{i=1}^{N-1} p_* - 1}, \tag{2}$$

where p_i is a discussion boards degree centrality and $p_* = max(p_i)$.

The centralization of an information diffusion network is a descriptive property of the network quantifying how centralized the information diffusion is. A network with evenly spread diffusions between all discussion boards has a small centralization while a network with the main bulk of diffusions happening between a subset of all discussion boards has a higher centralization. This is one of the key properties which relation to market data have been investigated. Information diffusion networks with higher centralization can be thought of as networks where the information is more semi-public.

4 Results and discussion

4.1 Statistics of gathered data

To better understand possible problems and possible restrictions it was important to have a firm knowledge of the data. In Table 1 the amount of messages posted

per discussion board is summarized. It is clear that the amount of data varies vastly between the different discussion boards. One can directly see that Yahoo! Finance by itself is responsible for approximately 70 % of all the messages posted. Also some boards are so inactive it might be worth considering if they are of any value in some sort of analysis at all, see Investor Discussion Board, HotStockMarket and Silicon Investor. From Table 2 it can be seen that all messages comprised in total approximately 55.3 million words of which around 1.4 million were keywords found in Loughran and McDonald's dictionaries.

Table 2: Total number of messages scraped from all discussion boards together with an approximation of the total number of words (found by finding all character combinations separated by space using regex) as well as the total number of words found in the two word lists from Loughran and McDonald [3].

Property	Quantity
Total message count	996 603
Approximation of total word count	55 262 936
Words found in positive word list	597 597
Words found in negative word list	799 203

In Figure 1 the distribution of messages during the average day is presented for each discussion board. It can be seen that the daily distribution is similar across all boards and also that the mean number of messages per day varies a lot between different message boards. An interesting observation is that the main bulk of messages seems to be posted during the stock exchanges open hours 09:30 to 16:00. This supports the hypothesis that people that are active on the discussion boards are traders. If this would not be the case, the assumption would be that people would tend to discuss more during spare time, for example in evenings. This further indicates that the information posted is relevant.

From Figure 2 the distribution of messages during the average week is presented for each discussion board. Also here one notices that the shapes of the distributions are similar and that the amount of messages vary widely between different discussion boards. A relatively small amount of messages is posted on the weekends with the main bulk of activity happening during the weekdays. A trend can also be seen that there tend to be more activity mid week than on mondays and fridays. This result in combination with what have earlier been discussed about Figure 1 further supports that the messages are relevant since they are posted mainly when the stock exchanges are open.

Table 1: Summary statistics of the total number of messages scraped segmented per discussion board. The descriptive statistics are on the stock dimension per discussion board (i.e. for example the max figure corresponds to the stock with most messages on the specific discussion board). Activity is defined as the number of stocks out of 500 where at least one message was found for the discussion board. The discussion boards are abbreviated with the following abbreviations: IH - InvestorsHub, YF - Yahoo! Finance, TL - The Lion, HSM - HotStockMarket, IV - Investor Village, TMF - The Motley Fool, RB - Raging Bull, S - Stockhouse, IDB - Investor Discussion Board, SI - Silicon Investor.

	IH	YF	TL	HSM	IV	TMF	RB	S	IDB	SI	Total
Min	0	0	0	0	0	0	0	0	0	0	0
Mean	253	1 407	37	6	101	11	17	156	0	5	1 993
Max	31 748	28 694	3 352	818	21 325	1 696	3 128	30 957	103	835	57 335
1st quartile	2	29	0	0	0	0	0	0	0	0	45
Median	4	118	2	0	0	0	0	0	0	0	159
3rd quartile	15	460	9	2	0	1	2	1	0	0	707
Total	126 509	703 471	18 554	2 855	50 281	5 307	8 637	78 245	222	2 522	996 603
Activity	441	490	421	157	22	153	211	168	29	64	

