The Network of Firms Implied by the News

G. Schwenkler and H. Zheng*

February 11, 2019†

Abstract

Data on business connections between firms are notoriously hard to collect although the network structure of firms matters for aggregate risks. This paper shows that readily available news provides sufficient information to infer a network of firms that resembles the true underlying network of firm connections. We develop a machine learning algorithm that takes news as input and outputs a network of firm interconnections. We empirically document that interconnectivity in our news-implied network captures contagion effects that affect asset prices and the macroeconomy. Our results enable the extraction of business networks from readily accessible data, facilitating the accurate measurement of risks.

Keywords: Networks, contagion, interconnections, machine learning, natural language processing.

*Schwenkler & Zheng are at the Department of Finance, Boston University Questrom School of Business. Schwenkler is corresponding author: Phone (617) 358-6266, email: gas@bu.edu, web: http://www.gustavo-schwenkler.com.

†We are grateful to Andrea Buffa, Kay Giesecke, Steven Kou, and seminar participants at Boston University.

1 Introduction

Firms are connected to each other through different types of business links: costumer-supplier relationships, trade credit, strategic relationships, banking, financing, and others. The literature has established that these links serve as spreaders of risk, affecting asset returns and the macroeconomy (see Acemoglu et al. (2012), Bai et al. (2015), Elliott et al. (2014), Fernando et al. (2012), Gabaix (2011), and, Jorion and Zhang (2009), among many others). Nevertheless, data on the interconnections between firms is notoriously hard to collect. Often, only incomplete and lagged data is available. In the U.S., corporations publish a fraction of their business links in their 10-K reports but this occurs only once a year. The Bureau of Economic Analysis reports on input-output links across sectors but these data are only updated once every 5 years. There are some interbank borrowing and lending data for the financial sector but these data are often proprietary and only available to central bankers. Constructing a timely network of interconnections between firms is extremely difficult given the currently available data sets.[1]

In this paper, we show that financial news published in common newspapers deliver sufficient information about the relationships between firms to be able to construct a timely and granular network of firm interconnections. We develop a machine learning algorithm that takes news data as an input and outputs a network of firm connections implied by the news. The news-implied networks we construct closely resemble business networks often extracted from alternative data sets. In contrast to the networks extracted from traditional data sources, however, our news-implied networks are available in high frequency and provide more detailed information about interconnections in the whole universe of firms. We show that measures of interconnectivity of the news-implied network positively relate to measures of financial risks. They capture contagion effects among corporations and predict periods of macroeconomic distress. Our news-implied networks thus enable the estimation of accurate measures of risks for the aggregate economy.

We exploit novel machine learning tools known as *natural language processing* (NLP) to extract from news data the names of corporations mentioned in the news and detect whether different corporations share a business link with each other. NLP has become increasingly popular in financial economics research; see Engle et al. (2019) and Jelveh et al. (2018) for recent applications of NLP for the analysis of climate change risk and the influence of political parti-

[1] A possibility is to consider variance decompositions for interconnected firms as in Diebold and Yılmaz (2014) and Demirer et al. (2018). However, networks derived from volatility data likely include risk premium effects that may not reflect the actual degree of connectivity under the physical measure.

sanship on economic research. Most commonly, NLP is used to estimate the sentiment of media content – that is, whether the media expresses mostly negative or mostly positive opinions – and how sentiment affects asset prices and macroeconomic factors; see Baker et al. (2012), Beber et al. (2015), Chen et al. (2014), Da et al. (2015), Das and Chen (2007), Engelberg et al. (2012), García (2013), Jegadeesh and Wu (2013), Tetlock (2007), and Shen et al. (2017), among others.[2] The application of NLP for sentiment analysis is mainly a univariate exercise: it extracts from a large dimensional data set of news article an aggregate measure of sentiment. In contrast, our approach extracts bivariate signals from the data. We use NLP to identify two firms that are connected to each other via a business link and assess how strong this relationship is. Our identifying assumption is that if two firms share a business connection, then the news should report about this business link in an article by mentioning the two firms in close proximity. The stronger the relationship is, the more often should the news report about this relationship in different articles. Based on these assumptions, we say there is a link between two distinct firms whenever these firms are mentioned within the same sentence of an article. We quantify the strength of the link by the frequency with which this link is identified in the data. Even though our strategy to identify business links from news data may seem stringent at first, it delivers plausible networks of firm connections that closely resemble networks that have been estimated from alternative data sets.

We obtain an extensive data set of financial news articles published by Reuters covering the years 2006 through 2013. Our data set includes over 100,000 news articles and spans over 7,000 firms. We apply our methodology to these news data and observe several interesting features. The network implied by our news data showcases a core consisting of large banks who are strongly interconnected. There are several smaller banks that are connected to these larger banks, making up a core-periphery structure for the financial sector. Core-periphery structures are often identified in empirical and theoretical studies of interbank networks; see Babus and Hu (2017), Farboodi (2017), Craig and von Peter (2014), in 't Veld and van Lelyveld (2014), and Gofman (2017). There are several non-financial clusters surrounding the financial sector, delivering a star architecture for the broader network of U.S. firms. A similar star structure was estimated by Acemoglu et al. (2012) for intersectoral trade networks based on input-output linkages from the Bureau of Economic Analysis. We analyze the dynamic evolution of the network over time. Some sectors become more or less prominent over time but the financial sector remains

[2]See Loughran and McDonald (2016) for a survey.

central and strongly connected. These observations provide empirical evidence for the centrality of financial sector in the U.S. economy, complementing the theoretical results of Bernanke et al. (1999) and Carvalho and Gabaix (2013) about the unique role played by the financial sector in exacerbating local shocks to the aggregate economy.

We compute monthly time series of interconnectivity measures for our news-implied networks and study how they relate to financial and macroeconomic factors in the United States. We first show that interconnectivity in our news-implied network is mostly unrelated to sentiment in the underlying news articles. These findings indicate that our news-implied network conveys information that is orthogonal to news media sentiment. We then show that the VIX is large and government bonds are expensive whenever the news-implied network becomes more strongly interconnected. We also show that the slope of the yield curve is positively steep in periods of strong interconnectivity. These results suggest that the degree of interconnectivity in our news-implied network reflects short-term risks that affect the aggregate stock market.

We question whether the degree of interconnectivity in the news-implied network predicts measures of macroeconomic distress. We find that interconnectivity negatively predicts industrial production and consumption growth as well as inflation at the monthly time horizon. This holds even after controlling for common predictors of these macroeconomic factors. We conclude that the risks reflected by the degree of interconnectivity in our news-implied network materialize and cause macroeconomic distress. In a final step, we assess why interconnectivity predicts periods of macroeconomic distress. For this, we obtain measures of credit risk contagion among U.S. corporation from Azizpour et al. (2018) and show that interconnectivity in our news-implied networks explains a significant fraction of the time variation of the contagion measure of Azizpour et al. (2018). These results demonstrate that links in our news-implied network reflect connections between corporations that serve as spreaders of risks across firms. These links facilitate the contagion of risk from one counterparty to the next, enabling the amplification of idiosyncratic shocks into macroeconomic distress as in the theoretical models of Acemoglu et al. (2012), Eisenberg and Noe (2001), Elliott et al. (2014), and Gabaix (2011).

The rest of this paper is organized as follows. Section 2 introduces our methodology. Section 3 describes our news data. Section 4 presents our empirical results and Section 5 concludes.

2 Methodology

We begin by introducing our methodology that takes news articles as input and outputs a network of firm connections implied by the news. Several steps go into our methodology. The first step consists of transforming the raw text data into a format that can be analyzed by our algorithms. Given the transformed data, we then proceed to extract firms from the text data and identifying relationships between different firms. Finally, given a set of relationships between firms, we construct a network of firms. We provide details for each one of these steps.[3]

2.1 Processing the raw text data

Consider a news article; see Figure 1 for an illustration. The news article can be viewed as a collection of sentences, each of which is a collection of words that play different roles in purveying information: nouns, verbs, adjectives, and so on. Our goal is to extract from the news articles tupels of the type $E = (F_1, F_2, T)$, where F_1 and F_2 are connected firms and T is a time stamp indicating the date when the connection was observed. Given that firms are by definition nouns, it is necessary that we take individual words from a news article and label each word as the part of speech they correspond to (verb, noun, etc.). This goal can be accomplished using natural language processing (NLP) algorithms, which are readily available nowadays. We use the Stanford *coreNLP* toolkit available in R (see Manning et al. (2005)). The coreNLP toolkit is one of the most popular natural language processing (NLP) toolkits among academics and practitioners. Atdag and Labatut (2013), Pinto et al. (2016), and Rodriquez et al. (2012) demonstrate the high accuracy of the coreNLP toolkit, which often outperforms available alternatives for the purpose of natural language processing.

