Mining Business Relationships from Stocks and News

Thomas Kellermeier[1], Tim Repke[1], and Ralf Krestel[1,2]

[1] Hasso Plattner Institute,
University of Potsdam, Germany
thomas.kellermeier@student.hpi.de
tim.repke@hpi.de
[2] University of Passau, Germany
ralf.krestel@uni-passau.de

Abstract. In today's modern society and global economy, decision making processes are increasingly supported by data. Especially in financial businesses it is essential to know about how the players in our global or national market are connected. In this work we compare different approaches for creating company relationship graphs. In our evaluation we see similarities in relationships extracted from *Bloomberg* and *Reuters* business news and correlations in historic stock market data.

Keywords: Market Analysis · Text Mining · Entity Relationships.

1 Introduction

Financial markets play a fundamental role in today's global market economy. The strategic decision making process nowadays is supported by large-scale data analysis and by monitoring stock markets, decision makers can learn a lot. By buying and selling stocks, investors influence their value, e.g. by offering more when demand is high, they drive the price per share up. Following the Efficient Market Hypothesis [2, 9], the investments can be interpreted as trust in positive future performance and in sum as an approximation of a company's intrinsic value. In the past, large efforts have been undertaken to understand and partially foresee the temporal evolution of stock markets and thereby reduce the risk of such investments for investors.

In this paper, we aim to identify relationships between publicly traded companies based on similar behaviour of their stock price movements and their mentions in business news.

Knowing about these inherent links may guide and support further analyses. For example, when assessing the credit risk of a new corporate client, the risk officer compares a number of factors to similar prior cases. To this end, we propose methods to construct weighted similarity graphs based on stocks and news texts that can be used to measure the similarity of companies.

Thoughtful preprocessing and feature selection is an important task for NLP [6]. By having methods to test or support NLP tasks with relationship information

from stock markets, more sophisticated models can be developed. Some potential applications for improving text-based features in this context: Concept maps [22]; bag-of-keywords [27]; sentiment WordNet [34, 17]. Incorporating relational features into artificial neural networks, most notably node embeddings or a graph convolutional layers has been shown to improve models in many areas of application [5].

Another use case for using business relationships and stock relationships is the observation and analysis of financial markets. In order to measure the value and credit risk of a company, competitors, suppliers, subsidiaries and other related companies might be considered to give a better assessment. Both business and stock relationships can be incorporated into a corporate graph and therefore support in understanding the complex net of relationships among companies.

In this work we provide an overview of approaches to generate weighted company relationship graphs from both text and stock data. Our comparison of these features yields important information about applicability of these approaches in different contexts.

2 Related Work

There are numerous approaches to model economic variables, for example utilizing econometrics, Natural Language Processing (NLP) and machine learning. Predicting a stock's future behaviour is among the main research topics. Although the stock market's future is unforeseeable, this problem receives great attention. It is a controversial topic and lacks a good theoretical foundation to justify forecasts on finance markets. As a naive example, one could take a model that always predicts rising stock prices – given enough time, this model will show great precision/recall scores thanks to global economic growth, which is only interrupted by crises where not even experts manage to project reliable time frames. By ignoring these issues and applying machine learning approaches in sandbox like environments, studies often do not fully reflect the problem space and therefore erroneously promise positive results. The related work presented in this section includes stock prediction, however we focus on their underlying methods and data to model financial markets.

With the high interest in prediction models a vast amount of data is available to researchers. Some studies focus on predicting the trend of stock prices via technical indicators like Bollinger Bands, momentum or Moving Average Convergence Divergence (MACD) [21, 23, 1]. Others consider the whole market instead of separate stocks by predicting a stock index or volatility index [6]. In this work we propose an approach to identify linked stock behaviour, so we focus on separate stocks instead of technical indicators or market indices. Although the studies above target other objectives, they need to tackle similar problems of data preparation to normalise the noisy and stochastic nature of stock markets.

Correlating stock prices requires intensive regression analysis, which is related to predictability of time series in the context of econometrics and statistics. Methods like cross-correlation, mutual information and Granger Causality are

proposed to measure and compare financial time series. Often enough, machine learning models are applied without dealing with systematic errors in the raw economic data in the assumption that those models will learn to ignore problematic issues like noise on their own. Studies with more economical or statistical background investigate the characteristics and quality of economic variables.

The selection of the dataset itself plays a crucial role in the feasibility of a classification or regression problem. Sun et al. [30] report 70 % accuracy for their matrix factorization model during training but only 51 % on test data which is not a significant improvement compared to random guessing. Lee et al. [21] present a Random Forest Classifier for predicting separate stock prices with an accuracy 22.2 % higher than random guessing. Although results like these sound promising, it could be the consequence of an unsafe evaluation. Among various factors influencing an experiments outcome, models without cross-validation are prone to the "lucky sample effect", as demonstrated by Hsu et al. [16]. The time series might be lacking ergodic properties and therefore report erroneously high accuracy.