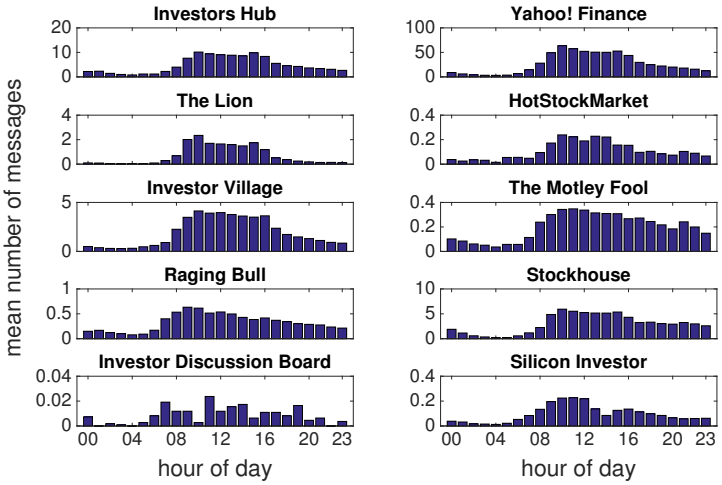

Figure 1: Bar plots showing the hourly distribution of the mean number of messages posted each hour of the day.

As earlier stated, the total amount of messages vary widely between the different discussion boards. When investigating the data it was also found that it varied greatly with time as can be seen in Figure 3 where a time series with the weekly number of messages over the three year period is presented for each discussion board. As can be seen, the activity on Raging Bull decreased substantially in the end of 2013. Furthermore, it can be seen that Stockhouse sprung to life in the beginning of 2014.

Some problems arose while collecting the data. Yahoo! Finance only indexes a maximum of 500 pages of discussions for a stock, where each page contains 20 discussions. Due to this 16 out of the 500 stocks were cut off early and not all messages over the three-year period could be crawled. Also two stocks contained some messages so long that it exceeded the Python library csv's max field size of 131072 characters, causing an error to be thrown. These messages were thought of as irrelevant and for simplicity they were ignored.

As one could expect from the data statistics showing how many messages there are in each discussion board, the run-time for the crawlers varied heavily. For the smaller boards the crawlers finished in about an hour. While for Yahoo! Finance it took a week to collect all the messages and Investor Hub around three days. The two main factors causing the long run times are network latency and server response time. These are typically in the order of tens to hundreds of milliseconds which might seem small. However, the sheer amount of requests causes the total runtime to escalate. If a faster runtime is wished for, parallelizing crawlers to better interleave the requests is a possible solution. Even so, the collection of the data for this project was a one-time operation and there were no need to decrease the runtime as it was feasible to keep the crawlers running for a week.

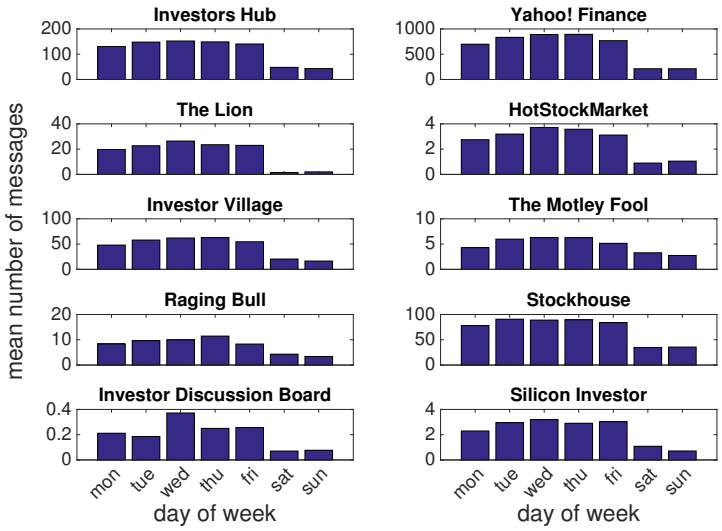

Figure 2: Bar plots showing the daily distribution of the mean number of messages posted on each day of the week.