A benefit of using the coreNLP toolkit is that it does not require preprocessing the data. We can feed in the raw news articles to the coreNLP algorithms, and it will output a matrix of words with its corresponding part-of-speech labels. Figure 1 illustrates the output delivered by coreNLP. It extracts words from the news article and it assigns a universal part-of-speech tag (*"upos"*).[4] The output contains one row per token, which can be either an individual word or a collection of words that naturally belong together. For each token, the output also keeps track of the document in which the token can be found (*"id"*), the number of the sentence (from top to bottom) in which the token can be found in the article (*"sid"*), and a unique identifier for

[3] For a more detailed overview of how to use text data for economic research, see Gentzkow et al. (2019).

[4] The definitions of the different upos tags can be found at `http://universaldependencies.org/u/pos/`.

the token within the article (*"tid"*). We refer to Arnold (2017) for a detailed overview of the coreNLP package and its output.

2.2 Identifying firms

We process the output of the coreNLP toolkit to identify firms from our text data. For this, we use a specially developed algorithm called *named entity recognition* (NER) that is available within the coreNLP toolkit (see Finkel et al. (2005) for an introduction). Given a collection of tokens extracted from the data, NER can identify whether a token refers to an "entity." It can also classify the type of entity the token is referring to. More precisely, the NER algorithm in the coreNLP toolkit aims to recognize named entities ("PERSON", "LOCATION", "ORGANIZATION", "MISC"), numerical entities ("MONEY", "NUMBER", "ORDINAL", "PERCENT"), and temporal entities ("DATE", "TIME", "DURATION", "SET"). Consider as an example the first sentence of the article in Figure 1: *"Several aspects of the tentative contract between General Motors Corp (GM.N) and the United Auto Workers union will be hard for Ford Motor Co. (F.N) and Chrysler LLC to match in labor talks expected to heat up in coming days, people familiar with the negotiations said."* The NER algorithm recognizes the following entities in this sentence (see Figure 2 for a sample output): (GM, ORGANIZATION), (Ford, ORGANIZATION), (Chrysler, ORGANIZATION), and (Tuesday, DATE). Even though the NER algorithm does not recognize United Auto Workers union as an entity, it performs well at recognizing all three corporations mentioned in the sentence. The NER algorithm of the coreNLP toolkit has been demonstrated to be highly accurate, with accuracy rates in the order of 80% (see Costa et al. (2017) and Dlugolinsky et al. (2013)). We feel confident delegating the recognition of firms to the NER algorithm given its demonstrated accuracy.

We run the NER algorithm on our data and collect all tokens with the classification "ORGANIZATION." We then remove from the set of recognized organizations all entities that are not corporations.[5] Even though the NER algorithm has high accuracy for recognizing whether a token is an organization or not, the output of NER may contain a non-negligible amount of error in identifying the name of the organization due to they way the article is written or just because different names can refer to one and the same organization (think of this as being "rounding"

[5]The label "ORGANIZATION" is not unique to corporations – government agencies and institutions may also be classified as organizations. Because of this, we manually remove any evident references to government and institutions, such as the Fed, the FDIC, the ECB, and the European Union. We are currently experimenting with ways to extend NER to recognize corporations within the class of organizations.

errors). For example, for an article about Toyota, NER may recognize "Toyota", "$700 million Deal by Toyota," "Toyota USA," "Toyota Motor Corporation," "Toyota Motor Credit," "Toyota Motors," and "Japan-based Toyota Motor Corporation" as different organizations. As human beings, we immediately realize that all of these tokens refer to the same firm, namely "Toyota." But how can we teach this to a computer? One way is to construct a dictionary containing all the names that people may use for each company. But constructing and maintaining such a dictionary would require a large amount of manual work. It would also yield a static dictionary that would have to be updated every time a new firm enters the market. To circumvent these issues, we propose an alternative procedure that is less arduous albeit less precise.

We first remove from the name assigned by NER to a firm all numbers, special symbols, adoptions, determiners, adverbs, and unreasonable postfixes such as "-based." We then remove all words that indicate business types (like "Co.," "Inc.," and "Ltd."). In the example above, this approach would leave us with "Toyota," "Deal Toyota," "Toyota USA," "Toyota Motor," "Toyota Motor Credit," "Toyota Motor," and "Toyota Motor". In a final step, we create a cluster of all firms with common words in their names and consider the most frequently mentioned entity name as the stem. Let F_S be the collection of firms in the cluster whose name is equal to the stem. For any firm $j \notin F_S$, we check whether the name of j is fully contained in the stem or the stem is fully contained in the name of j. If so, we add j to the F_S and we update the stem to be either the name of j or the prevalent stem, whichever is shorter. If not, then we remove j from the original cluster. We proceed iteratively until no more improvements of the firm name can be made. In our Toyota example, we would begin by calling "Toyota Motor" the stem given that it is the most prevalent name. We would then iterate through the firms named "Toyota," "Deal Toyota," "Toyota USA," and "Toyota Motor Credit." Given that "Toyota" is the shortest name fully contained in the stem, we would update "Toyota" to be the new stem. Then, because "Toyota" is contained in all other firm names in this cluster, we would update all other names to "Toyota" and terminate the iteration.

There are a few drawbacks of our approach. A first drawback is that we may often aggregate subsidiary firms and their mother firms (Toyota Motor Credit and Toyota in our example). This may not significantly influence our results if mother firms assume the risk of large losses of their subsidiaries, as is often observed in reality. A second drawback is that we may end up with the shortest nickname of a firm instead of its full official name. This poses difficulties when matching our firms to other databases, such as CRSP and Compustat. Furthermore, if a firm

has a short name and this firm is very prevalent in our data, then other less prevalent firms may be clustered with the more prevalent firm if they have longer names that include the short name. For example, "Delta" and "Delta Dental" may be merged together even though they refer to different firms. We find that such errors occurs only sporadically in the data so that they can be corrected manually. Finally, our method cannot identify abbreviations of companies (e.g., GM versus General Motors). For now, we manually adjust the most mentioned abbreviations like GM, GE, and others.

2.3 Identifying connections between firms

Once all firms are identified, we proceed to construct the tupels $E = (F_1, F_2, T)$ of connected firms. Our identifying assumption is that if two firms share some sort of business connections, then the news should report about this connection by mentioning both firms in the same news article within close proximity from each other. Based on this assumption, we identify a tupel $E = (F_1, F_2, T)$ whenever firms F_1 and F_2 are mentioned in the same sentence in an article published on date T. These tupels can be readily extracted from the output of the NER algorithm introduced in Section 2.2.

While our rule to identify connected firms may seem restrictive at first, our results show that our approach delivers reasonable networks of business connections that closely relate to macro and financial factors. We are in the process of extending our approach to introduce weighted links that allow us to differentially connect firms that are mentioned in one article. For example, if two firms are mentioned in the same article but not in the same sentence, then we could assign a lower weighted link to these two firms than if they were mentioned in the same sentence. While such an extension seems natural, we do not expect that our results will be significantly affected by implementing the extension.

2.4 Building the network

In a final step, we collect all tupels $E = (F_1, F_2, T)$ of firm connections and aggregate them across time. We plot the collection of business connection in network form using the package "igraph" in R. An example of a network implied by our news data can be found in Figure 4. Each node corresponds to a corporation. The size of the node is proportional to the number of times that corporation appears in one of the tupels $E = (F_1, F_2, T)$ either as F_1 or F_2. The width of the links between firm F_1 and firm F_2 is proportional to the number of times F_1 and

F_2 are identified to be in a relationship.

