From News to Company Networks:

Co-occurrence, sentiment, and information centrality

Thomas Forss
IAMSR
Åbo Akademi, Turku, Finland
RiskLab Finland
Arcada University of Applied Sciences
thomas.forss@abo.fi

Peter Sarlin
Department of Economics
Hanken School of Economics, Helsinki, Finland
RiskLab Finland
Arcada University of Applied Sciences
peter@risklab.fi

***Abstract*— To understand connectivity among companies in financial news, and their overall influence, we define an algorithm for ranking companies in networks of positive, negative and mixed co-occurrences. We collect a homogeneous set of financial articles from crowd sourced news to get a sentiment polarization between positive and negative news. We use this polarization to develop three types of sentiment networks by matching co-occurrences of companies in the Standard & Poor's 500 index. The entities are then ranked according to the information centrality measure and normalized by market capitalization. This text-to-network process allows qualitative and quantitative analysis of the top 25 nodes in each network on a quarterly basis over a period spanning 2011Q1 to 2016Q2.**

I. Introduction

While reading through different financial news sources we are presented with contradicting opinions and contradicting news articles on a daily basis. Looking at financial news as a whole to get an indication of the state of different companies is not an easy task. A person will just as likely end up coming to a faulty conclusion without following some kind of systematic and measurable approach. That is why many investors, money managers, and risk managers look at different types of indicators and rankings to help them make decisions.

Companies can be ranked and measured in many ways. The Standard & Poor's 500 (S&P 500) index is in itself a measure of importance as it is an index that reflects the performance of the most influential companies in the United States stock market. Maybe the most widely used measures of importance is comparing market capitalization of companies, which is a measure of how valuable a company is. Trading volume is another well known measure that is generally used with a short time frame and is used to measure trading activity of company stocks, options, and other financial instruments.

Different types of financial metrics are used for different purposes. Market capitalization for instance is used by many money managers, funds, and pension funds as a limitation on what companies they are allowed to invest in. Volume on the other hand can be used by traders to find patterns and unusual activity than can be used in winning trades.

Analysis of numerical financial data such as stock market price movements have stood at the center of much of the quantitative research in finance that was done up until the early 2000s. Methods that have been used range from exploratory analysis [1] and statistical methods [2] all the way to state of the art artificial intelligence and machine learning algorithms [3]–[5]. Descriptive, predictive, and prescriptive analytical methods have been used to get a better understanding of stock market movements and risk assessment during the last several decades.

In the middle of the previous decade research using financial news, using mainly text content, started to garner more attention both in media and among researchers. This has has lead to among other things new types of risk assessments [6] and predictive systems [7]. However, fairly little research into new measures that can be used to rank companies by flow of news over a specific time period has so far been suggested. In this research we begin work on a model for ranking companies according to news flow and news sentiment.

Textual analysis has been around for quite a long time. Some early research into financial news include [8]. The large bulk of research using financial news to create different indicators started only after the millenium changed. In 2008 Ötzgür et al. [9] studied the co-occurrence of people in a Reuters financial news data set. Other similar studies has since been done on ranking company co-occurrence in social networks [10], and more recently mapping the relations on banks in text using co-occurrences [11].

Another area of textual analysis that has gained popularity in financial news is methods for analysing sentiment. Researchers have shown that sentiment can be used to predict market movements [7] and that the platform and type of news sentiment have different impact and longevity.

In this research we are analysing, mapping, ranking, and visualizing the co-occurrence of companies in the S&P 500 index in crowd sourced financial news over the period 2011 – Q2 2016. We use a text-to-network approach to build networks of co-occuring company names, tickers, and parts of company names. This is done by matching regular expression of entities belonging to the S&P 500 to text in articles on a quarterly basis. Co-occurrences of company entities are then calculated for each quarter and from this we build undirected networks that show how companies are connected. We limit the scope of the research by looking at the top 25 companies based on a special closeness centrality measure. These networks are then visualized

You can also discuss how company connections change with time.

Minor: Labeled data like this isn't widely available

and qualitative analysis is done using graphs to support the quantitative measures.

For each quarter we create three different networks: one containing all quarterly data that we denote as a mixed network, one containing only data with positive (long) sentiment, and one containing data with only negative (short) sentiment. To determine the most important components in all the networks we calculate the flow of news between the nodes relative to the entire network. This is done using the closeness centrality measure information centrality, also known as current flow closeness centrality.

II. Data

There are many different types of financial text sources available that can argue for different perspectives on companies that also can have different biases. These biases are depending on how an author is positioned relative to a company as well as the mandated requirements for the report or article. The United States government for instance has a number of mandatory reports, such as the Q-8 and Q-10 that publicly registered companies have to file on an annual basis. There are also a number of registered analysts that follow companies and publish their analysis of the state of the companies accompanied by recommendations and suggestions. Then there is the main source of financial news, which is supplied by different companies such as Reuters, Bloomberg, Wall Street Journal and many others. They publish a mix of analyses, reports, and opinions on what has happened and what will happen. The authors of this type of articles are generally professional journalists and financial experts. Finally, we have a relatively new source of financial news, the crowd sourced news sites. Crowd sourced news sites are sites where anyone with relevant knowledge can post their own articles and analysis. Collectively the authors on such sites cover a very wide area of expertise, which is reflected in the multitude of articles.