Kim et al. [19] apply a rule-based classifier on stocks from the energy sector of the US stock market. They calculate the cross-correlation among pairs of stocks and predict trends by considering the lagged stock price of another highly cross-correlated stock. Even though they select only highly cross-correlating pairs, potential spurious correlation are not considered which might arise due to unfiltered autocorrelations. As already pointed out by Granger et al. [13], misspecification like omitted variables or autocorrelations can lead to spurious regression. Ruiz et al. [29] exploit these auto-dependencies of stock prices by training Auto Regression and Vector Auto Regression models with the Ordinary Least Squares (OLS) method to predict the daily closing price. Their approach combines stock price data with numerical features extracted from Twitter, e.g. number of retweets. They do not inspect their data for homoscedasticity, which is one of the requirements of the Gauss-Markov theorem for OLS [14].

Instead of proving the predictability of a new introduced features by feeding it into a prediction model, Vlastakis et al. [31] apply a regression analysis between the demand for market-related information and market variables like volatility and stock prices. The overall information demand is represented by Google's Search Volume Index for the search keyword S&P 500 and they conclude, that some relationships exist with a high certainty.

Kosapattarapim et al. [20] presents a very detailed procedure on inspecting Granger Causality between the stock exchange index of Thailand and the exchange rate between Thai Baht against US dollars. To ensure that the preconditions hold, they inspect unit roots and co-integration relationships. Their results indicate an unidirectional causality from stock prices to exchange rate.

For a more conscious inspection of non-linear auto-dependencies in finance time series, Dionisio et al. [7] compare the normalised mutual information with Pearson's r for several stock price indices. Referring to related work, they recall the assumption of a strong relationship between entropy, dependence and predictability. To exclude the linear auto-dependencies, they filter the data by

taking the residuals of an autoregressivemoving-average process. Unlike the linear correlation, mutual information still indicates a significant dependency on the residuals without any foreknowledge on the theoretical probability distribution or the type of dependency.

Since external information related to stock markets is not directly observable, meaningful proxies are extracted from unstructured data like forum posts, news, social media and SEC filings. The most popular text sources are financial news [27, 17, 34] because they are expected to represent new events influencing the financial market. Their importance is mostly reflected within the stock prices of the directly following days and loses its meaning over a longer time period [6]. Therefore, many scientists incorporate news for short-time prediction models [18]. They consider the news and the previous daily stock prices to predict the intraday price movement for the next day. Various methods for extracting abstract representations of news have been proposed over the last years.

Ding et al. [6] compare linear and non-linear approaches for prediction using events-based document representations without incorporating historical stock price data. They extract events by applying Open Information Extraction (OIE) on news articles from *Reuters* and *Bloomberg*. Thereby, each article is transformed into a tuple of subject, predicate verb and object.

Previous work usually only considers one company for predicting its future stock price as pointed out by Akita et al. [1]. They instead feed related articles and historical prices of ten companies within the same industry at once into a LSTM to predict the close prices of all ten companies by regression analysis. *Nikkei newspapers* are preprocessed using Paragraph Vectors which learns fixed-length feature vectors from variable-length texts. Their market simulation indicates that incorporating multiple companies from the same industry is very effective for stock price prediction.

A recent approach by Chen et al. [5] explores the setup and application of a corporate graph for a prediction model. A graph is proposed which contains nodes representing stock companies from Shanghai Stock Exchange and Shenzhen Stock Exchange and their shareholders. The weighted edges between those nodes indicate the shareholding ratio. They conclude that such relational data can improve the performance of stock prediction.

3 Relationships from News

In this work we propose to construct two company relationship graphs, one based on stock price movements, and the other on financial news. The literature on stock prediction has shown, that news articles impact financial markets [18] and can be used as a proxy for business relationship. If an announcement was made or any relevant information like a SEC filing was released by a company, financial news report it as fast as possible. Longer reports put it even into a bigger picture, provide some background information and refer to possible competitors. In this section we describe how we construct a relationship graph from a collection of news articles. Therefore, we introduce an effective baseline to extract company

names and match them to their respective stock symbol. Furthermore we discuss different approaches to assign weights to relationship edges in the graph.

Data. We use the financial news dataset [3] released by Ding et al. [6]. It contains news articles from Reuters [4] (106,519 documents) and Bloomberg [5] (448,395 documents) covering the time period from 2006-10-20 to 2013-11-26. We discard duplicate articles as well as articles with less than 300 characters.

