Credit and Capital Markets, Volume 52, Issue 2, pp. 229–251
Scientific Papers

Networks and News in Credit Risk Management

Ferdinand Graf and Martin Dittgen*

Abstract

The presumably most important function of a corporation is the establishment and management of connections to customers, suppliers, investors, debtors and competitors. All these connections may produce profits or bear risks. Hence, the isolated inspection of a corporation (or also a sovereign) may be insufficient. Instead, the economic environment of a corporation and its connections should be included in its valuation. Usually, this is done via manual and hardly standardized processes with their associated large efforts. This article presents a new method to analyze business news and to build up a network of corporations based on business news. To this end, we search in news articles from Reuters and Bloomberg for corporation names or synonyms and assume a connection exists between two corporations if the corporations are mentioned together frequently. Based on these connections, we (1) build up a network for the S&P500 companies, (2) identify groups therein to validate the approach manually and (3) test, whether corporations with many connections and a particularly favorable position in the network receive better rating grades compared to corporations with fewer connections and an average network position. The latter is equivalent to the question of whether a corporation's connections are a driver of the firm value. Moreover, we use the business news to measure a corporation's publicity and sentiment, and relate these to the corporation's rating as well. Our empirical results indicate that the network properties, the sentiment and the media attention are contained in respectively affect the rating grade. Hence, the incorporation of news in the firm valuation – as it is done by many financial institutions – is reasonable. The factors mentioned above increase the explanatory power of our regression model significantly. Since many corporations have sufficient news coverage for our approach but are not rated from a rating agency, and hence must be rated with internal models, our approach may support manual processes in financial institutions and reduce efforts and costs.

* Dr. Ferdinand Graf, d-fine GmbH, An der Hauptwache 7, 60313 Frankfurt, Ferdinand.Graf@d-fine.de, Martin Dittgen, d-fine GmbH, An der Hauptwache 7, 60313 Frankfurt, Martin.Dittgen@d-fine.de.

We would like to thank our colleagues and the anonymous referee for their valuable input.

Netzwerke und Unternehmensnachrichten in der Kreditrisikomessung

Zusammenfassung

Eine der zentralen aber oft unterschätzten Aufgaben von Unternehmen ist der Aufbau und die Pflege von Beziehungen zu Kunden, Lieferanten, Gläubigern, Investoren oder auch Konkurrenten, aus denen Profite und gegebenenfalls auch Risiken resultieren. Daher ist die isolierte Betrachtung eines Unternehmens (oder auch eines Staates) für dessen Bewertung oft nicht ausreichend. Stattdessen sollten das wirtschaftliche Umfeld eines Unternehmens und die Verbindungen eines Unternehmens direkt in dessen Bewertung einfließen. Deren zumeist qualitative, wenig standardisierte Analyse verursacht bei Kreditinstituten meist hohe Aufwände. Dieser Artikel beschreibt die Analyse von Unternehmensnachrichten und die Herleitung von Netzwerken dieser Unternehmen aus deren Unternehmensnachrichten. Hierzu suchen wir in Nachrichten von Reuters und Bloomberg nach Unternehmensnennungen und gehen von einer Verbindung zwischen zwei Unternehmen aus, wenn diese häufig in denselben Nachrichten genannt werden. Aufgrund dieser Verbindungen (1) erzeugen wir ein Netzwerk für die Unternehmen im S&P500, (2) identifizieren nicht-triviale Unternehmensgruppen und (3) testen, ob gut vernetzte Unternehmen eine bessere Bonitätsnote von den Ratingagenturen erhalten als weniger gut vernetzte Unternehmen. Letzteres ist gleichbedeutend mit der Fragestellung, ob eine gute, zentrale Positionierung eines Unternehmens in einem Netzwerk einen messbaren Mehrwert für das Unternehmen schafft, der sich im Rating niederschlägt. Darüber hinaus nutzen wir die Unternehmensnachrichten auch dazu um Kennzahlen abzuleiten, die die Aufmerksamkeit und die Stimmung der Nachrichtenlage unternehmensspezifisch messen und somit das wirtschaftliche Umfeld eines Unternehmens quantifizieren. Bezüglich dieser Kennzahlen überprüfen wir ebenfalls, ob sie einen messbaren Einfluss auf die Ratingnoten haben. Unsere Ergebnisse legen nah, dass sich sowohl Netzwerkeigenschaften als auch die Nachrichtenlage in der Bonitätseinschätzung niederschlagen. Diese Kennzahlen steigern den Erklärungsgrad unseres Shadow-Rating Modells erheblich. Da viele Unternehmen eine für unseren Ansatz hinreichende Nachrichtenabdeckung besitzen, aber kein Agenturrating, kann unser Ansatz besonders bei der Bewertung von Adressrisiken mit internen Modellen manuelle Prozesse ablösen und zu Effizienzsteigerungen führen.

Keywords: Business News, Network Analysis, Sentiment Analysis, Shadow-Rating Model

JEL Classification: G14, L14, D85

I. Introduction

Even though the first text documents originate from the Ancient Egypt, the data object ‘text’ is often excluded from quantitative empirical analysis in finance. This may be due to the missing or complex structures in text, which is therefore often cited as an example for unstructured data. There are almost numberless ways to report on one event via text. Moreover negations, humor, sarcasm and the ambiguity of words and sentences are huge challenges in the

automated evaluation of text. In addition, text passages may refer to other text passages and must be interpreted jointly, which may be hard for algorithms. All this makes the evaluation of text challenging.

Making things even worse, the number of sources for text increases continuously, and so does the amount of published text every day. Hence, the data that must be analyzed in a representative dataset is huge and the analysis is computationally expensive. *Das/Chen* (2007), *Tetlock* (2007) and *Loughran/McDonald* (2011) are presumably the most important studies in the financial literature overcoming all of these obstacles and document that news have a traceable impact on modern financial markets.

The connections between financial institutions – respectively sovereigns – and resulting contagion effects are often discussed in the literature, see *Eisenberg/Noe* (2001), *Elliott* et al. (2014), *Acemoglu* et al. (2015) and *Fagiolo* et al. (2007). Similar studies for non-financial corporations are rare with the exceptions for *Cossin/Schellhorn* (2007) and *Pozzi/Di Matteo/Aste* (2013). One reason for this may be that connections between corporations are usually associated with claims and liabilities between corporations, which are collected and consolidated by the regulatory authority only for financial institutions but not for non-financial corporations.

Credit rating agencies use various data sources and directly or indirectly pay attention to the information contained in business news. However, their precise approach how this information affects the rating grades is secrete, even though several studies analyzed it, see e. g. *Altman* (1968), *Kamstra* et al. (2001), *Bhojraj/Sengupta* (2003) and *Mählmann* (2011). We analyze a comprehensive set of business news and use it to derive a network for a part of the economy. Insights resulting from the news and the network are then integrated into regression models explaining the rating grade. This approach may further be used to derive rating grades for corporations without agency rating, which is known as shadow-rating, see *Ratha* et al. (2010). We think that both objects, i. e. 'text' and 'networks', will gain importance for markets and corporations due to the digitalization, mobile devices and new communications standards.