The news data used to test our methods was gathered as part of the research. While there are available financial data sets online such as Reuters data set RCV1 [12], these data sets are generally only labeled according to what type of news they belong to and as such the labeling of the data source is of limited use in this kind of research. Further, the data sources for such news articles are quite heterogenous as they cover any type of news that happen to fit within relatively loosely defined categories.

The data we gathered consists of crowd sourced financial news articles. These are articles that have to pass an editorial check, which is a control that they meet certain guidelines. Other than that, the articles can be written by anyone regardless of background and intent. Although some research into crowd sourced data has been done [13, 14], this type of financial news has not yet been studied to the same extent as traditional financial news. At the same time, crowd sourced sites have steadily gained in popularity among investors and users in recent years. For this research we gathered a specific set of articles containing author sentiment from the site SeekingAlpha.com [15]. Other crowd sourced sites that recently have gained popularity among financial news platforms include Investing.com and Stocktwits.com.

The data gathered consists of 17398 financial news articles from 2006 to the end of the second quarter in 2016. The articles are written by about 3600 unique authors. The articles gathered and studied are split and labeled by the following sectors: technology, finance, health care, consumer goods, basic materials, and services. Each article has been labeled by the author as either positive (long) or negative (short), with about 25% of the articles in the data labeled as negative and about 75% of the articles labeled as positive. The content of each article is an analysis made by the author of the article on why he or she is either positive or negative regarding the targeted company, commodity, sector, index, or a combination of different components.

Data structured in this format has the potential of containing useful information that is not found or simply overlooked in a more generalizing collection of financial news articles such as the RCV1 [12]. The articles in this data set also express an expectation of the future instead of simple reporting on things that has already happened. The specific intent of the collection of articles is to explain where the authors of the articles expect the targeted companies, indexes, or commodities are headed going forward. The time frame for the sentiment is not defined and varies between articles from very short term to long term.

III. Models

In this part we discuss the specifics of the approach used in the research in greater detail. We explain how we have chosen to process the data, which analytical methods we use, and discuss other topics of interest related to our methodology.

A. Parsing

In order to limit the scope of the research we decide to only look at the companies listed in the S&P 500 index. The total number of companies, tickers, and components occurring in the news would otherwise number in the thousands. The components of the S&P 500 index itself varies over time as new companies emerge and other companies go bankrupt, merge, or get acquisitioned. We chose to conduct our analysis on the components of the index that were present in the index during May 2016. In total for this period there were 503 components represented in the index. Some of the components can be of the same company that has several classes of stocks, for example ”GOOG” and ”GOOGL” both represent Alphabet Inc. These tickers are then represented as the same company when the parsing algorithm finds either of the matches.

In order to parse the data in articles we set up a number of regular expressions that search for either the company ticker, for instance ”AAPL”, or the company name ”Apple Inc.”. Whenever either ticker, full name, or part of the name is matched in an article we treat it as an occurrence of the company in that article. For instance, an article mentioning ”Goldman Sachs” would be recorded even though the full name of the company ”Goldman Sachs Group Inc” was not fully matched. A company occurring several times in an article is assumed to have no extra value and is not recorded more than once.

Some problematic instances have been identified by the algorithm. One such instance is when article authors use an abbrevation of a company, writing ”HP” instead of ”HP Inc” or simply using the wrong ticker for a company that they are

Minor: Only counts once for occurences? Might we improve by keeping full counts?

writing about. Some authors can mistakenly think that ”HP” is the ticker of HP Inc. (HPQ) when the ticker HP actually stands for Helmerich & Payne, Inc. We try to limit exposure to such scenarios through additional conditional regular expressions.

When developing the parsing algorithm there are some trade-offs that need to be considered. Due to the nature of the similarity between some company names as well as ticker symbols having different length we take the decision to be quite strict while parsing occurrences. During trial runs we concluded that we would rather miss a few true occurrences and by missing them create false negatives, than create false positives through identifying companies that were in fact not represented in the articles. That choice is made because creating false positives could mean creating false patterns in the data, patterns that in turn could propagate into faulty conclusions. Having false negatives could mean not finding a pattern that really exist in the data.

The algorithm is run on a quarterly basis, but is not limited to any specific timeframe. No link between quarters is defined other than that we look for the same set of entities in each quarter. When it comes to financial data, capitalizing words and character case is very important. If we didn't take character case into account we could only identify entities based on full company name. The implication of this is that article authors that misuse capitalizations of ticker symbols or use some form of made up abbreviations will not be matched by the algorithm and again add to the list of false negatives.

B. *Co-occurrence, sentiment and centrality*

We continue by building co-occurrence networks on the parsed data from news articles on a quarterly basis starting from Q1 2011 until the end of Q2 2016. A co-occurrence network is an undirected network where in our case each company entity represents a node in the network. The links between nodes, the edges, are represented by the number of occurrences the two nodes have had together in articles during a specific quarter. A co-occurrence is defined as two company entities being matched in the same article.