After filtering, there are 542,517 articles left. While Reuters articles are equally distributed over all covered years, most of the Bloomberg articles in our dataset were published after 2010.

Building a Relationship Graph. We assume that companies related in any context will be mentioned together in at least a few articles. In the following, the simultaneous mentioning of two companies within one article is called co-occurrence.

Fig. 1. Reuters article *"FedEx cancels Airbus A380 order, switches to Boeing"*, published on 2006-11-07, with recognised named entities.

Named Entity Recognition. First of all, we have to identify occurrences of every mentioned company. We observe a high recall of identified organisations. Although some entities might be falsely identified as an organisation, we see that later processing steps usually filter them out. We use SpaCy[6] to extract entities and only keep those that are classified as an organisation, which in the definition of SpaCy could be any company, agency or institution. Of the originally 40.7 million entities recognised in the over 500k articles, we keep 9.6 million organisation entities for further processing. Note, that these company mentions have numerous aliases and are not written consistently. For example, some authors might use a product of the company (e.g. *Google* instead of *Alphabet*) or just a representative such as the CEO (e.g. *Steve Jobs* instead of *Apple*). With the help of an aspect based product-centric model [24], we tested the performance of

[3] https://drive.google.com/drive/folders/0B3C8GEFwm08QY3AySmE2Z1daaUE
[4] https://www.reuters.com
[5] https://www.bloomberg.com
[6] https://spacy.io

our graph construction model using proxies like products or representatives. We found 19k occurrences of people as representatives of companies. When using those in our evaluation, we saw now no overall improvement. Thus, only direct mentions without proxies are considered in the remainder of this work.

Named Entity Linking. Out of all organisation entities, only those are of interest which can be linked to a company from the S&P 500 market index. The official company name is provided by the stock dataset and linked to the correct stock prices by its stock symbol. The names contain suffixes like *Inc.* or *Limited* which are usually left out in the news. To establish the link between occurrences in news and the full corporate name, we use regular expressions to normalise names and remove extensions.

If the regular expression reduced both strings to their least common sequence and they are completely equal, the extracted entity is assumed to match the examined stock company. In the end, 436k related company names were extracted which are distributed over 127k articles. Because occurrences were only found for 443 companies, the remaining companies are removed.

Sometimes, Bloomberg articles already include stock symbols in parentheses following the company mentions, as can be seen from Fig. 1 on the preceding page. Because this is the case for roughly 10 % of the cases, we can not solely rely on this information for linking entities. Instead, we use it for evaluating our approach since the information was added by the author and therefore is assumed to be accurate. Each stock symbol in parenthesis is compared to the stock symbol of our linked entity, resulting in over 99.8 % matches for the previously extracted 436k occurrences.

In the example in Fig. 1 on the previous page, two of the four organisations are linked to stock prices, whereas the remaining two are not considered since they are not components of the market index and thereby not contained by the data covered in this work.

Weighting Co-occurrence Edges. As mentioned before, we consider the co-occurrence of company mentions in the same article as a relationship. For more fine-grained interpretation of the resulting relationship graph, we add weights to the edges for which we propose three different metrics. First, we consider *Number of Articles* a co-occurrence appears in. This feature does not account for the number or distance of occurrences for two companies within one article. It rather measures the co-occurrences across the whole text corpus instead of weighting the connection within one article. The *Minimum Distance* takes the intra-article connection into account by calculating the distance for each possible pair of company names within one article. The shortest distance is kept for this article and averaged across all other distances in the corpus. This metric is based on the assumption that direct comparisons are drawn for strongly connected companies, which should be reflected by smaller distances on average. Lastly, the *Pairwise Distance* is a more sophisticated approach. It accounts for the multiple inter- and intra-article co-occurrences for which we use a scan line algorithm to traverse all mentions of these two companies in an article and pair these up while

avoiding a too high distance. Similar to the previous approach, all distances are averaged but, in addition, each pair is considered instead of only considering the best one for each article.

We evaluate the different weighting approaches later by comparing the resulting graphs to the relationships we find in stock data.

4 Relationships from Stocks

The second company relationship graph we use in our work is based on stock price movements. The naive approach to define relationships is the correlation between the time series of indicators such as daily open and close values of two stocks. However, the general performance of a marketplace influences the individually traded stocks and therefore simple approaches will find strong pairwise correlation between all stocks. Therefore, we have to ensure that the time series are independent of exogenous variables and free of autocorrelation and heteroscedasticity to the greatest possible extent. Such malicious properties distort statistical inference resulting in meaningless findings. In econometrics, statistical methods are applied to financial time series to deal with spurious correlations [33] and conclude with meaningful cross-correlation coefficients. In the following, we refer to cross-correlation when talking about correlation. In this section we will describe the data, methods to normalise the time series for removing external influences, and how we construct a weighted relationship graph from historic stock market data.