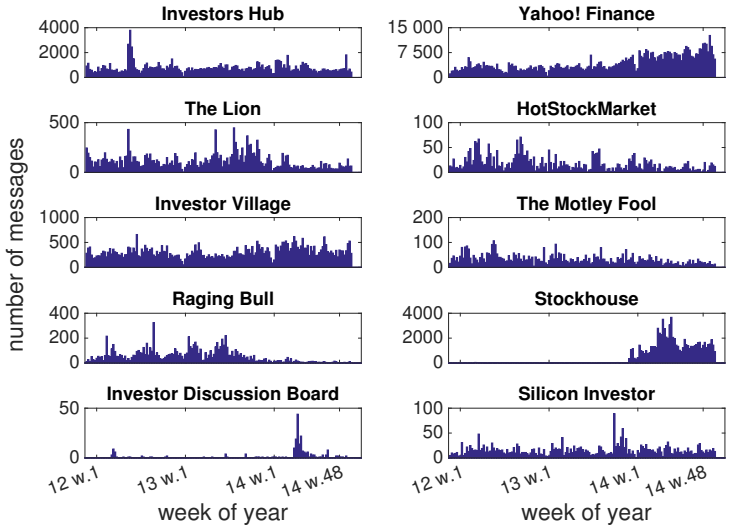

Figure 3: Bar plots showing amount of messages per week for all discussion boards over the three year period.

4.2 Message classification and Diffusion

Even with the, as earlier called, generous classification definition any diffusion at all could only be found for 120 out of 500 stocks using the keyword classification model. Using the positivity score classification model there were even less, only 60 out of 500. This when using a time delay limit $\Delta t = 5hours$ for the diffusion and for the positivity model a threshold of 0.7 and 0.3 for a positive respectively negative discussion. These limits are set after empirical testing to what seemed reasonable and are hence somewhat arbitrary. Potential future work could try investigating the consequences of different values.

The main reason for this is the lack of activity, both for some of the discussion boards overall and for certain stocks on every board. As seen in the statistics chapter some of the boards are not nearly as active as the bigger ones. For an example Yahoo! Finance! (the most active one) have some active discussion on 490 of the stocks while Investor Village (the least active one) have some active discussion on 22 of the stocks. Active here defined as at least one message for the stock. It might be worth reflecting over if some of the boards should be dropped.

The keyword classification model is a simple model but seemingly agreed well with the slightly more sophisticated positivity model. One problem with this model could be that some of the keywords are too common to keep. One of them is the word "good" that is such a commonly used word that it might be a good idea to filter it away, due to the high probability of it being used even if two discussions are unrelated.

The positivity measure for messages is no new idea, for an example see [4]. For future research one could adopt a more sophisticated machine learning model instead of looking at keywords. Some common algorithms for text classification are Naive-Bayes, Support Vector Machines and Decision Trees. Right now the keywords represents about 2 % of all the collected words, so practically 98 % of the data is left unused. There have also been cases where the messages have been classified in different categories, see Werner and Murray [5], where they classified messages as "sell", "buy" or "hold". The advantage of using different categories for classification is that a message being positive can depend on so much. If a stock doubles in value overnight some people will be happy, while someone who sold the stock the day before not so much.

4.3 Time series

In the information diffusion models we consider the positivity score and frequencies of keywords by the hour. This results in two time series for each stock and board, one for the hourly positivity score and one for the hourly keyword frequencies.

The stock data available is on daily basis and contains volume traded, closing price, return and adjusted return. By computing a daily overall positivity score of a stock it is possible to compare how the positivity relates to the return. The

hypothesis is that a positive correlation between the positivity score and the return will be found. However, this is not the case. But lets say there is a lot of positive buzz about a stock today, what will happen with the return tomorrow, or up to five days later? No such relationships were found. The percentage of stocks with *No or negligible relationship* was $> 85\%$, even when testing with aggregated return and with 0-5 days of lag.