II. Data

We analyze business news from the online archives of Reuters and Bloomberg within the time period 01.01.2007 to 31.12.2014. Our sample consists of 6.430.709 news articles. The Reuters archive is available under the URL http://www.reuters.com/resources/archive/us/yyyymmdd.html[1] and contributes 5.847.242 articles. The Bloomberg archive has the URL http://www.bloomberg.

com/archive/news/yyyy-mm-dd/[1, 2] and contributes 583.467 articles for the sub period 01.01.2010 to 31.12.2014. The articles are sufficiently uniformly distributed over time. The Reuters archive contains on average 2.026 articles per day and the Bloomberg archive 331 articles per day. On an average workday we observe in total 2.951 articles and on an average weekend-day 336 articles. We consider the two news providers as representative for the financial news universe, even though there are other important news providers like Financial Times or Wall Street Journal.

All corporations which were part of the S&P 500 Index or the Dow Jones Industrial Average for at least one day in our observations period are considered in our analysis. This yields a total of 666 corporations. For these corporations, we analyze Long-Term Foreign Currency Ratings from the rating agencies Moody's, Standard&Poors and Fitch. Moreover, we use financial ratios and stock returns as control variables in the regression analysis below, downloaded from http://financials.morningstar.com/ and https://finance.yahoo.com/, respectively.

III. Sentiment

We measure the sentiment of a news article with respect to a corporation with four individual indicators, namely raw-sentiment, commitment, information and relevance. For all four measures we assume that each word of a news article has an individual and corporation specific weight, which consists of two components.

The first component controls for the position of a word within the article for the following two reasons:

(1) Leakage of content: Important things are usually mentioned first, followed by supplemental information and discussions.

(2) Leakage of attention: Investors usually start reading an article from the beginning. However, some of them might stop reading before finishing the article.

Therefore, the first word in an article receives the highest weight of 1.0. The weight is continuously reduced for the following words. The last word receives a weight that is inversely proportional to $\log_{10}(|N|)$, where $|N|$ denotes the total number of words in the news article N. With this feature we control for the effect that the end of shorter news articles are more frequently read and may

[1] In both URLs the string yyyy is a placeholder for the year, mm for the month, and dd for the day the archive is requested.

[2] The Bloomberg homepage was restructured in the beginning of 2015 so that the archive is no longer available.

hence be more important than the end of rather long articles. The weight is given by:

$$W_1(N,p)=\left(1-\frac{p-1}{|N|}\right)^{x(N)} \text{ with } x(N)=\frac{\ln(\log_{10}(|N|)}{\ln(|N|)},$$

where p denotes the position of the word to which the formula is applied, i. e. 42 for the 42th word in the news article. Full-text search engines such as Elasticsearch make use of such weights to score and rank documents w.r.t. search requests, too.

The second component controls for the position of a word relative to the position of the company name, and models the effect that the same passage may have different importance for two corporations. This weight is calculated according to:

$$W_2(N,p,C)=\exp\left\{-\frac{1}{2}\left(\frac{\min\{k \mid N(p\pm k)\in H(C)\}}{12}\right)^2\right\},$$

where $N(q)$ denotes the qth word of news article N, C denotes the corporation and $H(C)$ the set of corresponding corporation names or synonyms.[3] The function is calibrated so that it assigns a weight of at least 0.7 to words within a 10-word distance to the corporation's name and a weight of not more than 0.1 to words outside of a 25-word radius. For a stylized example how the news are processed, see appendix I.

For each word and each corporation, both weights are multiplied, i. e. $W(N,p,C)=W_1(N,p)\cdot W_2(N,p,C)$. Afterwards, the words are compared to the word-lists positive, negative, strong and weak of Loughran and McDonald's 'Financial Sentiment Dictionary', see *Loughran/McDonald* (2011), and the corresponding word weights are summed up for each list, i. e.

$$\mathrm{WWC}(N,C,\mathrm{CAT})=\sum_{p\le|N|}W(N,p,C)\cdot 1[N(p)\in \mathrm{CAT}],$$

where $\mathrm{CAT}\in\{\mathrm{Pos},\mathrm{Neg},\mathrm{Str},\mathrm{Wea}\}$ denotes the set of all words on the corresponding word-list, and the function $1[\cdot]$ is one if and only if its argument is true and zero otherwise. The check $N(p)\in\mathrm{CAT}$ includes the Porter stemming algorithm, see *Porter* (1980), in order to make the analysis robust against various

[3] Note that a company may have more than one identifier and that the identifier may consist of more than one word, e. g. 'American Express' and 'AmEx' are identifiers for the corporation 'American Express Co'. Whenever a company name consists of a phrase, the words are concatenated by '_' and interpreted as a single word, e. g. 'American Express' is adjusted to 'American_Express'.

linguistic variations like singular/plural form, verb-tense etc. We also considered the corresponding word-list from the 'General Inquirer Dictionary'; however, results are not shown in the empirical section below.

The aggregated, weighted word counts $\mathrm{WWC}(\cdot)$ might hardly be comparable across the four word categories under consideration. Words in the category weak are expected to be rare since business news report predominantly on facts rather than speculating. In addition, the audience might appreciate a latent positive, strong tendency in news, which could be anticipated by the reporters and editors. Moreover, the representativeness of the documents used to build the word lists might imply a bias: For example the 'General Inquirer' was build using rather general English text, and the 'Financial Sentiment Dictionary' was derived from 10K reports. In order to account for a potential misbalance in the distribution of words across categories on our dataset the weighted word counts are adjusted. We use a linear regression model to explain the sum of weighted word counts in each category by the sum of all word weights in a news article, thus removing the bias via the regression formula:

$$\mathrm{WWC}(N, C, \mathrm{CAT}) = \alpha_{\mathrm{CAT}} + \beta_{\mathrm{CAT}} \cdot \#(N, C) + \varepsilon,$$

where $\#(N, C)$ denotes the sum of all word weights, i. e. $\#(N, C) = \sum_{p \leq |\mathbb{N}|} W(N, p, C)$. Finally, the estimated regression coefficients[4] $\widehat{\alpha_{\mathrm{CAT}}}$ and $\widehat{\beta_{\mathrm{CAT}}}$ are used to benchmark the sum of weighted word counts, i. e.

$$\mathrm{Adj\_WWC}(N, C, \mathrm{CAT}) = \mathrm{WWC}(N, C, \mathrm{CAT}) - \widehat{\alpha_{\mathrm{CAT}}} - \widehat{\beta_{\mathrm{CAT}}} \cdot \#(N, C).$$