A company without links to other companies, which means that it is never mentioned together with another company, would have no edges in the co-occurrence networks. A company that frequently appears in financial news would probably have connections to other companies that not necessarily can be seen simply from reading a specific article. Two seemingly unrelated companies can for instance be related to each other by both having close ties to a third company or a group of companies.

As we have extracted a sentiment value for each article we are able to distinguish between positive and negative co-occurrences. All occurrences in an article with positive sentiment are counted as positive occurrences, and vice versa for negative occurrences. In the data we gathered, there is no such thing as a neutral sentiment classification. To get a neutral representation, we instead combine both positive and negative occurrences into a mixed network. We thus build three networks for each quarter for an understanding of the difference between occurrences across sentiment. One consisting all positive occurrences, one containing all negative occurrences, and one containing a combination of positive plus negative occurrences.

Consider using words as edges instead of just co-occurence as edges

We label the network combining all instances as the mixed network.

Flow in our networks refer to how different nodes are connected to each other. Thus, the networks consist of weighted undirected links. With more data and more advanced language parsing, such as Google's Parsey McParseface, it would be possible to extract directed networks. This could be particularly useful if it was possible to identify pairs of occurrences where the sentiment differs depending on directionality and strength. Weighted edges between nodes in a network is generally a measure of how information spreads in the network. In the typical examples, such as the shortest path problem [16], an edge with a low value means the nodes are close to each other. In our case the inverse is true, the higher value an edge has the more information can travel through it. However, the flow we are interested in can be measured through different centrality measures. We are interested in a centrality measure that is able to account for all possible paths of news flow between all nodes in the networks, not only the shortest path.

Different types of centrality measures is what is generally used to compare nodes in a network. Relevant research on centrality in networks consists of: Identifying the most influential people in social networks [17], identifying how companies and their market cap is related in social networks [18], and identifying how banks are related in text [11].

Degree centrality is one of the simplest centrality measures, which uses the number of edges that a node has to measure centrality. Betweenness centrality is a centrality measure that represents how nodes in the shortest path between two nodes are connected. Eigenvector centrality is an influence measure that assumes some nodes contribute more to the measure than others, of which Googles PageRank algorithm is one version. Closeness centrality calculates the distance between nodes through the shortest path to determine which nodes are important. Information centrality is a version of closeness centrality that uses harmonic mean of resistance in a graph to calculate importance of nodes.

We are interested in finding out how information about one company affects other entities in the network, not simply the shortest path between nodes. This means that most of the centrality measures previously listed are not fitting. Information Centrality is one of the measures that allows us to measure all paths. We use information centrality in order to rank the entities in our networks. We use the same formula for calculating the centrality as in [11] with minor modifications to fit the different types of networks.

Nodes in the network are represented by n. For the mixed network we take all the occurrences:

$$n = n_{short} + n_{long}$$

For the long network we count only the positive occurrences:

$$n = n_{long}$$

For the short network we count only the negative occurrences:

$$n = n_{short}$$

Information centrality for the networks is then calculated as in [11]:

$$I(i) = \frac{n}{nC_{ij} + \sum_{j=1}^{n} C_{jj} - 2\sum_{j=1}^{n} C_{ij}}$$

In the pseudo-adjacency matrix, *w* is the link weight between nodes and *S(i)* is the node weight. That gives us the following formula:

$$C = B^{-1}, B_{ij} = \begin{cases} 1 + S(i), & if \ \ i = j \\ 1 - w_{ij}, & otherwise \end{cases}$$

To be able to rank company entities relative to each other we need to be able to calculate information centrality for all nodes in the network. The implication is that we need all the nodes to be connected in order to perform the calculations. As a side effect of having limited data and a large number of firms, there is a possibility that for each quarter we will have more than one network that are not directly linked to each other. To account for this, we use Laplace smoothing as proposed by [19]. This allows us to connect all nodes but also reduces the effects of false negatives on the network. A lower smoothing value, such as less than 0.5, allows the network to keep more of its characteristics, while a higher smoothing value would be mostly useful in comparing relative importance. We choose to test two different smoothing values of 0.1 and 1.0 to see if different values have a significant effect on the relative ranking of companies.

In order to get a better understanding of what the ranking represents we normalize the flow in the networks by market capitalization and we normalize information centrality to be between 0 and 1 as the centrality values are not linear. As the market capitalization of entities varies over time, we decide to use the market capitalization at the end of each quarter for the purpose of normalization of entities. A normalized ranking should represent a ranking of companies with proportionally the highest news flow. The non-normalized ranking better represents an absolute ranking where bigger companies get more space in financial reports, news, and statements. For the normalization we then construct the following formula:

$$J(i) = \left(\frac{I(i) - I_{min}(i)}{I_{min}(i) + I_{max}(i)} \right) * \frac{1}{m}$$

Where $J(i)$ is the normalized information centrality of node *i* and *m* the market capitalization for the node at the end of that specific time period. I_{min} is the minimum information centrality value among values for that quarter and I_{max} is the maximum information centrality value in the quarter.