We can aggregate the connections at different frequencies, allowing us to build time series of networks implied by the news data. For example, we can aggregate all tupels $E = (F_1, F_2, T)$ for which T falls in a specific month and roll this over month by month. Doing this would give us a monthly time series of news-implied networks. Our approach enables the construction of business networks at arbitrary frequencies. This is generally not possible with traditional data. Corporations report some of their business connections with other firms in their quarterly earnings. One could therefore use earnings reports data to construct quarterly business networks. To the extent that the news report about more business connections that is reported in earnings reports, our news-implied networks will be of higher frequency and more complete than analogous networks constructed from earnings reports.

3 Data

We obtain an extensive full-text news dataset from Ding et al. (2015). The data contains U.S. news articles from Reuters financial news published between October 20, 2006, and November 20, 2013. There are 106,521 articles in total. Table 1 provides summary statistics of the news articles in the data. We see that an average article is fairly large, including about 600 words and 21 sentences. There is also significant variability across articles: One article contains over 6000 words while others are fairly short with only one sentence. Figure 3 shows that the number of articles published in a new is fairly constant throughout our data sample (for the year 2006 we have a much shorter data sample).

We apply the methodology introduced in Section 2 to analyze the data. Our methodology recognizes 944,167 mentions of entities classified as "organizations." Figure 3 shows the number of recognized organization mentions in each year. Except for the year 2006 for which we have a shorter data sample, we see that our algorithm recognizes between 150,000 and 200,000 mentions in any given year. We also see that the number of organizations recognized in any given year is fairly constant. Of course, not every mention corresponds to a different organization. In total, our algorithm identifies 7,642 different firms during the time span covered by the data. The five most frequently mentioned firms are General Motors, Citigroup, J.P. Morgan Chase, Bank of America, and Goldman Sachs.

Table 1 indicates that an average article mentions a large number of firms (about 5). How-

ever, we are only able to recognize a low number of business connections from any given article: On average, we identify 0.38 connections per article because we restrict ourselves to connection being established only when two firms are mentioned in the same sentence. This limitation will be mitigated when we allow for weighted links between organizations that are mentioned in adjacent sentences in an article. Still, even in our restricted current setting, we are able to identify about 6,000 business connections in any given year considering the large amount of news articles published every year.

4 Empirical results

4.1 Full network

We plot in Figure 4 the full network of firms implied by all of the news articles in our data sample covering the years 2006 through 2013. For the sake of clarity, in the plot we only show the largest 50 nodes. These nodes represent the firms that most frequently appear in a connection with other firms in the whole data sample.

Just from this restricted representation, we can observe several interesting features. We first see that the big banks – Citigroup, Goldman Sachs, J.P. Morgan Chase, Bank of America, Morgan Stanley, UBS, and Barclays – represent some of the largest and most central nodes in our network, suggesting that the news reported very frequently about relationship between these major banks and other firms. The large banks are also highly interconnected, indicating that the news often reported about the relation between big banks. There are several smaller banks that lie on the periphery: Deutsche Bank, HSBC, Credit Suisse, RBS, ABN Amro, and Wells Fargo. Banks in the network of Figure 4 have a core-periphery structure with large banks being highly central and highly interconnected and smaller banks being connected to the larger banks on the outskirts. Such a core-periphery network is often observed in interbank data; see Afonso et al. (2013), Bech and Atalay (2010), Craig and von Peter (2014), in 't Veld and van Lelyveld (2014), and Gofman (2017), among others. Core-periphery networks have also been demonstrated to arise naturally in interbank network formation models; see Babus and Hu (2017) and Farboodi (2017). In contrast to the networks constructed from interbank data, which are often proprietary and available only to researchers and central bankers, our news-implied network can be extracted from data that is readily available to the public.

The network in Figure 4 also highlights the central position of the banking sector in the

general economy. We see that most non-financial firms (except General Motors) are located in the outskirts of the network, surrounding the large banks in the center. Several firms are only indirectly connected because they share a common link with one of the banks. For example, Chrysler and Microsoft are indirectly connected in the network of Figure 4 through a chain of intermediary banks that include Lehman Brothers, Credit Suisse, Morgan Stanley, and Barclays. To the extent that the links identified from the news data serve as spreaders of risk, the network in Figure 4 provide empirical support for the financial accelerator model of Bernanke et al. (1999). We show in 4.3 that several properties of our news-implied network relate to measures of macroeconomic and financial distress. Those findings suggest that the links in Figure 4 indeed serve as spreaders of risk across firms. Our observations about the centrality of the financial sector complement the results of Carvalho and Gabaix (2013), who empirically demonstrate that the large increases in macroeconomic volatility experienced in recent times are often due to instabilities in the financial sector.

Figure 4 exhibits several clusters associated with non-financial sectors. On the bottom left corner, we find a cluster of firms associated with the transportation sector. In the top part of the network we find a cluster associated with the technology and communications sectors. These clusters arise because the news often report about connections between firms in the same sector in addition to intersectoral relationships. The general architecture of the news-implied network resembles the star network of intersectoral connections estimated by Acemoglu et al. (2012) from input-output linkage data for the United States. The star network structure is not particular to the largest 50 firms we plot in Figure 4. We observe a similar structure in networks that include more firms; see Figure 5 for a sample network for the largest 500 nodes in the full data sample.

Input-output linkages are published by the Bureau of Economic Analysis. Linkages are measured across sectors and updated only every 5 years. In contrast, news articles are available in high frequency and they report about linkages across firms. Our news-implied firm networks can therefore provide more immediate and granular information about inter and intra-sectoral linkages and how they affect the macroeconomy. These kinds of insights are inaccessible from the Bureau of Economic Analysis data.

We also observe in Figure 4 that there are several largest nodes in the full network that correspond to non-financial firms: General Motors, Chrysler, Microsoft, Google, and Apple. These firms are highly important in their respective industries and contribute greatly to the U.S. economy. The news pick up on their importance and often report about their relationships

with other firms.

To wrap up our analysis of the full network, we study the distribution of the degree of the firms in our networks.[6] Figure 6 shows that the distribution of the degrees in our network is highly skewed with a heavy right tail: The smallest nodes in our news-implied network have 1 link (e.g., Freedom Bank, a small local bank in Indiana), the largest node has over 21 thousand links (General Motors), and a median node has 13 links (e.g., NBC Universal). The degree distribution in is well approximated by a power law with exponent equal to 1.83. Similar power laws have been estimated for cross-sectional distributions of other economic and financial data where inequality is a key feature, such as city sizes, firm sizes, and the degree distribution in intersectoral input-output networks; see Gabaix (2009) and Carvalho (2014).

4.2 Time series of networks

We proceed to construct time series of news-implied networks. In Figures 7 and 8 we plot yearly time series of networks implied by news articles in our data sample.[7] For each year between 2006 and 2013, we use the methodology of Section 2 to extract all business connections implied by articles published in that year and aggregate them to network plots. For simplicity, we only plot the connections between the largest 50 firms in every year, where the size of the firm is measured by how frequently that firm is found to be connected to a different firm. We do this just for yearly times series in Figures 7 and 8 for the sake of brevity. However, similar plots can be constructed for arbitrary frequencies – as high frequency as daily or hourly and as low frequency as quarterly or annual.

The time series of news-implied networks yield several interesting insights. We see that the architecture of the news-implied network can change drastically from year to year, according to how the news report about the relationships between different firms. We see that, in 2006 and 2007, the news-implied networks were relatively sparse with few clusters: a central one associated to the financial sector and some non-financial clusters dispersed in the periphery. Entering the financial crisis in 2008, the news-implied network became strongly interconnected with a strongly connected core made up of banks. In 2009, our news-implied network showcases an additional interconnected cluster of automobile firms, consistent with the prevailing crisis in the automobile sector that year. After 2010, when the great recession ended, the news-implied

[6] The degree of a firm is equal to the number of links that firm has with other firms in the network.