The benchmarked sums are then used to define the corporation specific measures raw-sentiment $\mathrm{RAW}(\cdot)$ and commitment $\mathrm{COM}(\cdot)$ for news articles:

$$\mathrm{RAW}(N, C) = \frac{\mathrm{Adj\_WWC}(N, C, \mathrm{Pos}) - \mathrm{Adj\_WWC}(N, C, \mathrm{Neg})}{|\mathrm{Adj\_WWC}(N, C, \mathrm{Pos})| + |\mathrm{Adj\_WWC}(N, C, \mathrm{Neg})|},$$

$$\mathrm{COM}(N, C) = \frac{1}{2}\left(1 + \frac{\mathrm{Adj\_WWC}(N, C, \mathrm{Str}) - \mathrm{Adj\_WWC}(N, C, \mathrm{Wea})}{|\mathrm{Adj\_WWC}(N, C, \mathrm{Str})| + \mathrm{Adj\_WWC}(N, C, \mathrm{Wea})}\right).$$

By construction, the raw-sentiment has the codomain [–1; 1], where the value 1 marks that more positive and fewer negative words than expected appear in

[4] The coefficients are estimated out-of-sample on news articles published either in 2016 or before our observation period. The estimated regression coefficients indicate a slightly positive and clearly strong, fact-orientated tendency in the news articles, which is consistent with our expectations. For the sake of brevity, the regression results are not shown in the empirical section.

the neighborhood of the corporation name. The value –1 marks that fewer positive words and more negative words than expected are used. Values between –1 and 1 indicate a mixture of both effects, i. e. more negative and more positive words than expected or fewer negative and fewer positive words. Then, the sign shows which effect dominates. The codomain of the commitment is [0; 1]. If a news article makes excessive use of subjunctive forms or other weak words, the commitment is close to zero. If many definitive words are used, it is close to one.

The third measure is the relevance of a news article for a corporation, which is defined as:

$$\mathrm{REL}(N, C) = \frac{\#(N, C)}{\sum_p W_1(N, p)}.$$

It also has the codomain [0; 1], where high values indicate a high relevance. Moreover, we request a news article to mention a corporation at least twice and to mention not more than 15 distinct corporations to receive a non-zero relevance measure. By this construction, we rule out news articles without a minimum focus.

Our fourth and last measure is the information. This measure compares a news article with all news articles on the considered corporation that were published up to 120 minutes earlier. For this task, we express the news article as a vector, denoted by $\vec{\cdot}$. The dimensions of the vector represent word stems according to the Porter stemmer, and the vector's magnitude in a dimension corresponds to the sum of the benchmarked word weights given a corporation.[5] The similarity between two news articles w.r.t. a company is now measured by the angle between the news articles expressed as vectors. Our measure of information for a news article and a corporation is defined by the most similar news article, i. e.:

$$\mathrm{INF}(N, C) = \min_{M \in S} \left\{1 - \cos\left(\vec{N}(C), \vec{M}(C)\right)\right\},$$

where S denotes the set of news on C that where published up to 120 minutes prior than news article N. This measure takes on values between zero and one, where values close to one indicate that relevant sections in the news articles w.r.t. a company are almost orthogonal to all previously released articles, and values close to zero that at least one very similar news article was published before.

[5] Since our word-weights are sensitive to the position of a word within the news article, this approach extends the bag-of-word approach, which may be criticized for ignoring the structure of the text completely.

We define the sentiment as the product of all four measures, i. e.

$$\mathrm{SEN}(N,C) = \mathrm{RAW}(N,C) \cdot \mathrm{COM}(N,C) \cdot \mathrm{REL}(N,C) \cdot \mathrm{INF}(N,C).$$

The sign of the sentiment is governed by the sign of the raw-sentiment and its codomain is [–1; 1]. A strong sentiment requires that all four measures must be clearly different from zero. Each single measure may set the sentiment to zero.

IV. Networks

It is highly unlikely that all corporations are of equal importance to the economy. Much more reasonable is that each corporation has an individual importance. Therefore, we obtain a network that abstracts the economy and – based on its structure – assigns specific weights to each network-node approximating the importance of the corporation to the economy. For computational reasons, we assume a static network and incorporate all news from our database. In addition, we consider connections between corporations as rather static and not as dynamically as e. g. a corporation's sentiment.

In a given news article we consider firstly how often a corporation is mentioned and secondly which other corporations are mentioned with it.[6] Consistent with the definition of our relevance measure and in order to remove false positives, we require that the name of the corporation under consideration has to appear at least twice in a news article. If a news article contains more than 15 different corporation names, we regard the news as potentially misleading and discard it.

To quantify the connection strength between two corporations we divide the number of news where both corporations are mentioned by the number of news containing at least one of the corporations. Therefore, the connection strength represents how often the corporations are mentioned together in the news expressed as a percentage.

Based on the connection strength we then construct two different networks:

(1) Discrete network: For this network, we consider non-weighted, binary edges. If two corporations are mentioned together in at least 5 news articles and the connection strength defined above is at least 5 %, we establish an edge. Otherwise, we consider no edge between the considered companies.

(2) Continuous network: For this network, we consider weighted edges. The edge-weight is identified with the connection strength implied by the news, thereby obtaining a continuous spectrum of different values.

[6] Even though connections may be a result of conflicting economic interest and harmonized economic interest, we aim at the impact of the dominant effect, which is expected to be positive (in line with 'competition is good for business').

Given one of the above networks, we proceed to measure the importance of each corporation in contrast to the rest. We consider the following metrics:

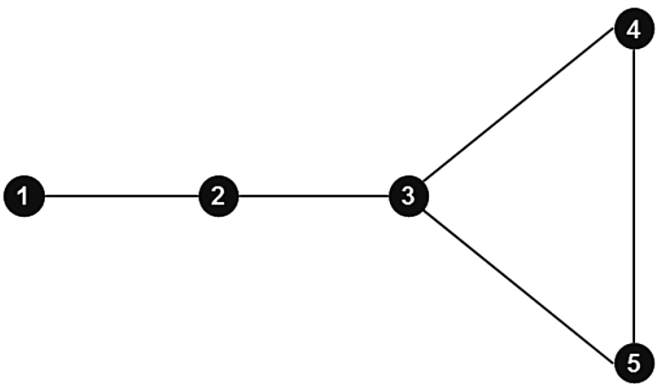
(1) Degree centrality: It measures how many direct connections a company has to other corporations in relation to all possible connections. For the continuous network it is obtained by taking the sum of the weights for each direct connection.

(2) Closeness centrality: It measures the inverse average geodesic distance between a corporation and all other corporations.

(3) Betweenness centrality: It measures how many corporation pairs that are not connected directly have a shortest geodesic path that includes the corporation under consideration standardized by all eligible pairs.

(4) Page-Rank centrality: It measures the importance of each corporation in relation to the importance of the corporations with which it is directly connected. This implies a recursive structure where the page-rank centrality is the value in stable equilibrium, after sufficient iterations.

(5) Local clustering: It measures how many corporation pairs that are directly connected to the corporation under consideration also share a direct connection with each other.



We obtain for node 3 a degree centrality of 3, a closeness centrality of 0.8, a Betweenness centrality of 4, a page-rank centrality of 0.2834 and a clustering of 0.33.

Figure 1: Stylized Example for an Undirected Graph with Non-weighted Connections

Figure 1 discusses an example of the measures above. All the measures can be extended to the continuous network setting. For a formal treatment we refer to *Jackson* (2008) or *van Steen* (2010). For the centrality measures high values indicate that a corporation has a high importance for the network, whereas large values in local clustering imply that a corporation's role in the network is essen-

tially replaceable by its neighboring corporations. Therefore, a high local clustering indicates that a corporation is redundant for the network and is of minor importance.