Data. In this work we use historic stock market data collected between 2010 and 2016 for stocks listed in the S&P 500 index[7]. Each stock's daily open, high, low, close and volume values (OHLCV) is given in US dollar, all values are already accounted for stock splits and adjusted to the last price. Thus, price values of affected stocks in 2010 are rectified to have the same meaning as in 2016. For linking stock prices with occurrences in the previously introduced news dataset, we will be using the company names in the securities table from the same published dataset. We discard data after 2013-11-29, so that relationships we extract are based on data from the same time period as our news corpus. However, the financial news before 2010 will be used, even though no stock prices are collected for this time period, since relational features are assumed to have a long-term impact, so incorporating information from news between 2006 and 2010 is helpful. To account for acquisitions, mergers, dual-class listings, or bankruptcies of components in the index, we only consider stocks that are part of the index by the end of 2016. In addition to the stock prices, overall measurements of the performance and the confidence at the NYSE for the same period as the stock prices need to be considered. Therefore, we use the CBOE Volatility Index (VIX)[8] and the S&P 500 index (GSPC)[9]. There appears to be

[7] https://www.kaggle.com/dgawlik/nyse, https://nemozny.github.io/datasets/
[8] https://www.kaggle.com/lp187q/vix-index-until-jan-202018
[9] https://www.kaggle.com/benjibb/sp500-since-1950

a strong negative relationship between volatility and stock market returns [10]. Most notable bursts of the VIX happened in 2010 and 2011 and are hypothesized to be caused by important steps during the European debt crisis [10]. Another burst is assumed to be a consequence of the *1000-point plunge* of the DOW Jones index on the 24th of August in 2015, which in turn was a consequence of a rout in the Chinese market pulling down stock markets all over the world [11].

The preselection ensures the complete stock prices of 467 companies for all 985 trading days from 2010-01-04 to 2013-11-29.

Normalising Stock Movements. The previously described data has to be transformed to fulfil preconditions for the correlation analysis, namely the time series consist of independently and identically distributed samples [11]. Therefore, we combine different methods to remove shared external influences and existing autocorrelation, ensure homoscedasticity and apply autoregressive models to remove any remaining irrelevant patterns.

Stationarity. An assumption of the methods we employ to detect relationships between the financial time series is, that the data is stationary. Stock prices however follow the ever growing market and are thus non-stationary and are assumed to contain a unit root [28]. The unit root, namely the influence of the market, prevents the series to return to stationary mean and can be accounted for by differencing the time series taking the absolute or relative differences between each sample [8]. In the following we will use relative differences to account for different levels of stocks. Hong et al. [15] provide empirical findings about recurring patterns in returns including that open-to-open returns are more volatile than close-to-close returns, while Wang et al. [32] provide evidence that intra-day (open-to-close) and overnight (close-to-open) returns have significantly different properties concluding that one shall not mix them. Lastly, Li et al. [23] consider daily open-to-close prices arguing that it is less prone to seasonality and the more volatile non-trading gaps across weekends and holidays. Results for our data coincide with related work and show a moderate negative skew (-0.1 in average) for open-to-open/close-to-close returns and a slightly positive skew (0.06 in average) for the distribution of intra-day returns. Because of their more normal like distributions, relative intra-day returns are used for the remainder of this work.

Homoscedasticity. An important assumption for correlation is, that the data is homoscedastic, that is homogeneity of volatility in the context of time series. Stock prices, however, are heteroscedastic [26], most likely because exogenous factors are left out. Some statistical methods can be employed to normalise the data, for example with the GARCH model [25] or other robust regression methods such as weighted least squares regression. In this work, we use the

[10] https://money.cnn.com/2011/08/08/markets/vix_fear_index/index.htm

[11] https://money.cnn.com/2015/08/24/investing/stocks-markets-selloff-china-crash-dow/index.html

Box-Cox transformation [3] for heuristic data stabilization by a simple power transformation. The model's power parameter is determined by maximising the log-likelihood function on the previously modified relative intra-day return. Since this general transformation is not a panacea to the problem, we also apply individually fitted autoregressive models on top of that.

Exogenous Variables. Financial markets are prone to a number of external influences which are usually not accounted for by the previously described regression models on stock prices. If the economy is doing well or experiences a period of uncertainty, this will be likely reflected in all stock prices. Hence, omitting exogenous variable is another reason for spurious correlations [12]. A greater correlation between stock prices can be caused by a shared external factor which is the common market performance in this case. To have a better representation for the intrinsic performance of a single stock, its returns need to be normalised regarding the shared performance of the market. Further, the underlying movement of a stock price might even more be biased by the sentiment of the according industry section. Stocks from the same industry will therefore have a high cross-correlation without revealing specific relationships. We counteract exogenous influences by subtracting the average return value of a stock's industry sector. Alternatively, we can normalise by the S&P 500 market index instead of the separate industry averages. The impact of this alternate normalisation step will be examined later when setting up the graph based on the extracted cross-correlations.