The board activity of a stock correlated to the traded volume was also investigated. In other words how the total amount of messages written per day relates to the volume traded. It can bee seen in Figure 4 how the correlation coefficients are distributed for the stocks. The result in Figure 4 shows that a majority of the stocks show a positive correlation while no negative is found. A correlation was found for 299 of the 500 stocks with the requirement of at least 30 days with messages. The average correlation is 0.39, which is a moderate positive relationship, see Table 4.3. In Section 4.5 we will see how this measure can be connected to network centralization.

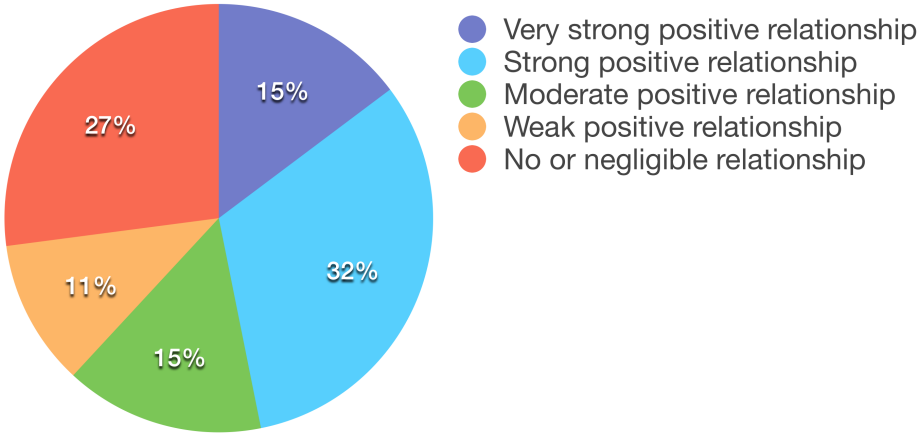

Figure 4: Distribution of the correlation between the daily amount of messages and volume traded for a stock. For correlation classification see table 4.3.

Table 3: Table of correlation coefficient classification [6].

Classification	Correlation coefficient r
Very strong positive relationship	+.70 or higher
Strong positive relationship	+.40 to +.69
Moderate positive relationship	+.30 to +.39
Weak positive relationship	+.20 to +.29
No or negligible relationship	+.01 to +.19
No or negligible relationship	-.01 to -.19
Weak negative relationship	-.20 to -.29
Moderate negative relationship	-.30 to -.39
Strong negative relationship	-.40 to -.69
Very strong negative relationship	-.70 or higher

4.4 Information diffusion network models

The biggest issue with generating the Information Diffusion Networks is that for the smaller stocks the diffusions occur rarely. Smaller stocks could sometimes only have discussions active in one or two discussion boards. Out of the 500 stocks only around 120 of them had any diffusions found at all (using the keyword classification model, for the positivity model there were even less). An example of a network can be seen Figure 5 for the Facebook stock. The arrows describe how many cases of diffusion there has been from one board to another. From Figure 5 one can see there has been 154 cases of diffusion from Investors Hub to Yahoo! Finance and 287 the other way around. In Figure 6 an example of a smaller stocks network is seen with a total of seven diffusions.