V. Analyses

We estimate the relationship between the most current rating decision (i. e. grade and outlook) between 01.01.2007 and 31.12.2014 per rating agency (i. e. Moody's, S&P and Fitch) for a corporation[7] and the corporation's network properties and medial position, and – as control variables – market and financial indicators. This approach is called 'shadow rating model' since it replicates the unknown rating model of the rating agencies and – once estimated – may be applied to corporations without external rating grade.

The dependent variable consists of two components, the rating grade and the outlook. We map the rating grade to an integer value according to Table 1. If a rating grade corresponds to more than one integer value, the grade is identified with the average of all eligible integers, e. g. Fitch's rating grade 'C' is mapped to 20.5. Afterwards, we reduce the result by 0.25 if the outlook is positive and increase it by 0.25 in case of a negative outlook. Hence, this transformation preserves the order imposed by the rating grade and refines it by the outlook. Even though we map the ratings to equally spaced numerical values, we allow for an exponential scaling as it is typically implied by rating scales via a Box-Cox transformation, see *Box/Cox* (1964), based on the control variables, which are discussed in the following section.

We consider the following financial ratios as control variables in our shadow rating model, all of which are taken from the most current financial statement with an accounting date that is at least three months before the corresponding rating decision: return on equity, free cash flow/sales, debt to equity ratio, short term debt to total debt, and revenue growth over the last three years as well as the total revenue as firm size measure. Moreover, we consider the stock return and the stock return volatility over the 30 days before the rating decision and include a binary variable for each rating agency.

To reflect the different structure of financial statements of financial institutions compared to non-financial corporations, financial institutions are either (1) flagged by a binary variable or (2) taken out of the sample. Binary variables for other industry sectors are not supported by the data. This set of variables cover all quantitative aspects traditionally incorporated in shadow rating models for large corporates and financial institutions.

[7] The restriction to the most current rating decision per corporation and rating agency stratifies the sample and prevents autocorrelation in the variables, *Ratha* et al. (2010).

Within the 30 days prior to a rating decision, we aggregate the sentiment of all news articles with respect to the corresponding corporation, standardize it with its standard deviation and denote it with $\text{Sentiment}(\cdot)$. Note that this variable is not restricted to the codomain [–1; 1] due to the division.

Table 1

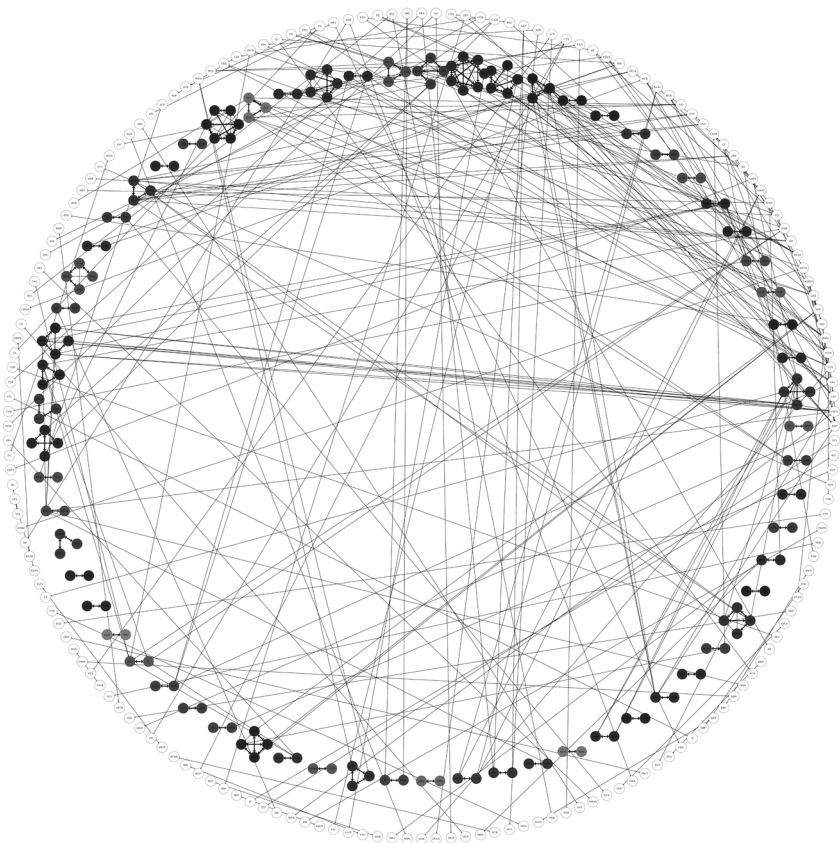
Mapping for Rating Grades from Moody's, S&P and Fitch on Integer Values

r(·)	Moody's	S&P	Fitch
1	Aaa	AAA	AAA
2	Aa1	AA+	AA+
3	Aa2	AA	AA
4	Aa3	AA–	AA–
5	A1	A+	A+
6	A2	A	A
7	A3	A–	A–
8	Baa1	BBB+	BBB+
9	Baa2	BBB	BBB
10	Baa3	BBB–	BBB–
11	Ba1	BB+	BB+
12	Ba2	BB	BB
13	Ba3	BB–	BB–
14	B1	B+	B+
15	B2	B	B
16	B3	B–	B–
17	Caa1	CCC+	CCC
18	Caa2	CCC	CC
19	Caa3	CCC–	CC
20	Ca	CC	C
21	Ca	C	C
22	C	SD	RD
23	C	D	D

Inspired by *Barber/Odean* (2008) and *Da/Engelberg/Gao* (2011), who document a causal relationship between the news attention of a corporation and its firm value, we also measure a corporation's media attention by

$$\text{Attention}(C,T) = \log(1 + \#\text{News}(C,T)),$$

where $\#\text{News}(C,T)$ denotes the number of news articles with non-zero relevance over a period of 30 days prior to T. The time intervals of 60 days and 90 days prior to a rating adjustment were analyzed as well, but imply weaker results. Therefore, we focus in the empirical section on the 30 day period.



A corporation is marked by a vertex and identified by the ticker symbol. A connection between two corporations is highlighted by an edge. See the pdf-file in the online appendix for a higher resolution and the centrality measures for each corporation.

Figure 2: Discrete Network

The centrality measures and the clustering coefficient are used a further explanatory variables. Since those statistics are static per corporation, no aggregation over time is required.

VI. Results

We find in 1,144,154 news articles out of 6,430,709 news articles at least one of our considered corporation name or synonym. This means that on average every 6th news article mentions at least one corporation under consideration. These news articles have an average length of 1,619 words with a standard deviation of 2,033 words.