C. *Qualitative analysis through visualization*

To give an overview of the networks as they change over time and to supplement quantitative centrality measures, we provide graphs as a visual representation of the networks. The visualizations are used to help give an understanding of how the data changes over time that would be hard to recognize by only viewing centrality measures or the raw data per se. Comparative analysis of quarterly cross-section snapshots, and color coding nodes based upon sectors, provides a dense view to a large amount of information in an easily understandable format.

As we parse in total approximately 500 firms, and hence have as many nodes in our network, we limit the scope to visualizing the top 25 companies according to information centrality for each quarter. This focuses our scope enough to take a closer look at which companies are regularly present in the top 25 positive, negative, and mixed co-occurrence networks. Further, in the visualizations we colour nodes by sectors of interest. We define the width of the edges in the visualizations of networks according to number of occurrences, and set the size of the nodes according to the total sum of the occurrences for the company in that quarter. This helps us illustrate how both individual companies and entire sectors are affected by the changes in news flow.

To find more specific patterns, we provide three comparisons between positive and negative networks for 2013Q4, 2014Q1 and 2014Q2 (see Figs. 1-3). An example of overall patterns would be the prominent but segmented position of technology and finance companies in the positive network in 2013Q4, whereas the negative network shows a more central role for consumer companies and a more complete network.

Likewise, one can observe that energy companies emerge as a fairly isolated segment in the positive network of 2014Q1.As it coincides with the oil price crash, what is interesting here, as shown in Fig. 2, is that instead of energy companies populating the top of the negative ranking, they instead jumped into the top of the positive ranking. Suggesting that investors sentiment regarding a certain cluster of energy companies in fact increased with the price drop. Assessing one quarter ahead, Fig. 3 shows again in 2014Q4 a less prominent position in both networks for energy companies.

IV. QUANTITATIVE ANALYSIS THROUGH CENTRALITY

For the quantitative part of the research, we look at and compare the top 25 ranking companies in the S&P 500 index according to information centrality. We test Laplace smoothing of 0.1 and 1 to see if changing smoothing has an effect on the average ranking as shown in Table 1. We can also in Table 1 see the companies that consistently can be found among the top 25 most central nodes for each of the three different network types: positive, mixed, and negative. Choosing a limitation of 25 allows us to capture not only the companies that consistently appear in the system but also some of the companies that move up and down in the ranking to see general trends each quarter.

	Average rank top 25					
	Laplace 0.1			Laplace 1.0		
Pos.	*Positive*	*Mixed*	*Neg.*	*Positive*	*Mixed*	*Neg.*
1	AAPL	AAPL	AAPL	AAPL	AAPL	AAPL
2	MSFT	GOOG	AMZN	MSFT	GOOG	GOOG
3	GOOG	MSFT	GOOG	GOOG	MSFT	AMZN
4	AMZN	AMZN	MSFT	AMZN	AMZN	MSFT
5	INTC	INTC	NFLX	INTC	INTC	NFLX
6	FB	FB	WMT	FB	FB	WMT
7	IBM	WMT	FB	IBM	IBM	FB
8	WMT	IBM	INTC	WMT	WMT	INTC
9	BAC	NFLX	IBM	BAC	NFLX	IBM
10	GS	BAC	YHOO	HPQ	BAC	YHOO
11	HPQ	YHOO	MS	WFC	HPQ	MS
12	WFC	GS	VZ	CSCO	CSCO	VZ
13	CSCO	HPQ	TGT	GS	YHOO	TGT
14	YHOO	CSCO	HPQ	JPM	GS	HPQ
15	JPM	MS	ORCL	YHOO	ORCL	ORCL
16	KO	ORCL	CSCO	ORCL	MS	CSCO
17	ORCL	JPM	T	IP	JPM	T
18	IP	WFC	GS	KO	VZ	CRM
19	BRK	IP	CRM	C	WFC	GS
20	NFLX	KO	JPM	NFLX	IP	JPM
21	MS	VZ	QCOM	GM	KO	QCOM
22	C	QCOM	DIS	MS	T	DIS
23	GM	C	IP	JNJ	QCOM	BBY
24	JNJ	T	BBY	VZ	C	COST
25	QCOM	GM	COST	BRK	GM	MCD

Table 1. Top 25 average rank for all companies by ticker from the S&P 500 list over the 2011 – Q2 2016 period. Split into two sets of Laplace smoothing of 0.1 and 1.0. Differences in rank between especially negative and positive networks are evident. The relative rank of companies doesn't significantly change using different smoothing coefficients.

From the order of entities in Table 1. we observe a number of patterns that could be of interest. We can look at the order of the companies that consistently appear among the highest ranking nodes. We can also follow individual entities that change in rank over time to see a general shift in centrality for that entity. In case the entities appearing in the top ranking barely move in terms of rankings, one could instead follow specific companies that fall or gain in rank between quarters. Third, we can compare companies between the different networks built on different sentiment to see which companies are found in one but not the other.

Average count of the edges in the different networks can be useful if we are interested in knowing the spread between occurrences. We get 1.76 average edges for mixed, 1.64 for the positive, and 1.49 for the negative networks. We conclude that the networks are small-world networks as defined in [16]. We find that the frequency of the articles changes as the data source grows with time and when volatility in the markets go up the frequency of articles tend to increase. From that we can conclude that the average number of edges also grows with time using this data source, which is partly but not only related to the number of published articles.