[7] Note that we have less data for the years 2006 and 2013 than for the other years.

networks again showcase a more common star structure as in Acemoglu et al. (2012), with banks located in the center and other sectors positioned around the financial sector.

We observe that some sectors become more prevalent over time while others become less prevalent. For example, we see that a small cluster of oil firms shows up in 2011 and vanishes soon after, reflective of the shock that the oil industry experienced after the Arab spring.[8] The technology sector cluster appears to become more prominent and interconnected in 2012, consistent with the fact that the technology sector outperformed other sectors that year.[9] In 2013, we see a cluster of airlines showing up, reflective of the boom experienced by the airline industry in 2013 and going into 2014.[10] All in one, the visual inspection of the time series of our networks suggest that our approach is able to extract from news data relevant signals about firms and sectors that are booming and busting over time.

We formally analyze the information contained in the time series of news-implied networks. For this, we consider first and second-order interconnectivity measures. These measures were introduced by Acemoglu et al. (2012) to quantify the impact of the structure of a production network on aggregate risk. The first-order interconnectivity measure is given by the coefficient of variation of the degree distribution in the network and indicates how strongly an average node is connected to other nodes through direct links. In contrast, the second-order interconnectivity measure highlights how strongly two nodes are indirectly connected through a third node. Acemoglu et al. (2012) demonstrate that aggregate volatility in a multisector production economy may be excessively high whenever the first-order or the second-order connectivity measures are high. In those cases, shocks to local sectors do not vanish quickly enough as they propagate through the network, resulting in large aggregate fluctuations. Based on the results of Acemoglu et al. (2012), we conjecture that if our news-implied networks indeed capture business connections between different firms that serve as spreaders of risk, the degree of interconnectivity of our news-implied networks should spike up during recessionary periods.

To validate our conjecture, we compute monthly time series of the first and second-order interconnectivity measures for our news-implied networks and plot the time series in Figure 9.[11] We see that our interconnectivity measures indeed spike up during the great recession. The time

[8] See http://www.economist.com/leaders/2011/03/03/the-2011-oil-shock.

[9] See http://www.theguardian.com/business/2012/dec/06/technology-sector-growing-faster-economy.

[10] See https://www.economist.com/gulliver/2013/12/27/good-times-for-the-airline-industry.

[11] For this, we compute analogous networks as those in Figures 7–8 but on a monthly basis and then compute the resulting interconnectivity measures.

series of interconnectivity measures appear to be persistent. To further assess the influence of the recession in 2008-2010 on our interconnectivity measures, we regress these on a recession indicator and lagged values of themselves. Table 2 summarizes our findings. We find that both time series are persistent, with first-order AR(1) coefficients in the order of 0.3. The recession indicator is significant and positively related to both interconnectivity measures. An F-test for the null hypothesis that the recession indicator does not influence an interconnectivity measure is rejected for both measures at the 99% confidence level. The recession indicator subsumes some of the persistence of the second-order interconnectivity measure.

Finally, we evaluate the relationship between interconnectivity and sentiment in our data. One concern about our interconnectivity measures is that they may be high in recessionary periods because those are periods in which the news becomes more pessimistic and reports about negative shocks affecting several firms at the same time.[12] If this were correct, then our algorithms would establish connections between firms and these connections would be found in articles with predominantly negative sentiment. We check whether this is the case by regressing our monthly interconnectivity measures on the the average news article sentiment in a given month and the recession indicator. Figure 10 plots the time series of the average article sentiment in our sample and Table 3 summarizes the results of our regressions. We find that sentiment explains only a small and mostly insignificant fraction of the time variation in our interconnectivity measures. These findings reject the notion that spikes in our interconnectivity measures are mainly driven by sentiment. They also suggest that interconnectivity in our news-implied networks conveys information that is orthogonal to the sentiment of the news articles from which we extract our networks.

4.3 Relation to macro and financial factors

We evaluate the information contained in the time series of news-implied network about financial and macroeconomic conditions. Figures 11 and 12 plot the time series of our interconnectivity measures in conjunction with relevant financial and macroeconomic factors. Table 4 provides summary statistics of our macro and financial factors. We see that the first-order interconnectivity measure is strongly correlated with the level and slope of the yield curve. The second-order interconnectivity measure appears to be more strongly correlated with equities and macroeconomic factors.

[12] Such a relationship between sentiment and recessions was empirically documented by García (2013).

4.3.1 Equities

We begin by analyzing the relationship between our interconnectivity measures and the stock market. Figures 11 and 12 suggest that the interconnectivity measures of our news-implied networks are somehow associated with market returns and volatility. To assess the significance of these relationships, we run contemporaneous regressions of the returns of the S&P 500 on the interconnectivity measures, the VIX, and several controls. Table 5 summarizes our findings. We find some evidence of a positive relationship between S&P 500 returns and the first-order interconnectivity measure when we control for the well-known leverage effect between returns and volatility.[13] However, this relationship vanishes when we also control for sentimentand other macro factors, with the loading on sentiment being significantly negative. These finding confirm similar results in the literature about the connection between sentiment and stock market returns (see García (2013) and Tetlock (2007)). They also validate our results from Section 4.2 about how our interconnectivity measures convey information that is different than sentiment.

Table 6 shows that the first and second-order interconnectivity measure are also positively related to the VIX even when controlling for the leverage effect, sentiment, and the state of the macroeconomy. Given that the VIX is a forward-looking measure of volatility, we interpret from the results of Tables 5 and 6 that the first-order interconnectivity measure contains information about stock market uncertainty that is not fully revealed by macroeconomic factors, past market returns, and sentiment. Our findings provide model-independent empirical evidence in support of the theoretical results of Acemoglu et al. (2012) and Gabaix (2011) that highlight that aggregate volatility is high in periods in which production networks become highly interconnected.

4.3.2 Macroeconomy

We study the information contained in our news-implied network about the state of the macroeconomy. We begin by analyzing the relationship between our interconnectivity measures and the yield curve. Our results for equities suggest that the interconnectivity measures are high when volatility is high. In those periods when stocks are risky, government bonds become attractive. Because of this, we expect to see a negative relationship between the interconnectivity measures and the level of the yield curve. Table 7 confirms this intuition. The results of Table 7 provide additional evidence that our news-implied networks capture information about periods of distress for the aggregate stock market.

[13] The leverage effect states that returns are low when volatility is high and vice-versa

Our empirical results so far validate the hypothesis that our news-implied networks capture information about aggregate market risk. Next, we ask whether market participants believe these risks are short-lived or long-lived. To answer this question, we regress the slope of the yield curve on contemporaneous realizations of our interconnectivity measures and other controls. If market participants believed the risks captured by our news-implied networks were short-lived, then the short end of the yield curve should be affected by changes in our interconnectivity measures but not the long end. In this case, we would expect to see a positive relationship between interconnectivity and the slope of the yield curve. In contrast, if the risks were viewed as long-lived, then the long end of the yield curve should react and not the short end. In this case, we would expect to see a negative relationship between interconnectivity and the slope of the yield curve. The regression results of Table 8 show that there is a positive relationship between the slope of the yield curve and the first-order interconnectivity measure even when controlling for relevant factors. These results suggest that the first-order interconnectivity measure contains information about short-lived risks that affect the aggregate market.

We ask whether the risks reflected in our news-implied networks materialize and affect the real economy. For this, we run predictive regressions for the industrial production growth rate, the consumption growth rate, and inflation.[14] Table 9 summarizes our findings. We see that the interconnectivity measures significantly predict industrial production, consumption growth, and inflation at the monthly frequency even after controlling for their lagged realizations. These findings suggest that the our news-implied networks contain information about financial risks that materialize in the short run and affect the macroeconomy.

4.3.3 Contagion

Up to this point, our findings indicate that the news contain and transmit relevant information about risks that affect the stock market and the macroeconomy. Our results suggest that the links reflected in our news-implied networks capture connections between firms that serve as spreader of risk. We now test whether this conclusion is supported by the data.

Several theoretical papers argue that contractual links between corporations can spread credit risk among counterparties, amplifying local shocks into aggregate distress through a mechanism known as *contagion*; see Acemoglu et al. (2012), Eisenberg and Noe (2001), Elliott et al.