Autoregression. Even after all previous normalisation steps, some autoregressive patterns may remain. Autoregressive models describe time series by a function of their past values along with an error term, known as the residual. The residuals are assumed to be free from linear autocorrelation [7] and therefore, if applicable, can be used to identify relationships between stocks. We use an autoregressive moving average model (ARMA) [29], that assumes lagged values and error terms for the same stationary and univariate time series. The model hyper-parameters are determined with the Box-Jenkins method [4]. We applied the ARMA model to 82 stocks for which autocorrelation was observable with this method. Even if the model wasn't applied to a time series, we refer to the unchanged data as the residuals in the following. On top, we apply a Generalised Autoregressive Conditional Heteroscedasticity Model (GARCH) on the residuals of the ARMA mean process to remove any remaining unstable volatility in the normalised stock prices. Because the actual volatility can not be observed directly and the data is ensured to be stationary, we use the squared residuals as a proxy. As the last step of our data normalisation we divide the ARMA residuals by the conditional volatility calculated with the GARCH model. For the 161 stocks where the GARCH model isn't applicable, we divide the residuals by the overall standard deviation to keep the resulting time series at the same scale.

Building a Relationship Graph. The data normalisation process is very involved to remove any influences on a stock price development that are not in-

herent to its intrinsic value. Obviously, some autocorrelation is still contained in the data since all economic decision making is somewhat linked. We conducted a series of tests to ensure our normalisation ensures preconditions for correlating stocks best as possible. Namely, we tested stationarity, data-distribution, homoscedasticity, structural breaks, autocorrelation, seasonality, and outliers. In our evaluation we saw only a few stocks failing some of the tests.

With all necessary statistical preconditions established, we calculate the pairwise cross-correlation between all normalised stock time series. For that, we use Pearson's r bivariate sample correlation coefficient. The average correlation between the 108,811 pairs of unprocessed stocks is very high with $r = 0.96$, but after pre-processing drops to a zero-centred normal distribution with $\sigma = 0.12$. Previously, we mentioned that the data can be normalised either with the average performance of stocks in one industry or the entire S&P 500 index. In our correlation analysis we only saw positive correlations with a median of $\tilde{r} = 0.2$ for market-wise normalisation. Because these correlations are not zero-centred, they are assumed to still have significant shared exogenous influences which pollute the individual correlation values. Concluding, the industry-wise normalisation, which ensures a zero-centred distribution, is preferable for removing exogenous influences.

We use the correlation factor as edge weights in the relationship graph. Different to the graph based on news, there are edges for all pairwise stocks.

5 Comparing Relationship Graphs from Stocks and News

In previous sections we described how we generate company relationship graphs from news and stocks with weighted edges describing the strength of relationship. With over 100k edges in the graph based on stocks and over 15k edges in the graph based on news, it is impossible to produce a meaningful visualisation. Therefore, we select only edges outside the respective 99.9th percentile in Fig. 2 on the facing page. For the news graph this leaves 90 edges between 98 nodes from all eleven industry sectors. For the stock graph this means we apply a threshold of $r = 0.368$ on the absolute correlation-based edge weights and are left with 109 edges among 123 different stocks from all industry sectors. A node's size is determined respectively to its total revenue in 2010 in order to indicate its importance, the colour coincides with the industry. The visualised graphs only consist of the most extreme values and might therefore not be representative for the entire graphs. However, by being the most extreme samples, this also means, that only the most certain relationships based on either type of data is shown.

Qualitative (Visual) Comparison. Comparing both graphs (visually), we see very different structures. Both graphs share 29 nodes and four edges only. Instead of many small decoupled sub-graphs for industry sectors the news graph reveals one big sub-graph consisting of 50 nodes from many different industry sectors. Some edges between companies sound comprehensible after investigating their business relationships. For example, *Microsoft Corp.* (*MSFT*) and *Vi-*