By construction, the discrete and the continuous network display very different properties. The discrete network possesses 401 connections for a total of 666 corporations and is hence rather sparse. There are 8,090 corporation pairs directly or indirectly connected. 321 corporations are isolated and not connected with other corporations. The average path length in this graph is 3.62 with a standard deviation of 2.1. The longest path ranges over 12 corporations; this appears five times. Figure 2 shows the full graph, where each corporation is marked by its ticker symbol. Corporations connected by a strong relationship, i. e. corporations that are mentioned together in 10 % of the eligible news articles or more, are assigned to groups, which are placed in the inner ring and colored identically. The pdf-file in the online-appendix provides additional information on each corporation via tooltips, e. g. the full corporation name, the centrality measures and the local clustering coefficients. A manual review of the identified corporation groups grades our approach as plausible.

The continuous network possesses 113,538 non-zero connections and is hence much more closely meshed than the discrete one. There are only 7 corporations without any connection to another corporation. However, the average connection strength is 0.002 with a standard deviation of 0.009 and appears as rather weak.

Our sample contains 377 corporations out of 666 with at least one rating in the observation period.[8] These corporations break down into 64 corporations with ratings from all three rating agencies, 130 corporations with ratings from two agencies and 183 corporations with ratings from one agency. 50 corporations are financial institutions which are in our regression flagged by a binary variable or alternatively not included. A thorough analysis and model focused exclusively on financial institutions was not considered due to the limited sample size.

[8] Corporations without external rating are not excluded from the network construction since a corporation's value may also be susceptible to connections to non-rated corporations.

Table 2

Descriptive Statistics for Sentiment and Attention (Rows 1 and 2), Measures for the Discrete Network (rows 3 to 7) and for the Continuous Network (rows 8 to 12) for the Sample with Financial and Non-financial Institutions

Description	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	StdDev	%==0
Sentiment	–4.716	0.0000	0.3520	1.267	0.8530	51.207	5.5590	0.1055
Attention	0.0000	1.946	2.944	2.949	3.970	6.609	1.5516	0.0692
Degree	0.0000	0.0000	1.0000	1.6310	2.0000	8.0000	2.1275	0.4283
Closeness	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	<0.0001	0.4283
Betweenness	0.0000	0.0000	0.0000	11.3900	0.0000	205.8000	37.218	0.7701
Page Rank	0.0000	0.0000	0.0021	0.0017	0.0029	0.0056	0.0017	0.4283
Clustering	0.0000	0.0000	0.0000	0.2194	0.3333	1.0000	0.3671	0.6961
Degree	0.0000	0.4393	0.6822	0.8032	1.1190	2.1410	0.5107	0.0205
Closeness	0.0000	0.0035	0.0041	0.0040	0.0048	0.0055	0.0009	0.0205
Betweenness	0.0000	110.000	724.000	1657.000	2362.000	11314.000	2276.106	0.1087
Page Rank	0.0000	0.0011	0.0015	0.0017	0.0023	0.0038	0.0009	0.0205
Clustering	0.0000	0.6906	0.7251	0.7168	0.7688	0.8926	0.1192	0.0205

The shadow rating model is estimated on 635 observations if financial institutions are included and on 531 observations if financial institutions are excluded. Table 2 shows the distributional properties for the variables derived from the news articles considering the sample with both financial and non-financial institutions. Whereas the measures in rows 3 to 7 are derived from the discrete network, the measures in rows 8 to 12 correspond to the continuous network.

The distribution of $\text{Sentiment}(\cdot)$ is marginally skewed to the right, indicating that news have a positive attitude. For about 7 % of the observations there is no news article published before a rating decision which results in an attention equal to zero. Due to the sparseness of the discrete network, the centrality measures and the clustering coefficient are censored at zero for a significant proportion of the observations. In the case of the continuous network, the censoring is removed almost completely and the measures fluctuate enough to receive a smooth distribution. For the sake of comparison during the regression analysis all variables shown in Table 2 are standardized to the range [0; 1] and the sentiment to the range [–1; 1] by dividing each variable by its largest realization.

As it is unlikely that the relationship between the dependent variable and our explanatory variables is linear, we apply a Box-Cox transformation to the dependent variable, i. e. we allow for a power transformation, which is estimated

by the control variables exclusively and without the network properties, the medial attention and sentiment. With this approach we prevent the regression estimates on the network and media variables in the subsequent shadow rating model from overfitting effects caused by the Box-Cox transformation and isolate the relationship between the information extracted from the news and the rating grades.

The Box-Cox regression yields an exponent for the dependent variable of 0.7474 considering the full sample including financial institutions, and of 0.7878 if financial institutions are excluded. Both estimates imply a concave transformation of our equally spaced dependent variable, which is compatible with a typical (logarithmic) rating scale. For the following analysis, the transformed dependent variable is multiplied by –1 so that a positive regression coefficient for a variable in the shadow rating model implies a positive relationship between the creditworthiness and the variable.

The regression results, i. e. the estimated coefficients, the heteroscedasticity robust p-values, the regression's adjusted R2, Breusch-Pagan[9] and Likelihood-Ratio test, and the relative reduction in the 10-fold cross validation (hereafter: CV) error[10], are shown in Tables 4 and 5. Cross-validation is used to assess predictive performance: it tests whether a given model would perform well on yet unseen data. An increase in the CV-error when adding a new covariate indicates overfitting of the model to the training set. A lower CV-error generally indicates a better predictive power, see *Hastie/Tibshirani/Friedman* (2008) for a thorough discussion.

As comparison benchmark, the second column of both tables shows the estimates for a regression model based on the control variables exclusively, which is then enriched in a first step by the variables sentiment and attention, and afterwards by a company's network properties. The Likelihood Ratio test is applied using the simple model as the null-model. In order to facilitate the comparison of the CV-errors we report the reduction of the CV-error of a model compared to the error of the simple model.

Models 3 to 6 correspond to the discrete network, models 7 to 10 are based on the continuous network. Each of these models incorporates exactly one centrality measure, so that collinearity issues do not arise. More precisely, the variance inflation index is – for all regressions and all variables – less than 4, thus indicating that no significant covariance among the explanatory variables is present.

[9] The Breusch-Pagan tests for heteroscedasticity and shows whether the heteroscedasticity robust p-value are justified, which is true for all following regressions.

[10] In order to compare the different models we used the same 10-fold partitions for all regression models. In each test-fold the mean squared error was measured and averaged.

The control variables are throughout all regressions statistically significant with a plausible sign of the estimated regression coefficient with the exception of the stock return, which is not significant. When the news information variables are introduced to the regression model, the p-values of the debt structure and the revenue growth increase and indicate that the model starts to saturate.

The sentiment is significant if financial institutions are excluded, see Table 5. Then, corporations with positive sentiment receive a better rating grade than corporations with neutral or negative sentiment. Since the stock return over the same time frame is not significant, this might indicate that news articles contain valuable information that is not fully included in market prices and contradict strong market efficiency, see also *Heston/Sinha* (2016). Corporations with high news attention receive – ceteris paribus – a better rating grade, too, which is also consistent with *Barber/Odean* (2008). The results remain basically the same if the word lists from the 'Financial Sentiment Dictionary' are exchanged against the word lists from the 'General Inquirer'. These results are not shown here.