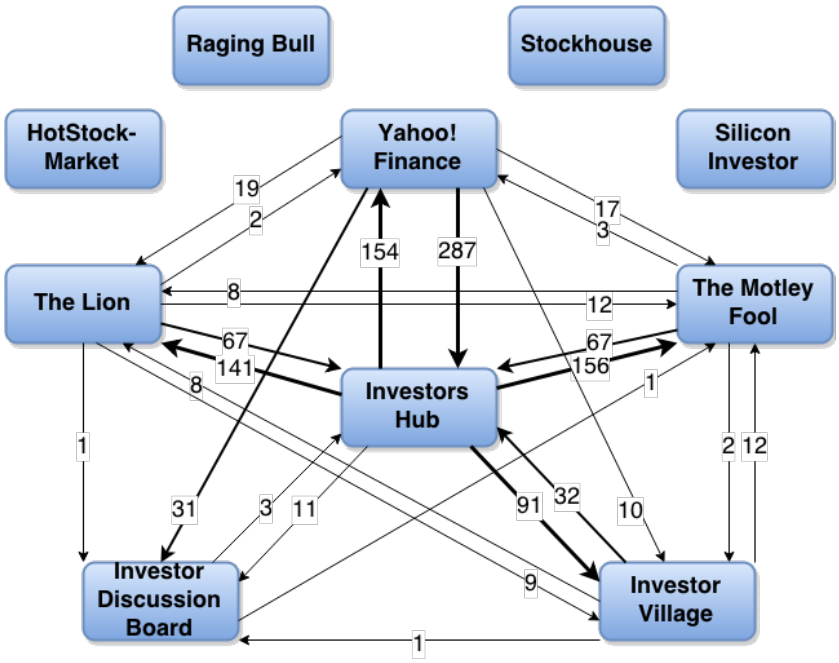

Figure 5: The Information Diffusion Network of the Facebook stock using the keyword classification model. The four isolated nodes had no ingoing or outgoing diffusions over the three-year period. This network has a centralization of 0.8811.

The networks for the stocks all had centralization values C between $0.8 - 1.0$. Since $C \in [0, 1]$ it might look as all the networks have a similar structure but this is not necessarily true. The problem is that the activity of Investors Hub and Yahoo! Finance is higher than anywhere else. This often leads to a high number of diffusions between these two and the centralization always being high. However this doesn't mean that the measure is bad. As long as the ordering is correct, having values in $[0.8, 1.0]$ or $[0, 1]$ leads to the same results. The main issue is that few stocks are as active as Facebook, compare Figure 5 and Figure 6.

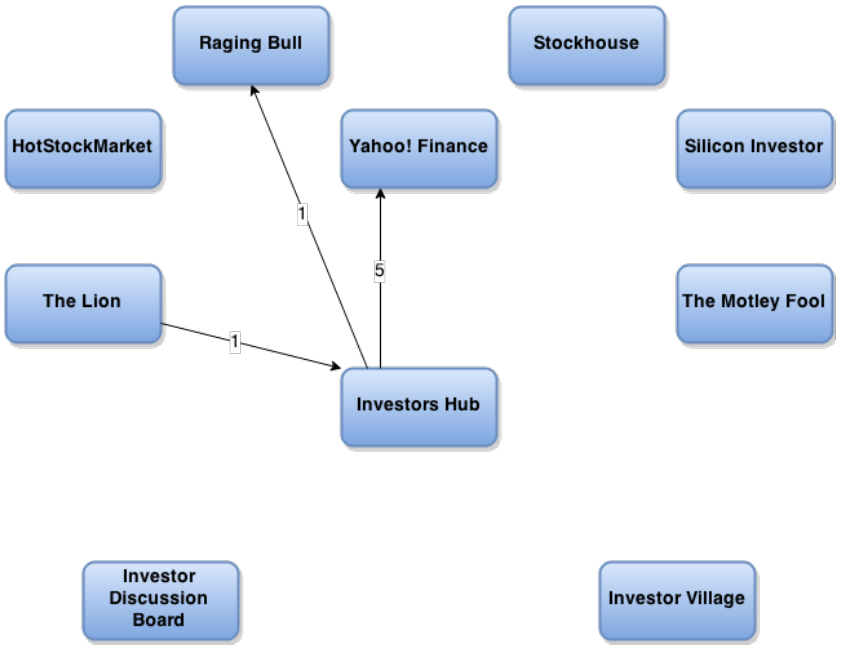

Figure 6: The Information Diffusion Network of the Pizza Inn Holdings stock using the keyword classification model. The seven isolated nodes had no ingoing or outgoing diffusions over the three-year period. This network has a centralization of 0.9815.