[14] We neglect GDP growth here due to the lack of data (we only have 28 quarterly observations). We also ignore the recession indicator for now as there was only one recession in our data sample.

(2014), and Gabaix (2011), among others. Recenlty, Azizpour et al. (2018) analyze the timing of defaults of U.S. corporations between 1960 and 2012 and find that contagion is a key driver of U.S. aggregate credit risk. Azizpour et al. (2018) identify contagion as the part of the aggregate default rate that cannot be explained by observable and latent systematic risk factors and instead can be explained by the timing and magnitude of past defaults.[15] Azizpour et al. (2018) find that the measure of contagion they extract from U.S. aggregate default timing data is significantly large and time-varying. The contagion measure of Azizpour et al. (2018) is particularly large during periods of default clustering, when it can explain up to 90% of the aggregate default rate in the United States.

If, indeed, the links in our news-implied networks captured connections between firms that serve to spread credit risk, then we would expect to see a positive association between our interconnectivity measures and the contagion measure of Azizpour et al. (2018). We proceed to validate this conjecture. We obtain monthly data on the contagion measure of Azizpour et al. (2018). Figure 13 plots the contagion measures and compares it to our interconnectivity measures. There appears to be significant correlation between our interconnectivity measures and contagion.

We analyze this relationship by regressing the measure of contagion of Azizpour et al. (2018) on our interconnectivity measures and other factors that drive contagion in the data; see Table 10. We find that our interconnectivity measures explain a large fraction of the time variation of the contagion measure of Azizpour et al. (2018). Both t and F-test reject the null hypothesis that interconnectivity does not explain contagion. Consistent with the notion of contagion of Azizpour et al. (2018), which is more concerned with the direct impact of a default on affected firms, we find that it is the first-order interconnectivity measure that most closely relates to contagion. First-order interconnectivity contains significant information about contagion even after controlling for risk factors that generally drive contagion in the data. All in one, these results demonstrate that the links in our news-implied networks reflect connections between firms that facilitate contagion of credit risk across counterparties.

[15] More precisely, Azizpour et al. (2018) formulate the aggregate default rate as an additive function of observable and latent systematic factors, and a self-exciting component that jumps whenever a default occurs. The jump is proportional to the logarithm of the amount of debt defaulted on. The impact of past defaults on the current aggregate default rate fades out exponentially.

5 Conclusion

We demonstrate that the news contains timely and granular information about interconnections between firms that is otherwise inaccessible from alternative news sources. We develop a methodology that takes news articles as input, aggregates them over time, and extract a network of firm connections implied by the news. A node in our network corresponds to a firm mentioned in the news and the links between nodes reflect how frequently two firms are mentioned in relationship to each other. We construct interconnectivity measures for our news-implied networks and show that the interconnectivity measures capture information about contagion effects that impact asset prices and the macroeconomy. The results of our paper enable the development of measures of risks that accurately reflect the network interconnections of firms.

References

Acemoglu, Daron, Vasco M. Carvalho, Asuman Ozdaglar and Alireza Tahbaz-Salehi (2012), 'The network origins of aggregate fluctuations', *Econometrica* **80**(5), 1977–2016.

Afonso, Gara M., Anna Kovner and Antoinette Schoar (2013), Trading partners in the interbank lending market, Technical Report 620, Federal Reserve Bank of New York.

Arnold, Taylor (2017), 'A tidy data model for natural language processing using cleannlp', *The R Journal* **9**(2), 248–267.

Atdag, Samet and Vincent Labatut (2013), A comparison of named entity recognition tools applied to biographical texts, *in* 'Proceedings of the 2nd International Conference on Systems and Computer Science', IEEE, pp. 228–233.

Azizpour, S., K. Giesecke and G. Schwenkler (2018), 'Exploring the sources of default clustering', *Journal of Financial Economics* **129**(1), 154–183.

Babus, Ana and Tai-Wei Hu (2017), 'Endogenous intermediation in over-the-counter markets', *Journal of Financial Economics* **125**(1), 200–215.

Bai, Jennie, Pierre Collin-Dufresne, Robert S. Goldstein and Jean Helwege (2015), 'On bounding credit-event risk premia', *Review of Financial Studies* **28**(9), 2608–2642.

Baker, Malcolm, Jeffrey Wurgler and Yu Yuan (2012), 'Global, local, and contagious investor sentiment', *Journal of Financial Economics* **104**(2), 272–287.

Beber, Alessandro, Michael W. Brandt and Maurizio Luisi (2015), 'Distilling the macroeconomic news flow', *Journal of Financial Economics* **117**(3), 489 – 507.

Bech, Morten L. and Enghin Atalay (2010), 'The topology of the federal funds market', *Physica A: Statistical Mechanics and its Applications* **389**(22), 5223–5246.

Bernanke, Ben S., Mark Gertler and Simon Gilchrist (1999), The financial accelerator in a quantitative business cycle framework, Handbook of Macroeconomics, Elsevier, Amsterdam, pp. 1341–1393.

Carvalho, Vasco M. (2014), 'From micro to macro via production networks', *Journal of Economic Perspectives* **28**(4), 23–48.

Carvalho, Vasco and Xavier Gabaix (2013), 'The great diversification and its undoing', *American Economic Review* **103**(5), 1697–1727.

Chen, Hailiang, Prabuddha De, Yu (Jeffrey) Hu and Byoung-Hyoun Hwang (2014), 'Wisdom of crowds: The value of stock opinions transmitted through social media', *The Review of Financial Studies* **27**(5), 1367–1403.

Costa, C. M., G. Veiga, A. Sousa and S. Nunes (2017), Evaluation of stanford ner for extraction of assembly information from instruction manuals, *in* '2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)', pp. 302–309.

Craig, Ben and Goetz von Peter (2014), 'Interbank tiering and money center banks', *Journal of Financial Intermediation* **23**(3), 322–347.

Da, Zhi, Joseph Engelberg and Pengjie Gao (2015), 'The sum of all fears investor sentiment and asset prices', *The Review of Financial Studies* **28**(1), 1–32.

Das, Sanjiv R. and Mike Y. Chen (2007), 'Yahoo! for amazon: Sentiment extraction from small talk on the web', *Management Science* **53**(9), 1375–1388.

Demirer, Mert, Francis X. Diebold, Laura Liu and Kamil Yılmaz (2018), 'Estimating global bank network connectedness', *Journal of Applied Econometrics* **33**(1), 1–15.

Diebold, Francis X. and Kamil Yılmaz (2014), 'On the network topology of variance decompositions: Measuring the connectedness of financial firms', *Journal of Econometrics* **182**(1), 119 – 134.

Ding, Xiao, Yue Zhang, Ting Liu and Junwen Duan (2015), Deep learning for event-driven stock prediction, *in* 'Proceedings of the 24th International Conference on Artificial Intelligence', AAAI Press, pp. 2327–2333.

Dlugolinsky, Stefan, Marek Ciglan and Michal Laclavik (2013), Evaluation of named entity recognition tools on microposts, *in* '2013 IEEE 17th International Conference on Intelligent Engineering Systems (INES)', pp. 197–202.

Eisenberg, Larry and Thomas H. Noe (2001), 'Systemic risk in financial systems', *Management Science* **47**(2), 236–249.

Elliott, Matthew, Benjamin Golub and Matthew O. Jackson (2014), 'Financial networks and contagion', *American Economic Review* **104**(10), 3115–3153.

Engelberg, Joseph E., Adam V. Reed and Matthew C. Ringgenberg (2012), 'How are shorts informed?: Short sellers, news, and information processing', *Journal of Financial Economics* **105**(2), 260–278.

Engle, Robert, Stefano Giglio, Heebum Lee, Bryan Kelly and Johannes Stroebel (2019), 'Hedging climate change news', *Review of Financial Studies* . forthcoming.

Farboodi, Maryam (2017), Intermediation and voluntary exposure to counterparty risk. Working Paper.

Fernando, Chitru S., Anthony D. May and William L. Megginson (2012), 'The value of investment banking relationships: evidence from the collapse of Lehman Brothers', *The Journal of Finance* **67**(1), 235–270.