The regression coefficients for the local Clustering, Closeness and Page-Rank centrality are not (always) statistically significant if the discrete network is considered, which is presumably due to the large number of disconnected corporations and the strong censoring. However, for the Degree and Betweenness centrality we measure a statistically significant and economically plausible effect. Considering the weighted network, all network properties are statistically highly significant and imply a reasonable economic impact on the creditworthiness, i. e. corporations with a more central position receive – ceteris paribus – a better rating grade, and corporations that are redundant for the network (indicated by a high clustering) receive a lower rating grade.

Table 3

Regression Results Based on the Full Sample Including Financial Institutions

			Discrete Network				Continuous Network			
Model	1	2	3	4	5	6	7	8	9	10
Description	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)
Degree	--	--	0.5289 (0.008)	--	--	--	0.7922 (0.0022)	--	--	--
Closeness	--	--	--	0.1254 (0.5635)	--	--	--	1.0700 (0.0026)	--	--
Betweenness	--	--	--	--	0.9361 (0.0000)	--	--	--	1.1234 (0.0000)	--
Page Rank	--	--	--	--	--	0.0543 (0.6801)	--	--	--	0.8530 (0.0014)
Clustering	--	--	–0.2891 (0.0141)	–0.1138 (0.2749)	–0.1034 (0.3285)	–0.1145 (0.3031)	–1.1182 (0.0000)	–1.6317 (0.0000)	–0.9273 (0.0000)	–1.1670 (0.0000)
Sentiment	--	0.2705 (0.3602)	0.2610 (0.3757)	0.2407 (0.4159)	0.2323 (0.4346)	0.2588 (0.3845)	0.3601 (0.2278)	0.3782 (0.213)	0.3745 (0.204)	0.3642 (0.223)
Attention	--	0.9149 (0.0000)	0.7825 (0.0000)	0.9219 (0.0000)	0.7749 (0.0000)	0.9426 (0.0000)	0.5900 (0.0030)	0.7350 (0.0003)	0.7879 (0.0000)	0.5894 (0.0031)
30 days tock return	0.1870 (0.1319)	0.2003 (0.1031)	0.2316 (0.0634)	0.2068 (0.0957)	0.2256 (0.0662)	0.2086 (0.0933)	0.2078 (0.0806)	0.1854 (0.1188)	0.2131 (0.076)	0.2087 (0.0788)
30 days tock return volatility	–0.1002 (0.0062)	–0.1066 (0.0041)	–0.0964 (0.0110)	–0.1044 (0.0051)	–0.0986 (0.0082)	–0.1048 (0.0052)	–0.1022 (0.0058)	–0.1030 (0.0055)	–0.0980 (0.0072)	–0.1023 (0.0057)
Return on equity	0.0176 (0.0000)	0.0168 (0.000)	0.0167 (0.0000)	0.0168 (0.0000)	0.0172 (0.0000)	0.0167 (0.0000)	0.0167 (0.0000)	0.0164 (0.0000)	0.0166 (0.0000)	0.0168 (0.0000)
Free cash flow to sales	0.0121 (0.0001)	0.0097 (0.0015)	0.0084 (0.0081)	0.0096 (0.0022)	0.0085 (0.0044)	0.0096 (0.0020)	0.0093 (0.0023)	0.0102 (0.0006)	0.0103 (0.0006)	0.0094 (0.0019)
Debt-to-equity ratio	–0.3308 (0.0000)	–0.3434 (0.0000)	–0.3313 (0.0000)	–0.3377 (0.0000)	–0.3176 (0.0000)	–0.3420 (0.0000)	–0.3455 (0.0000)	–0.3376 (0.0000)	–0.3407 (0.0000)	–0.3470 (0.0000)
Short term debt to total debt	0.0323 (0.0178)	0.0312 (0.0188)	0.0318 (0.0149)	0.0310 (0.0200)	0.0314 (0.0159)	0.0311 (0.0176)	0.0321 (0.0120)	0.0320 (0.0123)	0.0355 (0.0036)	0.0321 (0.0121)
3 year revenue growth	0.0143 (0.0010)	0.0134 (0.0015)	0.0132 (0.0015)	0.0130 (0.002)	0.0125 (0.0021)	0.0131 (0.0019)	0.0135 (0.0011)	0.0127 (0.0025)	0.0127 (0.002)	0.0135 (0.0012)
Revenue	0.0076 (0.0000)	0.0051 (0.0000)	0.0046 (0.0008)	0.0051 (0.0001)	0.0044 (0.0010)	0.0052 (0.0001)	0.0042 (0.0011)	0.0048 (0.0001)	0.0030 (0.0176)	0.0043 (0.0007)
Adj. R2	0.3267	0.3516	0.3573	0.3509	0.3694	0.3505	0.3699	0.3708	0.3885	0.3706
LR Test	--	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Breusch-Pagan	0.0015	0.0109	0.0336	0.0161	0.0237	0.0110	0.0563	0.0336	0.0591	0.0581
Redu. CV-error	--	0.0347	0.0420	0.0326	0.0614	0.0313	0.0598	0.0596	0.0883	0.0600

The dependent variable is the Box-Cox transformed creditworthiness $\hat{y}$, i. e. the mapped rating/outlook y to the power of 0.7474, and the full regression equation is:

$$\hat{y} = a + b \times Cent + c \times Clust + d \times Sent + e \times Atten + \vec{f} \times \overrightarrow{Contr} + \varepsilon, \quad \hat{y} = \frac{y^{\lambda} - 1}{\lambda - 1} \; with \; \lambda = 0.7474.$$

Here *Cent* denotes exactly one centrality measure and *Clust* is the clustering coefficient. *Sent* is the sentiment and *Atten* the media attention both during the 30 days preceding the rating adjustment.