For a better understanding of what the implication of changes in information centrality are, we plot the 25 highest ranked firms on a quarterly basis for each network. As can be seen from Figures 4-6, the centrality measures are on average slightly higher for the positive network than for the negative network. Understandably that difference can be attributed to the positive network containing a higher edge average. However, the angle of the trends between quarters seem to be containing new information. For instance, for the period around 2015Q2, the negative network has a steeper upward angle than the positive network. And in the period from the start of 2016Q1 until the end of 2016Q2, the positive network has a steeper upward angle than the negative network. That could be implying that the change in author sentiment is reflected differently in different networks. Moreover, information centrality in itself captures also properties of the network structure, rather than just edge average. This points to the fact that the peaks in centrality in 2015Q2 and 2016Q2 are also attributed to the network structure processes that increases connectivity along multiple paths.

In Table 2, we see that Apple Inc. is the only company that appears as top ranked in the positive network. Bank of America Corp appears once and Apple Inc. the rest 21 times at the top of the mixed rankings. In terms of negative ranking, we have Amazon.com, Inc. appearing four times as the most central, as it is generally more volatile. In the negative ranking Netflix Inc. and Alphabet Inc. both appear as the highest ranked one time each. The rest of the 16 quarters Apple Inc. again is represented as the highest ranked company.

As Apple Inc. seems to dominate the rankings we decide to look at a way of normalizing the ranking based on the size of the companies. In Table 3, we recalculate the ranking using two normalizations. The first is a direct normalization by market cap. As expected, we can observe that this favours smaller companies in the S&P 500 index as the information centrality measure is not linear. The typical range in our information centrality measurements have been between 0.11 and 0.05. Qorvo Inc. places first in this second ranking even though it has few occurrences, suggesting that normalizing only by market cap does not necessarily provide a stable ranking. Not to favour smaller companies, the second approach is done by first normalizing information centrality to between 0 and 1. After that we normalize the ranking by market cap, which provides a more

Quarters with highest flow per company				
Company	*Positive*	*Mixed*	*Negative*	*Total*
AAPL	21	22	16	59
AMZN	0	0	4	4
GOOG	0	0	1	1
BAC	1	0	0	1
NFLX	0	0	1	1

Table 2. Companies with highest information flow per network. Apple dominates all networks. Amazon has the second highest total due to ranking first four times in the negative network. Alphabet places mostly in second place just after Apple and is not showed as ranking first more than once.

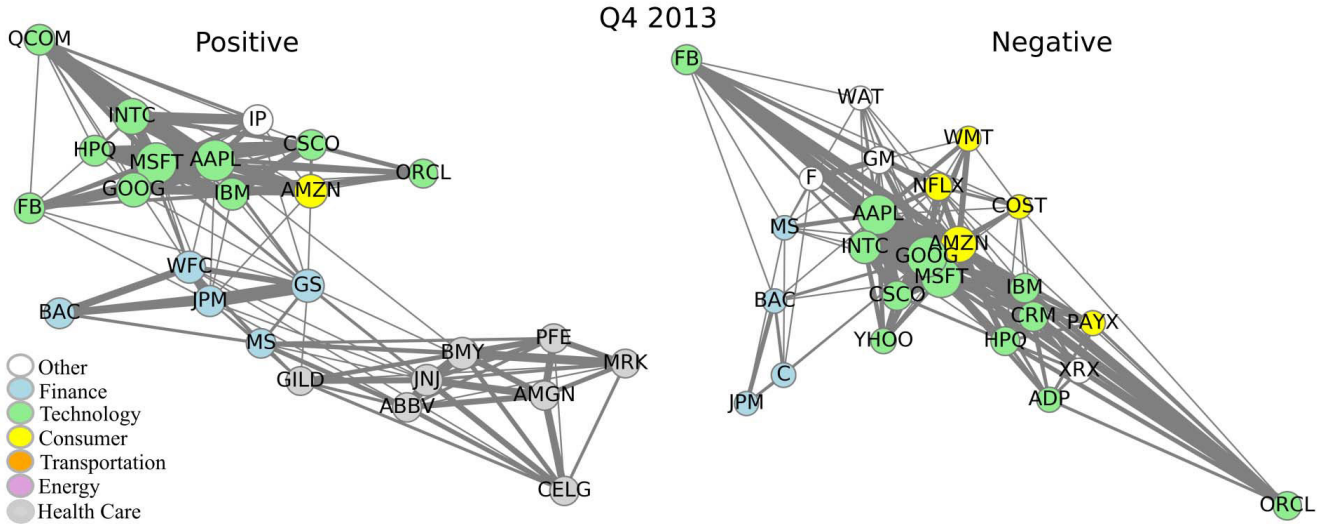

Fig. 1. Quarter 4 2013 Co-occurrence networks for top 25 positive and negative nodes. Several health care sector companies represented in the top 25 positive nodes but not in the negative indicating something news worthy is happening in that sector. Thicker edges represent more co-occurrences between two entities and larger sized nodes indicate larger total sum of company occurrences. Length of edges in the networks are not representing any added value.