Finkel, Jenny Rose, Trond Grenager and Christopher Manning (2005), Incorporating non-local information into information extraction systems by gibbs sampling, *in* 'Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)', pp. 363–370.

Gabaix, Xavier (2009), 'Power laws in economics and finance', *Annual Review of Economics* **1**(1), 255–294.

Gabaix, Xavier (2011), 'The granular origins of aggregate fluctuations', *Econometrica* **79**(3), 733–772.

García, Diego (2013), 'Sentiment during recessions', *The Journal of Finance* **68**(3), 1267–1300.

Gentzkow, Matthew, Bryan T. Kelly and Matt Taddy (2019), 'Text as data', *Journal of Economic Literature* . forthcoming.

Gofman, Michael (2017), 'Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions', *Journal of Financial Economics* **124**(1), 113–146.

in 't Veld, Daan and Iman van Lelyveld (2014), 'Finding the core: Network structure in interbank markets', *Journal of Banking & Finance* **49**, 27–40.

Jegadeesh, Narasimhan and Di Wu (2013), 'Word power: A new approach for content analysis', *Journal of Financial Economics* **110**(3), 712 – 729.

Jelveh, Zubin, Bruce Kogut and Suresh Naidu (2018), Political language in economics. Working Paper, Columbia Business School.

Jorion, Philippe and Gaiyan Zhang (2009), 'Credit contagion from counterparty risk', *The Journal of Finance* **64**(5), 2053–2087.

Loughran, Tim and Bill McDonald (2016), 'Textual analysis in accounting and finance: A survey', *Journal of Accounting Research* **54**(4), 1187–1230.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David Mc-Closky (2005), The Stanford CoreNLP natural language processing toolkit, *in* 'Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations', ACL, pp. 363–370.

Pinto, Alexandre, Hugo Gonçalo Oliveira and Ana Oliveira Alves (2016), Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text, *in* 'Proceedings of the 5th Symposium on Languages, Applications and Technologies (SLATE '16)', Vol. 51, pp. 3:1–3:16.

Rodriquez, Kepa Joseba, Mike Bryant, Tobias Blanke and Magdalena Luszczynska (2012), Comparison of named entity recognition tools for raw OCR text, *in* 'Proceedings of KONVENS 2012', ÖGAI, pp. 410–414.

Shen, Junyan, Jianfeng Yu and Shen Zhao (2017), 'Investor sentiment and economic forces', *Journal of Monetary Economics* **86**, 1 – 21.

Tetlock, Paul C. (2007), 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of Finance* **62**(3), 1139–1168.

Variable	Mean	Std dev.	Min	Max
Number of words per article	583	359	19	6658
Number of sentences per article	20.61	13.34	1	253
Number of unique organizations identified in an article	5.35	3.82	1	49
Number of connections identified in an article	0.38	3.1	0	1430

Table 1: Summary statistics of news articles in our data set. We consider 106,521 news articles from Reuters financial news published between October 20, 2006, and November 20, 2013. We only keep news articles whose news ID starts with "USN".

	First-order	First-order	Second-order	Second-order
Intercept	*** 0.593	*** 0.646	104.757	*** 127.24
	(6.638)	(7.404)	(6.586)	(8.210)
Lagged interconnectivity	** 0.302	* 0.220	0.312	0.130
	(2.893)	(2.119)	(3.007)	(1.252)
Recession indicator		*** 0.083		*** 23.500
		(2.969)		(4.161)
Number of observations	84	84	84	84
Adjusted R^2	0.082	0.162	0.088	0.240
F-test		** 8.815		*** 17.315

Table 2: Regressions of the first and second-order interconnectivity measures on the recession indicator and their lagged values. The time series are monthly. We construct a recession indicator from the NBER recession dates. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	First-order	First-order	Second-order	Second-order
Intercept	*** 0.603	*** 0.646	*** 109.758	*** 127.706
	(6.696)	(7.407)	(6.853)	(8.217)
Lagged interconnectivity	** 0.289	* 0.215	** 0.277	0.123
	(2.745)	(2.072)	(2.653)	(1.176)
Average article sentiment	−0.108	0.143	−38.849	21.955
	(−0.930)	(1.025)	(−1.672)	(0.822)
Recession indicator		** 0.104		*** 26.936
		(2.987)		(3.828)
Number of observations	84	85	85	85
Adjusted R^2	0.080	0.162	0.108	0.237

Table 3: Regressions of the first and second-order interconnectivity measures on average article sentiment, the recession indicator, and lagged values of themselves. The time series are monthly. We construct a recession indicator from the NBER recession dates. We construct our measure of sentiment article by article. For each article, we use the sentiment annotator in the coreNLP toolkit to evaluate the sentiment of each sentence and then take the average across the sentiment of all sentences in an article. For each month, we compute an average article sentiment measure as the average sentiment across all articles in that month. We standardize the monthly average measures using the full same mean and standard deviation. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

Variable	Mean	Std dev.	Min	Max
S&P 500 returns	5.385	19.844	-42.508	52.160
VIX	22.736	10.324	10.820	62.640
Level of yield curve	0.990	1.645	0.010	5.030
Slope of yield curve	1.942	1.023	-0.480	3.430
GDP growth	0.707	0.781	-1.900	1.400
Industrial production growth	0.014	0.865	-4.300	1.400
Consumption growth	0.085	0.281	-0.900	0.900
Inflation	0.136	0.081	-0.079	0.432

Table 4: Summary statistics of our financial and macroeconomic factors. We obtain all data from the St. Louis Fed's FRED database. The S&P 500 returns are month-to-month simple returns. The VIX in a given month is evaluated as the average VIX observed during that month. The level of the yield curve is measured via the 3-month Treasury bill secondary rate, while the slope is constructed as the spread between the fixed-maturity yields of the 10-year and the 1-year Treasury bills. Both level and slope are evaluated at the end of a month. All macroeconomic time series are seasonally adjusted annualized rates. GDP growth is measured from quarter to quarter. Consumption growth is the monthly growth rate of the real per capita consumer expenditure index published by the Bureau of Economic Analysis (account code DPCERX). Inflation is the month-to-month change rate in the personal consumption expenditures index excluding food and energy (BEA account code DPCCRG). Industrial production growth is given by the month-to-month growth rate in the industrial production index of the Board of Governors of the Federal Reserve System. All time series are monthly except for GDP growth. We construct a monthly GDP growth time series from quarterly observations by interpolating with the most recent observation.

Regression	(1)	(2)	(3)	(4)	(5)
Intercept	0.252	2.050	0.252	−2.192	0.951
	(0.067)	(0.619)	(0.067)	(−0.714)	(0.233)
1st-order IC	0.194		0.194	** 10.611	8.254
	(0.045)		(0.044)	(2.717)	(1.969)
2nd-order IC		−0.011	−0.035	0.021	0.023
		(−0.499)	(−0.933)	(0.667)	(0.712)
VIX				*** −0.281	*** −0.303
				(−6.504)	(−4.905)
Article sentiment					** −16.174
					(−2.805)
Recession indicator					−1.522
					(−0.996)
GDP growth					0.593
					(0.671)
Yield curve level					** −0.840
					(−2.684)
Number of observations	85	85	85	85	85
Adjusted R^2	−0.012	−0.009	−0.014	0.326	0.376

Table 5: Regressions of the S&P 500 returns on contemporaneous values of the interconnectivity measures, the VIX, and macro controls. "IC" stands for interconnectivity. Because the first and second-order interconnectivity measures are highly co-linear, we replace the second-order connectivity measure with its residuals after being regressed on the first-order interconnectivity measure whenever both measures are included in a regression. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

Regression	(1)	(2)	(3)	(4)	(5)
Intercept	−8.696	−10.118	−8.696	−8.389	* 12.880
	(−1.082)	(−1.472)	(−1.117)	(−1.321)	(2.112)
1st-order IC	*** 37.065		*** 37.065	*** 37.302	* 17.582
	(3.942)		(4.072)	(5.025)	(2.567)
2nd-order IC		*** 0.217	* 0.202	* 0.159	0.035
		(4.830)	(2.558)	(2.453)	(0.623)
S&P 500 returns				*** −1.221	*** −0.709
				(−6.504)	(−4.168)
Article sentiment					−12.798
					(−1.522)
Recession indicator					−0.354
					(−0.133)
GDP growth					*** −6.718
					(−5.078)
Number of observations	85	85	85	85	85
Adjusted R^2	0.148	0.210	0.201	0.469	0.663