Table 4

Regression Results Based on the Sample Excluding Financial Institutions

			Discrete Network				Continuous Network			
Model	1	2	3	4	5	6	7	8	9	10
Description	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)	Coef. (P-val)
Degree	--	--	0.825 (0.0011)	--	--	--	1.2446 (0.0002)	--	--	--
Closeness	--	--	--	0.0064 (0.9808)	--	--	--	1.4324 (0.0018)	--	--
Betweenness	--	--	--	--	0.907 (0.0002)	--	--	--	1.4395 (0.0000)	--
Page Rank	--	--	--	--	--	0.188 (0.2722)	--	--	--	1.3764 (0.0001)
Clustering	--	--	–0.4172 (0.0025)	–0.1234 (0.3141)	–0.1319 (0.3008)	–0.1779 (0.1804)	–1.2796 (0.0000)	–1.9441 (0.0000)	–0.9584 (0.0005)	–1.356 (0.0000)
Sentiment	--	0.7188 (0.014)	0.6735 (0.0221)	0.6701 (0.0233)	0.6557 (0.0262)	0.6947 (0.0199)	0.9133 (0.0018)	0.9178 (0.0013)	0.8446 (0.0028)	0.9251 (0.0016)
Attention	--	0.8286 (0.0001)	0.6269 (0.0013)	0.8696 (0.0000)	0.6913 (0.0004)	0.8136 (0.0001)	0.3073 (0.1766)	0.5932 (0.0109)	0.6427 (0.0028)	0.2764 (0.2267)
30 days tock return	0.1999 (0.2906)	0.1648 (0.3851)	0.1985 (0.2977)	0.1734 (0.3650)	0.2111 (0.2678)	0.1795 (0.3497)	0.1982 (0.2807)	0.1727 (0.3537)	0.2332 (0.2106)	0.1986 (0.2770)
30 days tock return volatility	–0.3235 (0.0000)	–0.3354 (0.0000)	–0.3228 (0.0000)	–0.3365 (0.0000)	–0.3246 (0.0000)	–0.3335 (0.0000)	–0.3154 (0.0000)	–0.3146 (0.0001)	–0.2993 (0.0001)	–0.3131 (0.0000)
Return on equity	0.0171 (0.0000)	0.0159 (0.0000)	0.0158 (0.0000)	0.0160 (0.0000)	0.0167 (0.0000)	0.0158 (0.0000)	0.0154 (0.0000)	0.0150 (0.0000)	0.0157 (0.0000)	0.0154 (0.0000)
Free cash flow to sales	0.0269 (0.0000)	0.0231 (0.0000)	0.0201 (0.0001)	0.0231 (0.0000)	0.0204 (0.0001)	0.0227 (0.0000)	0.0229 (0.0000)	0.0245 (0.0000)	0.0233 (0.0000)	0.0233 (0.0000)
Debt-to-equity ratio	–0.4512 (0.0000)	–0.4405 (0.0000)	–0.4250 (0.0000)	–0.4442 (0.0000)	–0.4289 (0.0000)	–0.4428 (0.0000)	–0.4229 (0.0000)	–0.4218 (0.0000)	–0.4418 (0.0000)	–0.4255 (0.0000)
Short term debt to total debt	0.0347 (0.0350)	0.0315 (0.0499)	0.0330 (0.0356)	0.0307 (0.0561)	0.0321 (0.0418)	0.0320 (0.0425)	0.0325 (0.0323)	0.0323 (0.0349)	0.0364 (0.0125)	0.0326 (0.0313)
3 year revenue growth	0.0166 (0.0029)	0.0153 (0.0047)	0.0142 (0.0074)	0.0148 (0.0063)	0.0139 (0.0084)	0.0145 (0.0072)	0.0155 (0.0033)	0.0144 (0.0068)	0.0145 (0.0058)	0.0154 (0.0036)
Revenue	0.0082 (0.0000)	0.0064 (0.0001)	0.0055 (0.0015)	0.0065 (0.0002)	0.0055 (0.0012)	0.0063 (0.0002)	0.0053 (0.0012)	0.0063 (0.0001)	0.0041 (0.0129)	0.0054 (0.0008)
Adj. R2	0.3633	0.3852	0.3959	0.3838	0.3975	0.3850	0.4097	0.4061	0.4234	0.4116
LR Test	--	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Breusch-Pagan	0.0000	0.0000	0.0001	0.0000	0.0001	0.0000	0.0007	0.0001	0.0002	0.0007
Redu. CV-error	--	0.0209	0.0381	0.0121	0.0391	0.0195	0.0540	0.0506	0.0757	0.578

The dependent variable is the Box-Cox transformed creditworthiness $\hat{y}$, i. e. the mapped rating/outlook y to the power of 0.7878, and the full regression equation is:

$$\hat{y} = a + b \times Cent + c \times Clust + d \times Sent + e \times Atten + \vec{f} \times \overrightarrow{Contr} + \varepsilon, \quad \hat{y} = \frac{y^{\lambda} - 1}{\lambda - 1} \; with \; \lambda = 0.7878$$

Here *Cent* denotes exactly one centrality measure and *Clust* is the clustering coefficient. *Sent* is the sentiment and *Atten* the media attention both during the 30 days preceding the rating adjustment.

The information gained from news articles improves the model's goodness of fit by up to 8 % in terms of adjusted R2 from 36.3 % to 42.3 %. About one third of the increase is coming from a corporation's sentiment and attention and the other two third from the corporation's position in the network. The Likelihood-Ratio test also reveals that the models with news and network information significantly outperform the baseline model with traditional risk factors only.

The cross validation error (including financial institutions) of the models reduced by 3.5 %, compared to the base model, with only sentiment and attention included. Including the continuous network with Betweenness centrality, this reduction improved to 8.8 %. A similar improvement is seen if financial institutions are excluded. Cross validation favors the continuous models, i. e. increasing the reduction of 6.1 % for the discrete network to 8.8 % for the continuous model (including financial institutions using the Betweenness centrality).

VII. Interpretation

News articles are manifold and inform about any subject. Beside the subject, news articles contain valuable context provided by the author that helps interpreting the subject and give insights into the relationship between corporations. We presented strong evidence that this information is reflected in rating grades published by the major rating agencies, and hence, should be integrated in internal rating models and early warning systems as well.

To this aim, information that is often called "qualitative information" can be extracted from unstructured documents with state-of-the-art methods and does not require a manual and costly analysis, as was shown in this work. Moreover, the information gained from news articles may be integrated directly into rating systems as our analysis above indicates. Due to numerical tractability the network measures used are based on all news items and have no dynamics. Hence, they are observed contemporaneously with the rating grades, so that the out-of-sample accuracy of this approach cannot be answered completely and is open for future research. However, the cross validation analysis indicates that the network properties of a corporation are rather stabile.

We find that the continuous, fine-meshed network outperforms the discrete, spare network in predictive power. On the other hand, the former one implies a higher degree of complexity and abstraction. In addition, a simple intuitive visualization as shown in Figure 2 and the pdf-file in the online appendix is not possible for the fine-meshed continuous network. In practice one could implement the continuous network for numerical models and provide a representation of the discrete network for visualization of the key underlying structure of the network.

VIII. Summary and Outlook

In this paper we show that a network, obtained by simple rules from unstructured business news articles, provides a useful and structured factor for statistical models. Utilizing this factor in a shadow-rating model leads to a significant increase in the accuracy of our credit risk assessment.

Our approach could possibly be extended on a larger dataset with regards to the non-directed approach we took when constructing the network. It might proof beneficial to classify the connections into categories such as conflicting economic interest (e. g. offering of substitutable goods) and harmonized economic interest (e. g. cooperation, sponsor).

References

Acemoglu, D./*Ozdaglar*, A./*Tahbaz-Salehi*, A. (2015): Systemic Risk and Stability in Financial Networks, American Economic Review 105(2), 564–608.

Altman, E. I. (1968): Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, The Journal of Finance 23(4), 589–609.

Barber, B. M./*Odean*, T. (2008): All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors, The Review of Financial Studies 21(2), 786–818.

Bhojraj, S./*Sengupta*, P. (2003): Effect of Corporate Governance on Bond Ratings and Yields: The Role of Institutional Investors and Outside Directors, Journal of Business 76(3), 455–475.

Box, G. E. P./*Cox*, D. R. (1964): An Analysis of Transformations, Journal of the Royal Statistical Society. Series B (Methodological).