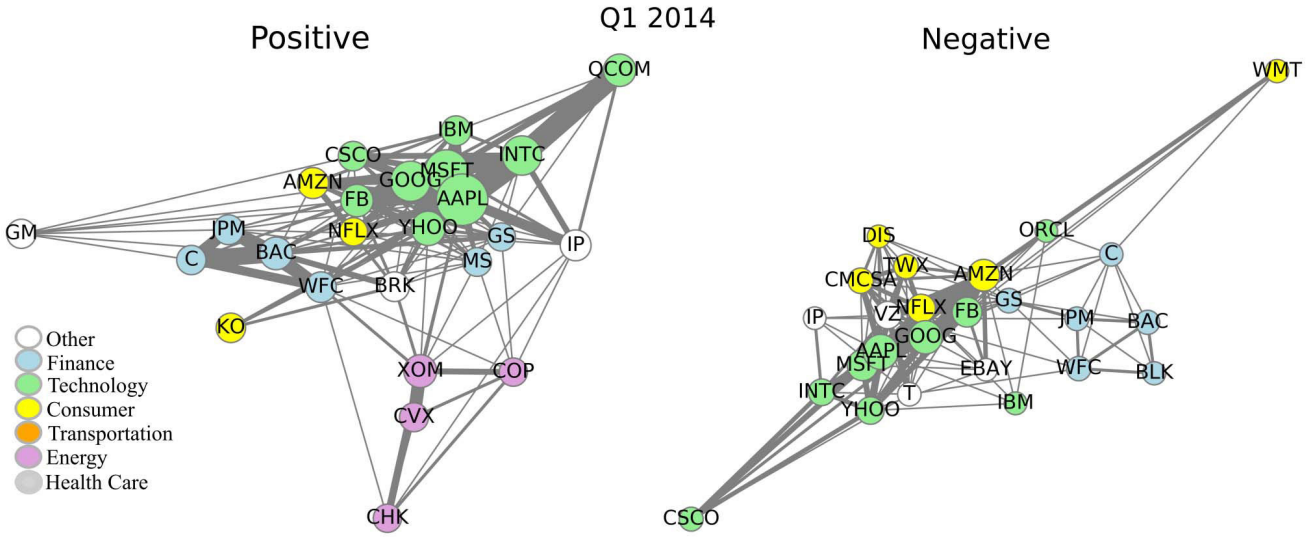

Fig. 2. Quarter 1 2014 Co-occurrence networks for top 25 positive and negative nodes, 4 nodes from energy sector jumps into top 25 of positive but not negative as a response to falling oil prices while health care sector falls from top 25. The default assumption would have been that reduced performance of the energy sector due to lower oil prices would be reflected in negative sentiment.

stable ranking with a mix of both smaller and larger companies. The normalization of market cap is based on companies' market capitalization at the end of each quarter. From Table 3, we see that when recalculating the ranking Apple falls from first place to not being in the top 25 ranking anymore. In the double normalized ranking Netflix Inc. places at top of the average ranking for all three networks.

CONCLUSION

Determining which entities in a stock market that are most central can be used as a way of understanding which companies are the market movers. That in turn can be used to model risks and which companies have an unexpectedly high news flow compared to for example their market cap. By using a sentiment, co-occurrence, and information centrality ranking we can also determine which companies influence each other and which companies don't have any direct relations in news.

We have developed a ranking system that uses financial data to rank the companies in the S&P 500 index both for positive and negative news. The ranking system was tested on a quarterly basis but could just as well be used on shorter or longer time frames. The ranking changes depending on the available data as well as the efficiency of parsing algorithms. Further, the absolute