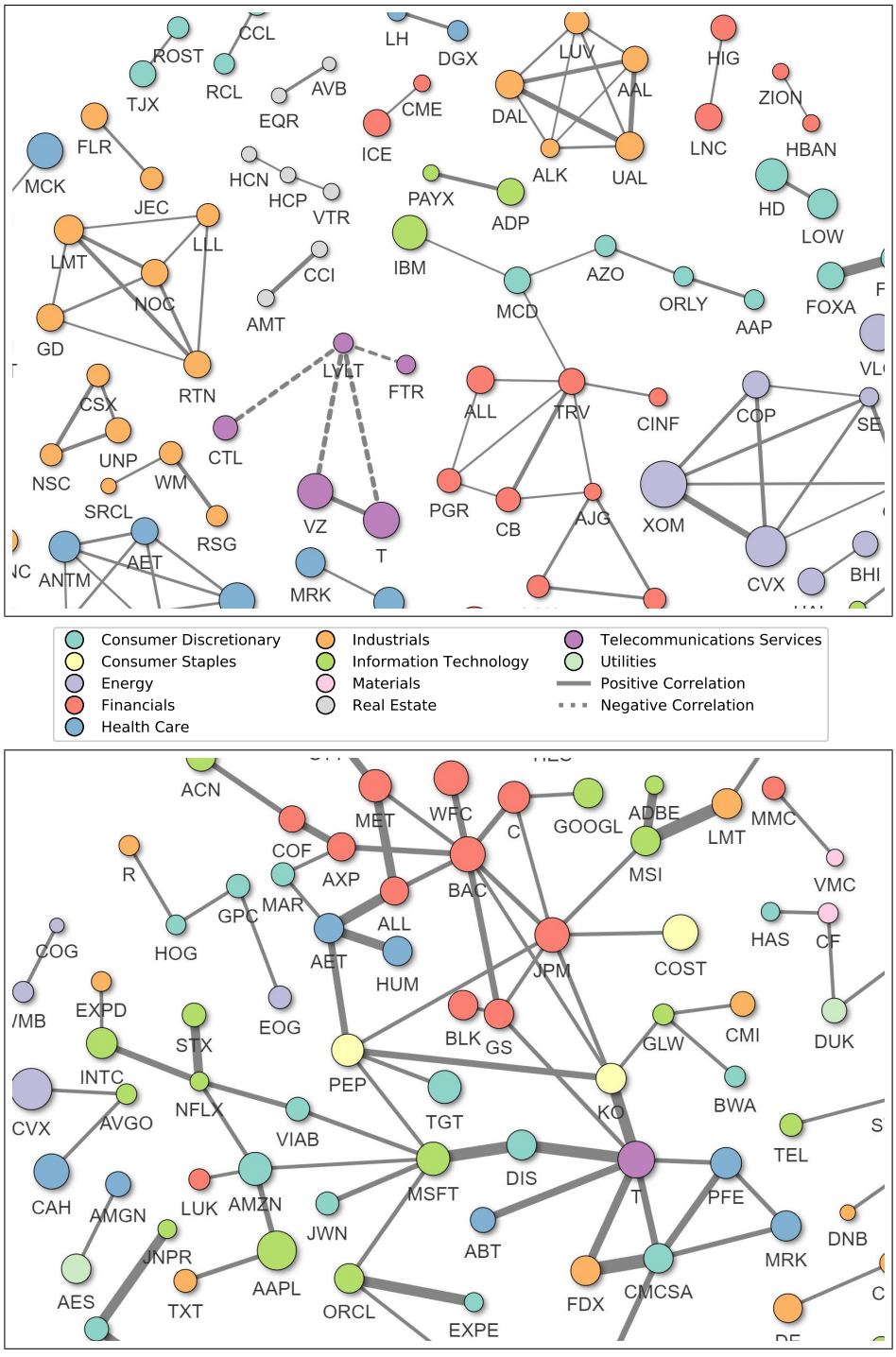

Fig. 2. Excerpts of relationship graphs based on stock similarity (top) and news articles (bottom). Only edges with weights above the 99.9th percentile are drawn.

acom Inc. announced a long-term strategic alliance for corporate segments like game development and advertisement in 2007. To mention another example, a high pairwise distance can be observed for both streaming providers *Netflix Inc.* (*NFLX*) and *Amazon.com Inc.* (*AMZN*) which both benefit from the increased demand in this segment of the market. Over the whole graph, only 28 edges are connections among companies originating from the same industry. The greatest industry cluster can be observed for companies from the sector *Financials* which includes insurance companies (e.g. *American International Group, Inc.*), investment banks (e.g. *JPMorgan Chase & Co.*) and financial service providers (e.g. *Citigroup Inc.*). Some of them are densely connected with other industries which can be argued by their investments in stocks of other companies.