Table 6: Regressions of the VIX level on contemporaneous values of the interconnectivity measures, the S&P 500 returns, and macro controls. "IC" stands for interconnectivity. Because the first and second-order interconnectivity measures are highly co-linear, we replace the second-order connectivity measure with its residuals after being regressed on the first-order interconnectivity measure whenever both measures are included in a regression. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

Regression	(1)	(2)	(3)	(4)	(5)
Intercept	*** 5.979	*** 4.870	*** 5.979	*** 6.001	*** 5.261
	(4.664)	(4.189)	(4.640)	(4.425)	(4.369)
1st-order IC	*** −5.883		*** −5.883	*** −6.882	*** −5.734
	(−3.924)		(−3.904)	(−4.359)	(−4.075)
2nd-order IC		** −0.026	−0.005	−0.013	−0.012
		(−3.373)	(−0.394)	(−0.995)	(−1.025)
Recession indicator				* 0.951	−0.659
				(2.099)	(−1.290)
Inflation				* 4.579	** 5.123
				(2.291)	(2.866)
S&P 500 return					−0.035
					(−1.011)
Article sentiment					*** −9.255
					(−5.195)
Number of observations	85	85	85	85	85
Adjusted R^2	0.146	0.110	0.138	0.202	0.394

Table 7: Regressions of the level of the yield curve on contemporaneous values of the interconnectivity measures, the recession indicator, GDP growth, and the S&P 500 returns. "IC" stands for interconnectivity. Because the first and second-order interconnectivity measures are highly co-linear, we replace the second-order connectivity measure with its residuals after being regressed on the first-order interconnectivity measure whenever both measures are included in a regression. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

Regression	(1)	(2)	(3)	(4)	(5)
Intercept	−1.129	−0.643	−1.129	−1.212	−0.810
	(−1.414)	(−0.898)	(−1.412)	(−1.305)	(−0.852)
1st-order IC	*** 3.621		*** 3.621	*** 3.906	** 3.405
	(3.878)		(3.872)	(3.735)	(3.191)
2nd-order IC		*** 0.017	0.007	0.010	0.009
		(3.651)	(0.878)	(1.088)	(0.988)
Recession indicator				−0.408	0.028
				(−1.093)	(0.068)
GDP growth				−0.101	−0.191
				(−0.517)	(−0.927)
S&P 500 return					0.019
					(0.706)
Article sentiment					* 2.911
					(2.223)
Number of observations	85	85	85	85	85
Adjusted R^2	0.143	0.128	0.141	0.133	0.166

Table 8: Regressions of the slope of the yield curve on contemporaneous values of the interconnectivity measures, the recession indicator, GDP growth, and the S&P 500 returns. "IC" stands for interconnectivity. Because the first and second-order interconnectivity measures are highly co-linear, we replace the second-order connectivity measure with its residuals after being regressed on the first-order interconnectivity measure whenever both measures are included in a regression. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

Regression	Industrial production		Consumption		Inflation	
Intercept	0.012	0.029	0.087	*** 0.807	*** 0.096	** 0.232
	(0.132)	(0.041)	(2.683)	(3.457)	(5.807)	(3.412)
Previous value	** 0.343	0.299	−0.013	−0.059	** 0.303	** 0.282
	(3.305)	(2.731)	(−0.114)	(−0.541)	(2.893)	(2.712)
1st-order IC		−0.018		** −0.845		* −0.157
		(−0.021)		(−3.114)		(−2.055)
2nd-order IC		* −0.018		0.000		−0.000
		(−2.608)		(−0.022)		(−0.194)
Number of observations	84	84	84	84	84	84
Adjusted R^2	0.107	0.156	−0.012	0.075	0.082	0.106
F-test		* 3.413		* 4.849		2.126

Table 9: Predictive regressions for inflation and the growth rates of GDP, industrial production, and consumption based on lagged values of the interconnectivity measures and themselves. "IC" stands for interconnectivity. Because the first and second-order interconnectivity measures are highly co-linear, we replace the second-order connectivity measure with its residuals after being regressed on the first-order interconnectivity measure whenever both measures are included in a regression. All regressions are in monthly frequency except for GDP, for which we run a quarterly regression using the average value of the interconnectivity measure in the previous 3 months. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

Regression	(1)	(2)	(3)	(4)	(5)
Intercept	** −51.600	* −37.354	** −51.358	*** −76.912	* −43.841
	(−2.843)	(−2.284)	(−2.841)	(−3.614)	(−2.630)
1st-order IC	*** 112.410		*** 111.938	* 47.558	
	(5.311)		(2.378)	(2.396)	
2nd-order IC		*** 0.533	0.235		
		(5.027)	(1.274)		
Yield curve level				*** 9.741	*** 9.468
				(4.047)	(3.807)
Yield curve slope				*** 22.861	*** 24.221
				(5.961)	(6.173)
GDP growth				2.272	2.523
				(0.479)	(0.515)
S&P 500 return				0.373	0.470
				(0.607)	(0.741)
VIX				* −1.046	* −1.064
				(−2.311)	(−2.273)
Credit spread				*** 39.817	*** 43.500
				(4.272)	(4.574)
Recession indicator				* −15.774	−14.199
				(−2.128)	(−1.859)
Number of observations	74	74	74	74	74
Adjusted R^2	0.272	0.250	0.278	0.589	0.560
F-test	*** 28.211	*** 25.272	*** 15.039	* 5.653	

Table 10: Regressions of the contagion measure of Figure 13 on contemporaneous values of the interconnectivity measures, GDP growth, the S&P 500 returns, VIX, and the U.S. corporate credit spread. "IC" stands for interconnectivity. Because the first and second-order interconnectivity measures are highly co-linear, we replace the second-order connectivity measure with its residuals after being regressed on the first-order interconnectivity measure whenever both measures are included in a regression. We measure the corporate credit spread as the difference between Moody's Seasoned BAA and AAA corporate bond yields. We obtain data on these corporate bond yields from the St. Louis Fed's FRED database. The values in parentheses give t-statistics. The F-test is for the null hypothesis that a regression that includes the interconnectivity measure fits the data as well as an otherwise equivalent regression that does not control for interconnectivity. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

-- GM-UAW contract seen hard to match fully by rivals
-- By David Bailey
-- Tue Oct 2, 2007 3:47pm EDT
-- http://www.reuters.com/article/2007/10/02/us-gm-uaw-idUSN0242907020071002

DETROIT (Reuters) - Several aspects of the tentative contract between General Motors Corp (GM.N) and the United Auto Workers union will be hard for Ford Motor Co. (F.N) and Chrysler LLC to match in labor talks expected to heat up in coming days, people familiar with the negotiations said.