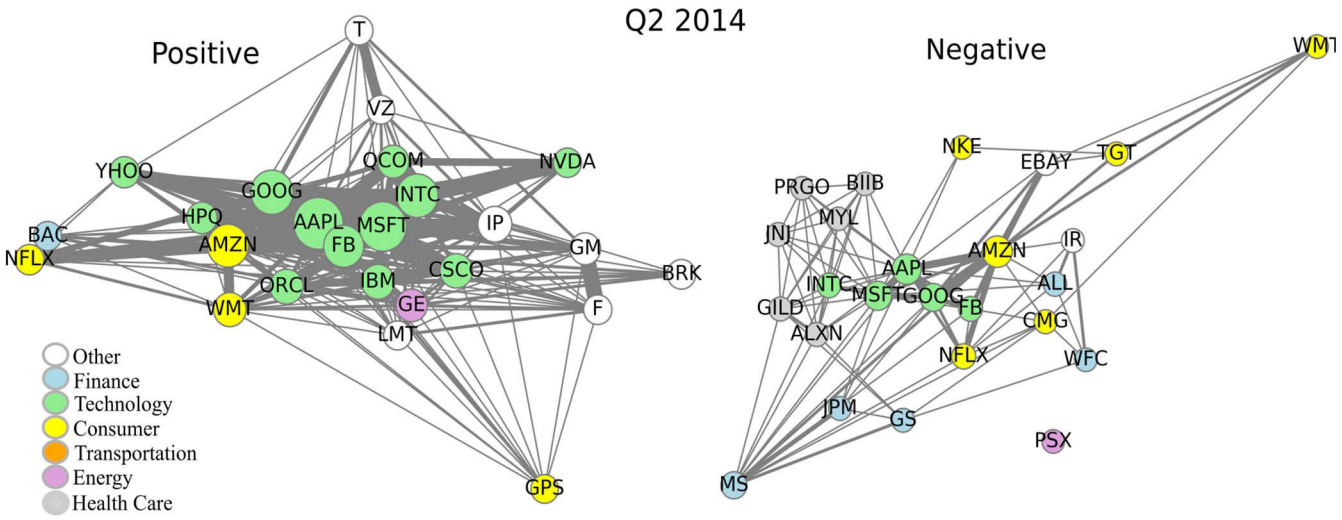

Fig. 3. Quarter 2014 Co-occurrence network of top 25 positive and negative companies. Oil related energy companies fall from the top 25 representation again and health care sector companies enter the negative top 25. PSX shown as alone in the network because the Laplace smoothing links of 0.1 are not drawn to reduce noise. That means that Phillips 66 (PSX) is currently the most negative of another cluster of nodes only loosely connected to the larger network.

measures in the ranking can change with time as more news becomes available. While the ranking can be used to measure how much visibility a company is getting, the relative value to other companies is more interesting as the absolute values change with available news for a given period.

The developed rankings provide two useful, yet different, approaches. The first non-normalized ranking shows representations of which companies have the highest absolute ranking based on news flow. The second type of ranking measure deviations of negative or positive news flow relative to their size. With further extensions, the rankings could potentially be used to identify companies that are signalling weakness or strength. Likewise, the sentiment and co-occurrence information could be aggregated to provide a joint measure of sentiment in the network as a whole.

Just as the market conditions change over time as we have seen with a long period of low interest rates since 2009, so does the news regarding the market. There seem to be more useful information here that can be extracted, and methods that can be further developed. The sentiment for each article in the research could be automatically calculated and then compared to the authors own labelling. Further, we could then compare for example the different Reuters data sets with our current data set to find out if crowd sourced article sentiment match sentiment in other types of financial articles. Further research could also be done on using the ranking as basis for an automated trading system for either long only or a long/short approach. Another research avenue would be to look more into company specific risks and model which news specific companies are vulnerable to.

ACKNOWLEDGMENT

Thomas Forss thanks Liikesivistysrahasto and Handelshögskolan at Åbo Akademi for the financial support.

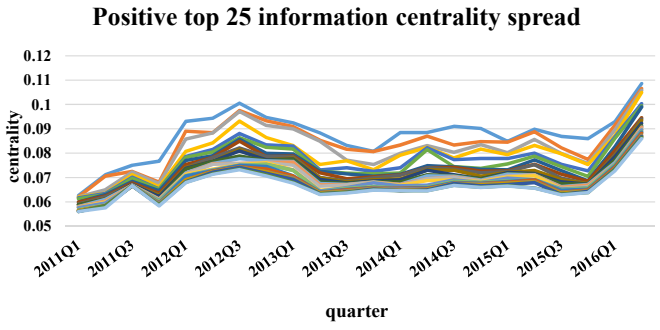

Figure 4. Line graph of information centrality for the top 25 firms in the positive network for each quarter over the entire time frame

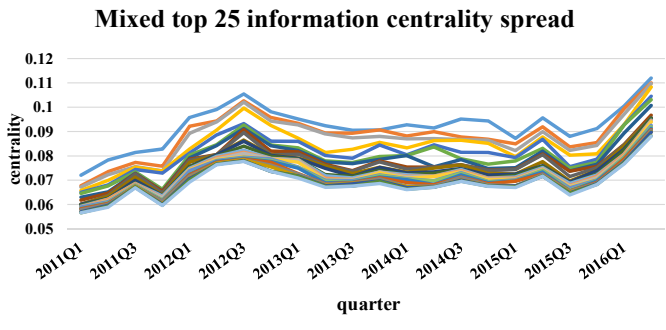

Figure 5. Line graph of information centrality for the top 25 firms in the mixed network for each quarter

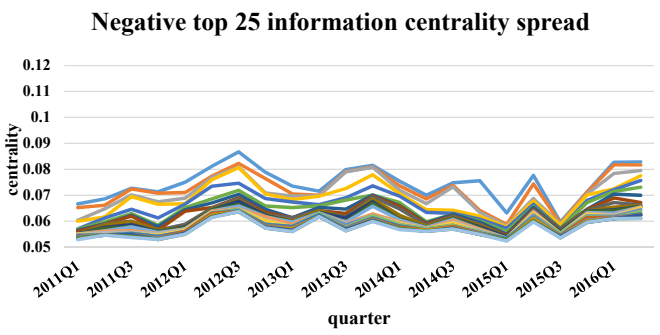

Figure 6. Line graph of information centrality for top 25 negative firms for each quarter