In the stock graph, a large proportion of high correlations are observed among companies belonging to the same industry sectors. Only eight edges between two different sectors are present in this visualisation. In terms of inter-industry connections, the node *MCD* (*McDonald's Corp.*) in the center of the graph is the strongest one since it is connected to nodes from three different industries. Investigation by financial news did not reveal an underlying relationship with the connected companies. Instead, this stock appears to be an appropriate strong representative component of the market performance and therefore is strongly linked to other important representative components like *IBM*. From the 109 edges selected for the graph, only four represent a negative correlation which all originate from the industry sector *Telecommunications Services*. Further, it should be noted that there a three companies for which each one compromises two stocks. Because two stocks of the same company are almost equal, these three correlation pairs reveal the highest r.

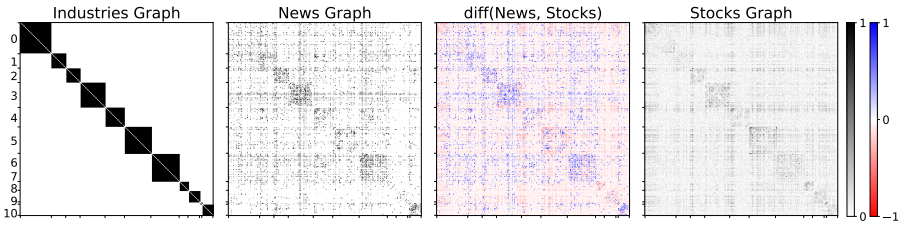

Fig. 3. Adjacency matrices for the industries graph, stocks graph, news graph and their difference. The industries graph contains only edges between companies of the same industry. The industry sectors are as follows (top to bottom): Consumer Discretionary; Consumer Staples; Energy; Financials; Health Care; Industrials; Information Technology; Materials; Real Estate; Telecommunications Services; Utilities

Quantitative Comparison. As denoted previously, the visualisations of the graphs are not entirely representable. The adjacency matrices in Fig. 3 reveal

Table 1. Pearson's r for comparing graphs based on co-occurrence counts, minimum and pairwise distances in news with normalised and raw stock correlations

		News		
		Count	Min-Dist	Pairwise-Dist
Stocks	Normalised	0.0945	0.1124	**0.1284**
	Raw	-	-	-0.0037

some overall patterns. The edges between companies from the same industry sectors are stronger for both graphs. For the stocks graph, the companies in the sector *Health Care* have a low correlation to companies from any other sectors. This pattern is still persistent after taking the difference of both graphs. Further, there are strong disagreements observable between both graphs regarding the sectors *Industrials*, *Information Technology* and *Utilities*. As already seen in the qualitative evaluation, the companies in the sector *Financials* usually have high edge weights with other companies in the news graph.

In order to measure the compatibility of both graphs and their different variations, we conduct a comparison of the stock correlations and the extracted business relationships. The number of edges in the unfiltered graphs are incompatible for comparison. Thus, we use only companies and relations that appear in both graphs resulting in 417 companies and 86,736 unique bidirectional edges. We use the absolute correlation as weights in the stock graph and adjust the scale of edge weights. Table 1 shows the correlation of graphs weighted by different metrics. The graph of relationships extracted from news weighted by pairwise distances of company mentions in the texts has the highest similarity to the correlations between normalised stocks. The correlation graph of raw stock prices is expected to contain almost exclusively spurious correlations, but is included for comparison purposes. In the graph from raw stocks, almost all companies are highly correlated, hence the low similarity to all other graphs. While business news are dominated by reports about new alliances or financing deals, stocks reflect the actual effect that this has on the market and also investors reactions that go beyond what gets featured in news. This can clearly be seen in our visual comparison of most prevalent company relationships and also in our quantitative analysis.

6 Conclusions

In this work we have demonstrated two methods to create a graph of company relationships. We extracted company mentions from business news and proposed three approaches to add edge weights as an indicator of how strong a particular relation is. In our second approach, we extracted the company relationships from historic stock market data, for which we proposed extensive pre-processing steps to ensure that autoregressive and external influences do not invalidate the results. Based on four years historical stock prices and seven years financial news,

we found evidence supporting the hypothesis that both graphs show similarities. However, we had to introduce limitations and assumptions, as business relationship and intrinsic value are not directly observable. Through the methods presented in this paper, we introduced proxies for these information in the form of weighted graphs. We examined how well a stock price can be described by stock prices of related companies to understand to what extent stock movements are determined by business relationships. As there is no complete collection of business relationships, we used co-occurrences of company mentions in news articles. In our evaluation, the edge weights based on pairwise distances are most similar to stock correlations.

We see a lot of potential use cases for company relationship graphs in downstream tasks, for example as additional information in entity embedding models, extending knowledge bases, or as a supporting feature in market analysis.