4.5 Network relation to stock data

Tests have been conducted to see how network centralization relates to stock data. Through comparing network centralization and market capitalization (cap.) a negative correlation was found, see Figure 7. Market capitalization is the total value of the shares outstanding, in other words the share price times the total number of shares of the stock. One can see that small stocks are more centralized while large stocks are more decentralized. This is consistent with the hypothesis that smaller stocks, with fewer holders, are discussed in a more narrow network, while the discussions of large stocks, with a lot of holders, are more evenly distributed in the network.

As previously stated, see Figure 4, for some stocks there is a significant correlation between the number of messages written and the volume traded each day. One hypothesis is that more centralized networks, which tends to be smaller stocks, have a higher correlation between the traded volume and the number of messages written. Meanwhile decentralised networks, which tend to be larger stocks, should be more stable and the trade should not be connected to the number of messages written at the same degree. However, when testing how these correlations relates to network centralization no evident correlation was found.

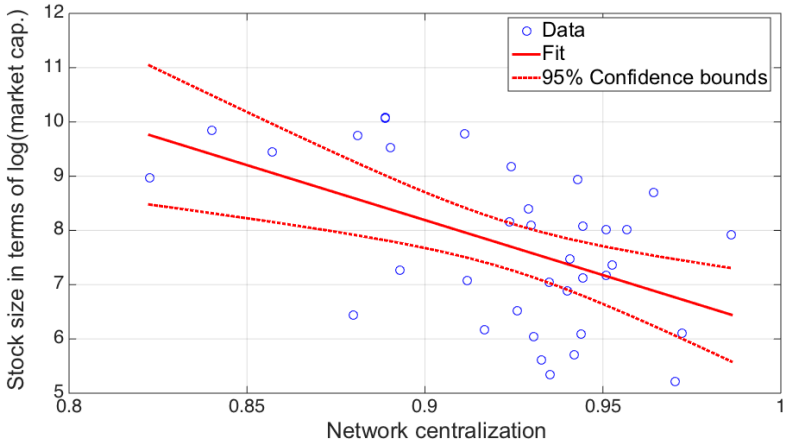

Figure 7: Logarithm of stock market capitalization plotted against network centrality. A statistically significant linear regression is included.

5 Summary and conclusions

Market capitalization of a stock has been shown to be inversely correlated with the network centralization measure from our developed information diffusion model. This agrees with the hypothesis that smaller stocks have more central information networks.

No correlation could be found between the positivity of the discussions and the return of the stocks, with a lag set to 0-5 days. This agree with earlier research on the area [4]. However a strong correlation between the daily amount traded and the amount of messages posted could be found. That people tend to post more when they trade might not come as a surprise, but strengthens the idea that there is a connection between the market and online discussions.

Moreover the result motivates further research on the novel area to discover more implications of the centrality measure. Our data can serve as a firm ground for future research to find these implications.

6 Acknowledgements

We would like to thank Johan Walden of University of California at Berkeley for being our supervisor during this project. Also Maya Neytcheva of Uppsala University for being our course coordinator and the IT department of Uppsala University for providing us with the necessary computing tools.

References

[1] Han N. Ozsoylev and Johan Walden. 2011. *Asset pricing in large information networks.* Journal of Economic Theory 146:2252-80.

[2] Han N. Ozsoylev, Johan Walden, M. Deniz Yavuz and Recep Bildik. 2014. *Investor Networks in the Stock Market.* The Review of Financial Studies 27:1323-1366.

[3] Tim Loughran and Bill McDonald. 2011. *When is a Liability not a Liability.* Journal of Finance, V66, pp. 35-65.

[4] Linhao Zhang. 2013. *Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation* Unpublished honor thesis, The University of Texas at Austin

[5] Werner Antweiler and Murray Z. Frank 2004. *Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards* The Journal of Finance Vol. 59, No. 3, pp. 1259-1294

[6] Pearson's Correlation - A Rule of Thumb. (n.d.). Retrieved January 30, 2015, from *http://faculty.quinnipiac.edu/libarts/polsci/Statistics.html*