1	QRVO	QRVO	FSLR	NFLX	NFLX	NFLX
2	WRK	WRK	DNB	NVDA	BBY	BBY
3	URI	FSLR	PKI	BBY	NVDA	NVDA
4	FSLR	URI	OI	IP	IP	YHOO
5	DNB	DNB	GT	QRVO	YHOO	IP
6	PKI	PKI	LM	GPS	GPS	CMG
7	PBI	PBI	NFX	DNB	DNB	DO
8	OI	OI	HBI	YHOO	SPLS	SPLS
9	AIZ	AIZ	URBN	JNPR	JNPR	AVGO
10	AVY	AVY	PHM	FSLR	AVGO	M
11	ETFC	FLIR	SEE	SPLS	CMG	CRM
12	FLIR	ETFC	MLM	SEE	FSLR	JNPR
13	TE	HAR	SNA	AVGO	SEE	MLM
14	HAR	TE	CVC	DPS	QRVO	DNB
15	PDCO	PDCO	TSS	UA	DO	TGT
16	LEG	LEG	GAS	EA	UA	RIG
17	GT	GT	FTR	MU	EA	KSS
18	LM	LM	FL	HPQ	MU	AMZN
19	PBCT	PBCT	NDAQ	GT	M	GPS
20	NFX	ZION	AN	HRS	GT	HPQ
21	URBN	NFX	ENDP	DO	HPQ	TWC
22	ZION	URBN	TSO	AA	CHK	MS
23	HBI	HBI	LVLT	CHK	COH	COH
24	PHM	TGNA	DO	CMG	CRM	SEE
25	TGNA	PHM	SWKS	HRB	DPS	CA

Table 3. Top 25 average rank for all companies from the S&P 500 list over the 2011 – Q2 2016 period when normalized. To the left we normalize by market cap, to the right we first normalize information centrality between 0 and 1 and then we normalize by market cap.

REFERENCES

[1] Capon, N., Farley, J. U., & Hoenig, S. *Determinants of Financial Performance: A Meta-Analysis*. Manag. Sci., vol. 36, no. 10, pp. 1143–1159, Oct. 1990.

[2] E. I. Altman. *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*. J. Finance, vol. 23, no. 4, pp. 589–609, Sep. 1968.

[3] D. B. Keim and R. F. Stambaugh. *Predicting returns in the stock and bond markets*. J. Financ. Econ., vol. 17, no. 2, pp. 357–390, Dec. 1986.

[4] R. R. Trippi and E. Turban, Eds., *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance*. New York, NY, USA: McGraw-Hill, Inc., 1992.

[5] B. Back, T. Laitinen, and K. Sere. *Neural networks and genetic algorithms for bankruptcy predictions*. Expert Syst. Appl., vol. 11, no. 4, pp. 407–413, Jan. 1996.

[6] P. Sarlin and T. A. Peltonen. *Mapping the state of financial stability*. J. Int. Financ. Mark. Inst. Money, vol. 26, pp. 46–76, Oct. 2013.

[7] J. Bollen, H. Mao, and X. Zeng. *Twitter mood predicts the stock market.* J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011.

[8] P. S. Jacobs and L. F. Rau. *SCISOR: Extracting Information from On-line News*. Commun ACM, vol. 33, no. 11, pp. 88–97, Nov. 1990.

[9] A. Özgür, B. Cetin, and H. Bingol. *Co-occurrence network of reuters news*. Int. J. Mod. Phys. C, vol. 19, no. 5, pp. 689–702, May 2008.

[10] Y. Jin, M. Ishizuka, and Y. Matsuo. *Ranking Companies Based on Multiple Social Networks Mined from the Web*. INTECH Open Access Publisher, 2010.

[11] S. Rönnqvist and P. Sarlin. *Bank networks from text: interrelations, centrality and determinants*. Quant. Finance, vol. 15, no. 10, pp. 1619–1635, 2015.

[12] Lewis, D. D., Yang, Y., Rose, T., and Li, F. *RCV1: A New Benchmark Collection for Text Categorization Research*. Journal of Machine Learning Research, 5:361-397, 2004.

[13] Zhao, Y., & Zhu, Q. (2014). *Evaluation on crowdsourcing research: Current status and future direction*. Information Systems Frontiers, 16(3), 417-434.

[14] Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., & Zhao, B. Y. (2015, February). *Crowds on wall street: Extracting value from collaborative investing platforms.* In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 17-30). ACM.

[15] "*Stock Market Insights*," *Seeking Alpha*. [Online]. Available: http://seekingalpha.com/. [Accessed: 27-May-2016].

[16] Floyd, R. W. (1962). *Algorithm 97: shortest path*. Communications of the ACM, 5(6), 345.

[17] K. Stephenson and M. Zelen. *Rethinking centrality: Methods and examples*. Soc. Netw., vol. 11, no. 1, pp. 1–37, Mar. 1989.

[18] Y. Jin, Y. Matsuo, and M. Ishizuka. *Ranking Companies on the Web Using Social Network Mining.* Web Mining Applications in E-commerce and E-services, I.-H. Ting and H.-J. Wu, Eds. Springer Berlin Heidelberg, 2009, pp. 137–152.

[19] "*An empirical study of smoothing techniques for language modeling*." [Online]. Available: http://dl.acm.org/citation.cfm?id=981904. [Accessed: 02-Jun-2016].

[20] Watts, D. J., & Strogatz, S. H. (1998). *Collective dynamics of 'small-world' networks*. nature, 393(6684), 440-442.