References

1. Akita, R., Yoshihara, A., Matsubara, T., Uehara, K.: Deep learning for stock prediction using numerical and textual information. International Conference on Computer and Information Science pp. 1–6 (2016)
2. Bachelier, L.: Theory of Speculation. Annales scientifiques de l'École normale supérieure (1900)
3. Box, G.E.P., Cox, D.R.: An Analysis of Transformations. Journal of the Royal Statistical Society: Series B (Methodological) (1964)
4. Box, G.E.P., Jenkins, G.M.: Time Series Analysis Forecasting And Control. Journal of Time Series Analysis (4 1970)
5. Chen, Y., Wei, Z., Huang, X.: Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction. In: Proceedings of the ACM International Conference on Information and Knowledge Management. pp. 1655–1658. ACM (2018)
6. Ding, X., Zhang, Y., Liu, T., Duan, J.: Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. p. 14151425 (2014)
7. Dionisio, A., Menezes, R., Mendes, D.A.: Mutual information: A measure of dependency for nonlinear time series. In: Physica A: Statistical Mechanics and its Applications (2004)
8. Engle, R.F., Granger, C.W.J.: Co-Integration and Error Correction: Representation, Estimation, and Testing. Econometrica (1987)
9. Fama, E.F.: Random walks in stock market prices. Financial analysts journal pp. 55–59 (1965)
10. Fleming, J., Ostdiek, B., Whaley, R.E.: Predicting stock market volatility: A new measure. Journal of Futures Markets (1995)
11. Franke, J., Härdle, W.K., Hafner, C.M.: Statistics of Financial Markets (2010)
12. Granger, C.W.J.: Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica (1969)
13. Granger, C.W.J., Newbold, P.: Spurious regressions in econometrics. Journal of Econometrics (1974)
14. Hallin, M.: Gauss-Markov Theorem in Statistics. In: Wiley StatsRef: Statistics Reference Online. Springer (2014)

15. Hong, H., Wang, J., Heaton, J., Holden, C., Lo, A., Slezak, S., Stein, J., (the, R.S.: Trading and Returns Under Periodic Market Closures (1998)
16. Hsu, M.W., Lessmann, S., Sung, M.C., Ma, T., Johnson, J.E.: Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert Systems with Applications (2016)
17. Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., Ngo, D.C.L.: Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications (2015)
18. Khadjeh Nassirtoussi, A., Aghabozorgi, S., Wah, T., Ngo, D.: Text mining for market prediction: A systematic review. Expert Systems with Applications pp. 7653–7670 (2014)
19. Kim, S.: A Cross-Correlation-Based Stock Forecasting Model. In: Proceedings of The National Conference On Undergraduate Research (2016)
20. Kosapattarapim, C.: Granger causality between stock prices and currency exchange rates in Thailand. In: AIP Conference Proceedings. vol. 1905, p. 50025 (3 2017)
21. Lee, H., Surdeanu, M., MacCartney, B., Jurafsky, D.: On the Importance of Text Analysis for Stock Price Prediction. In: Proceedings of the Language Resources and Evaluation Conference. pp. 1170–1175 (2014)
22. Li, B., Chan, K.C., Ou, C., Ruifeng, S.: Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Information Systems (2017)
23. Li, X., Xie, H., Chen, L., Wang, J., Deng, X.: News impact on stock price return via sentiment analysis. Knowledge-Based Systems pp. 14 – 23 (2014)
24. Lipenkova, J.: A system for fine-grained aspect-based sentiment analysis of chinese. In: ACL-IJCNLP. pp. 55–60. ACL (2015)
25. Millo, G.: Robust Standard Error Estimators for Panel Models: A Unifying Approach. Journal of Statistical Software (2017)
26. Morgan, I.G.: Stock Prices and Heteroscedasticity. The Journal of Business (1976)
27. Peng, Y., Jiang, H.: Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016)
28. Lopez de Prado, M.: Advances in Financial Machine Learning. Wiley (2018)
29. Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A.: Correlating Financial Time Series with Micro-blogging Activity. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. ACM (2012)
30. Sun, A., Lachanski, M., Fabozzi, F.J.: Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. International Review of Financial Analysis (2016)
31. Vlastakis, N., Markellos, R.N.: Information demand and stock market volatility. Journal of Banking and Finance (2012)
32. Wang, F., Shieh, S.J., Havlin, S., Stanley, H.E.: Statistical analysis of the overnight and daytime return. Physical Review E (2009)
33. Yule, G.U.: Why do we Sometimes get Nonsense-Correlations between Time-Series?–A Study in Sampling and the Nature of Time-Series. Journal of the Royal Statistical Society (1926)
34. Zhai, Y., Hsu, A., Halgamuge, S.K.: Combining News and Technical Indicators in Daily Stock Price Trends Prediction. In: Advances in Neural Networks ISNN 2007. pp. 1087–1096. Springer (2007)