Cossin, D./*Schellhorn*, H. (2007): Credit Risk in a Network Economy, Management Science Vol. 53, 1604–1617.

Da, Z./*Engelberg*, J./*Gao*, P. (2011): In Search of Attention, Journal of Finance 66(5), 1461–1499.

Das, S. R./*Chen*, M. Y. (2007): Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web, Management Science 53(9), 1375–1388.

Eisenberg, L./*Noe*, T. H. (2001): Systemic Risk in Financial Systems, Management Science 47 (2), 236–249.

Elliott, M./*Golub*, B./*Jackson*, M. O. (2014): Financial Networks and Contagion, American Economic Review 104(10), 3115–3153.

Fagiolo, G./*Reyes*, J./*Schiavo*, S. (2007): On the Topological Properties of the World Trade Web: A Weighted Network Analysis, Working Paper.

Hałaj, G./*Kok*, C. (2013): Assessing Interbank contagion Using Simulated Networks, Working Paper.

Hastie, T./*Tibshirani*, R./*Friedman*, J. (2008): The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics.

Heston, S. L./*Sinha*, N. R. (2016): News versus Sentiment: Predicting Stock Returns from News Stories, Working Paper.

Jackson, M. O. (2008): Social and Economic Networks, Princeton University Press.

Kamstra, M./*Kennedy*, P./*Teck-Kin*, S. (2001): Combining bond rating forecasts using logit, The Financial Review 37, 75–96.

Loughran, T./*McDonald*, B. (2011): When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Journal of Finance 66, 35–66.

Mählmann, T. (2011): Is there a relationship benefit in credit ratings?, Review of Finance, 1–36.

Porter, M. F. (1980): An algorithm for suffix stripping, Program 14, 130–137.

Pozzi, F./*Di Matteo*, T./*Aste*, T. (2013): Spread of risk across financial markets: better to invest in the peripheries, Scientific Reports 3.

Ratha, D./*De*, P. K./*Mohapatra*, S. (2010): Shadow Sovereign Ratings for Unrated Developing Countries, World Development 39, 295–307.

Tetlock, P. C. (2007): Giving Content to Investor Sentiment: The Role of Media in the Stock Market, Journal of Finance 62, 1139–1167.

van Steen, M. (2010): An Introduction to Graph Theory and Complex Networks, Maarten van Steen.

Appendix I

In this stylized example we restrict ourselves to one news article that can be accessed in full with the following URL: http://www.reuters.com/article/us-chevron-results-idUSWNAS836520070727.

The news article consists of 513 words (with whitespace separated strings) and mentions six corporations. However, only two corporations are mentioned at least twice, namely Chevron Corp and Dynegy Inc., for which the news article is analyzed in the following. Table A1 shows an extraction of the news article and the corresponding word weights. The weight W_1 depends only on the position of a word within the news article that is captured through p. The weights $W_2(\text{CVX})$ and $W_2(\text{DYN})$, respectively, depend only on the word's distance to the nearest corporation identifier. The product of W_1 and $W_2(\text{CVX})$ gives $W(\text{CVX})$, and the product of W_1 and $W_2(\text{DYN})$ gives $W(\text{DYN})$.

We check for every single word if it appears on Loughran and McDonald's positive or negative word list. If a word is on the positive list, the corresponding word weights $W(\text{CVX})$ and $W(\text{DYN})$ are colored in green, if it is on the negative list, they are colored in red. Afterwards, we add up $W(\text{CVX})$ of all positive words and of all negative words, which determine the raw-sentiment and render the news as positive for Chevron Corp. The same is done for $W(\text{DYN})$, which also marks the news as positive for Dynegy Inc.

The relevance measure is based on sum over the column W_1 and the sum over the column $W(\text{CVX})$ and $W(\text{DYN})$, respectively. It shows that the news article has a higher relevance for Chevron Corp. than for Dynegy Inc.

The sum of all positive words weights for Chevron Corp. gives 1.92 and the sum of all negative word weights gives 0.01. Summing up all word weights for Chevron Corp. gives 143.23. The latter value and the coefficients in Table A1 are used to standardize the former values. For the positive words the standardized value is –0.62, indicating that there are slightly too few positive words used, and for negative words –2.20, also indicating that negative words are definitely underrepresented. This gives a raw-sentiment for Chevron Corp. of 0.56. The commitment measure is calculated in an analogous manner.

Table A1

An Extraction of the News Article that is Located Under http://www.reuters.com/article/us-chevron-results-idUSWNAS836520070727, and the Word Weights According to the Previously Introduced Formulas

p	Word	W_1	W_2 (CVX)	W (CVX)	W_2 (DYN)	W (DYN)
1	Chevron	1,0000	1,0000	1,0000	0,0181	0,0181
2	Corp	0,9997	0,9965	0,9962	0,0228	0,0228
3	(CVX.N)	0,9994	0,9862	0,9856	0,0286	0,0285
4	posted	0,9991	0,9692	0,9683	0,0356	0,0355
5	a	0,9988	0,9460	0,9448	0,0439	0,0439
6	better	0,9984	0,9169	**0,9154**	0,0539	**0,0538**
7	than	0,9981	0,8825	0,8808	0,0657	0,0656
8	expected	0,9978	0,8435	0,8417	0,0796	0,0794
9	24	0,9975	0,8007	0,7987	0,0956	0,0954
10	percent	0,9972	0,7548	0,7527	0,1142	0,1138
11	rise	0,9969	0,7066	0,7044	0,1353	0,1349
12	in	0,9965	0,6570	0,6547	0,1593	0,1588
13	quarterly	0,9962	0,6065	0,6042	0,1863	0,1856
14	earnings	0,9959	0,5561	0,5538	0,2163	0,2154
15	on	0,9956	0,5063	0,5041	0,2494	0,2483
16	Friday	0,9953	0,4578	0,4557	0,2855	0,2842
17	on	0,9950	0,4111	0,4090	0,3247	0,3230

18	higher		0,9946		0,3666	0,3646		0,3666	0,3646
19	profit		0,9943		0,3247	0,3228		0,4111	0,4088
20	from		0,9940		0,2855	0,2838		0,4578	0,4551
21	its		0,9937		0,2494	0,2478		0,5063	0,5031
22	refineries		0,9933		0,2163	0,2148		0,5561	0,5524
23	and		0,9930		0,1863	0,1850		0,6065	0,6023
24	a		0,9927		0,1593	0,1582		0,6570	0,6522
25	gain		0,9924		0,1353	**0,1343**		0,7066	**0,7013**
26	from		0,9920		0,1142	0,1133		0,7548	0,7488
27	the		0,9917		0,0956	0,0948		0,8007	0,7941
28	sale		0,9914		0,0796	0,0789		0,8435	0,8363
…	…		…		…	…		…	…
164	quarter		0,9407		0,0657	0,0618		0,0000	0,0000
165	dropped		0,9403		0,0796	**0,0748**		0,0000	**0,0000**
166	about		0,9399		0,0956	0,0899		0,0000	0,0000
…	…		…		…	…		…	…

Appendix II

Online-Only-Appendix
https://doi.org/10.3790/ccm.52.2.A97