The adoption of second-tier wages for new hires at GM represents an attractive concession for Ford and Chrysler, but the structure of a retiree health-care trust could prove difficult to transfer, sources familiar with the matters said on Tuesday. The establishment of a Voluntary Employees Beneficiary Association trust, or VEBA, was a centerpiece of the UAW's agreement with GM, allowing the automaker to take some $50 billion of liabilities off its books. Privately held Chrysler has been focused on cash flow since Cerberus acquired the automaker over the summer, to the point that it has been taking daily cash flow reports. The GM-UAW health-care trust would not provide savings until 2010, when the new trust is expected to take over some $3 billion in annual retiree health care payments from the top U.S. automaker. Ford and Chrysler would be hard-pressed to match the bump-up in pension payments to their retirees that GM has agreed to give to its UAW retirees under the tentative contract, people familiar with the talks said. UAW President Ron Gettelfinger said on Friday he expected to assess the state of talks with both Ford and privately held Chrysler after local UAW leaders unanimously recommended that workers approve the GM contract. Gettelfinger wants the agreement with GM to serve as a basic pattern for talks with Ford and Chrysler in keeping with a long-held tradition that has kept all three Detroit-based automakers on a similar labor-cost footing. The union's deal with GM includes a second-tier wage for new hires outside the production line, a health-care trust for retirees and some job security. UAW VOTES ON GM CONTRACT CONTINUE The UAW may not resume full negotiations with Ford or Chrysler until it completes the ratification, or has enough of a favorable indication from the voting at GM locals first, one person close to the talks said. Subcommittees for the UAW and Chrysler had been meeting this week, but there was no indication when full talks would resume, said the person, who asked not to be named because of the private nature of the talks. In the meantime, negotiators at Ford and Chrysler have been poring over the details in the UAW contract with GM. The UAW and GM reached a tentative four-year contract last week to end a two-day national strike -- the first full-scale walkout by the UAW against GM since 1970. The union wants to wrap up the ratification voting by October 10. A majority of the UAW members at GM must approve of the contract for the agreement to be ratified. The more than 73,000 GM hourly workers represented by dozens of UAW locals across the United States have begun voting on the contract. Members of UAW Local 174 near Detroit voted Monday in favor of the contract after a heavy turnout among the 250 to 300 members in one of the first tests of the new contract. A local in Lansing, Michigan, was voting on Tuesday, while other major locals had scheduled informational meetings and votes for later in the week and running into next week. Local 174 members said job security promises, the hiring of temporary workers as permanent employees and a better understanding of the impact of the health-care trust on retirees may have tipped the scale toward ratification. GM gave the UAW job guarantees, made 3,000 temporary workers permanent and promised to insource some jobs done by contractors in addition to the health-care trust. The automaker gave binding commitments to its 16 U.S. assembly plants through the four-year contract, but three do not have binding commitments beyond that and GM expressly excluded two powertrain plants and a service parts operation from a moratorium on plant closings and sales. A local representing workers at an assembly plant in Orion, Michigan, that has no GM commitments beyond 2013 is scheduled to vote on Wednesday on the contract. (Additional reporting by Kevin Krolicki and Poornima Gupta)

(a) Example of a news article in our data.

A tibble: 759 x 8

	id	sid	tid	word	lemma	upos	pos	cid
	<chr>	<int>	<int>	<chr>	<chr>	<chr>	<chr>	<int>
1	doc1	1	1	DETROIT	DETROIT	NOUN	NNP	5
2	doc1	1	2	-LRB-	-lrb-	.	-LRB-	14
3	doc1	1	3	Reuters	Reuters	NOUN	NNP	15
4	doc1	1	4	-RRB-	-rrb-	.	-RRB-	22
5	doc1	1	5	-	-	.	:	24
6	doc1	1	6	Several	several	ADJ	JJ	26
7	doc1	1	7	aspects	aspect	NOUN	NNS	34
8	doc1	1	8	of	of	ADP	IN	42
9	doc1	1	9	the	the	DET	DT	45
10	doc1	1	10	tentative	tentative	ADJ	JJ	49
11	doc1	1	11	contract	contract	NOUN	NN	59
12	doc1	1	12	between	between	ADP	IN	68
13	doc1	1	13	General	General	NOUN	NNP	76
14	doc1	1	14	Motors	Motors	NOUN	NNPS	84
15	doc1	1	15	Corp	Corp	NOUN	NNP	91
16	doc1	1	16	-LRB-	-lrb-	.	-LRB-	96
17	doc1	1	17	GM.N	GM.N	NOUN	NNP	98
18	doc1	1	18	-RRB-	-rrb-	.	-RRB-	103
19	doc1	1	19	and	and	CONJ	CC	105
20	doc1	1	20	the	the	DET	DT	109
21	doc1	1	21	United	United	NOUN	NNP	113
22	doc1	1	22	Auto	Auto	NOUN	NNP	120
23	doc1	1	23	Workers	Workers	NOUN	NNPS	125
24	doc1	1	24	union	union	NOUN	NN	133
25	doc1	1	25	will	will	VERB	MD	139
26	doc1	1	26	be	be	VERB	VB	144
27	doc1	1	27	hard	hard	ADJ	JJ	147
28	doc1	1	28	for	for	ADP	IN	152
29	doc1	1	29	Ford	Ford	NOUN	NNP	156
30	doc1	1	30	Motor	Motor	NOUN	NNP	161
31	doc1	1	31	Co.	Co.	NOUN	NNP	167
32	doc1	1	32	-LRB-	-lrb-	.	-LRB-	171
33	doc1	1	33	F.N	F.N	NOUN	NNP	173
34	doc1	1	34	-RRB-	-rrb-	.	-RRB-	177
35	doc1	1	35	and	and	CONJ	CC	179
36	doc1	1	36	Chrysler	Chrysler	NOUN	NNP	183
37	doc1	1	37	LLC	LLC	NOUN	NNP	192
38	doc1	1	38	to	to	PRT	TO	196
39	doc1	1	39	match	match	VERB	VB	199
40	doc1	1	40	in	in	ADP	IN	205
41	doc1	1	41	labor	labor	NOUN	NN	208
42	doc1	1	42	talks	talk	NOUN	NNS	214
43	doc1	1	43	expected	expect	VERB	VBN	220
44	doc1	1	44	to	to	PRT	TO	229
45	doc1	1	45	heat	heat	VERB	VB	232

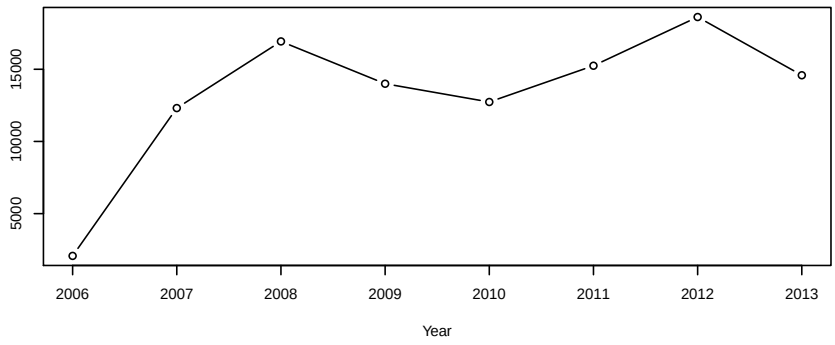
... with 714 more rows

(b) Output of the coreNLP toolkit.

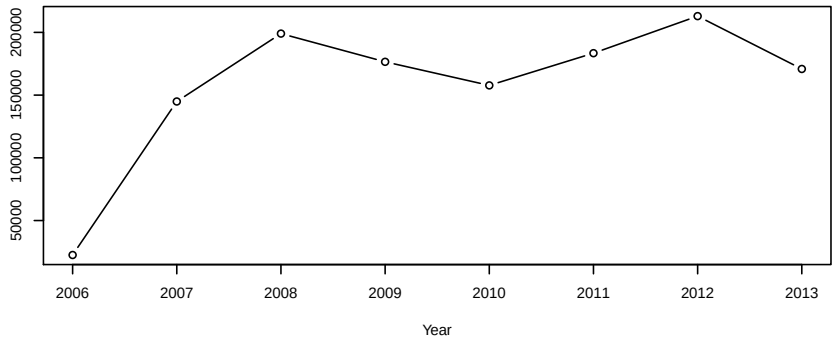
Figure 1: Sample of a news article in our data and the output delivered by the coreNLP toolkit when given the article as input.

```
# A tibble: 4 x 7
  id      sid   tid tid_end entity_type  entity   entity_normalized
  <chr> <int> <int>   <int> <chr>        <chr>    <chr>
1 doc1      1    10      10 ORGANIZATION GM       ""
2 doc1      1    16      16 ORGANIZATION Ford     ""
3 doc1      1    18      18 ORGANIZATION Chrysler ""
4 doc1      1    41      41 DATE         Tuesday  XXXX-WXX-2
```

Figure 2: Sample output of the named entity recognition (NER) algorithm of the coreNLP toolkit.



(a) Number of articles per year.



(b) Number or organizations identified in news data of a given year.

Figure 3: Time series of the number of articles and number of organizations in any given year of our data sample.

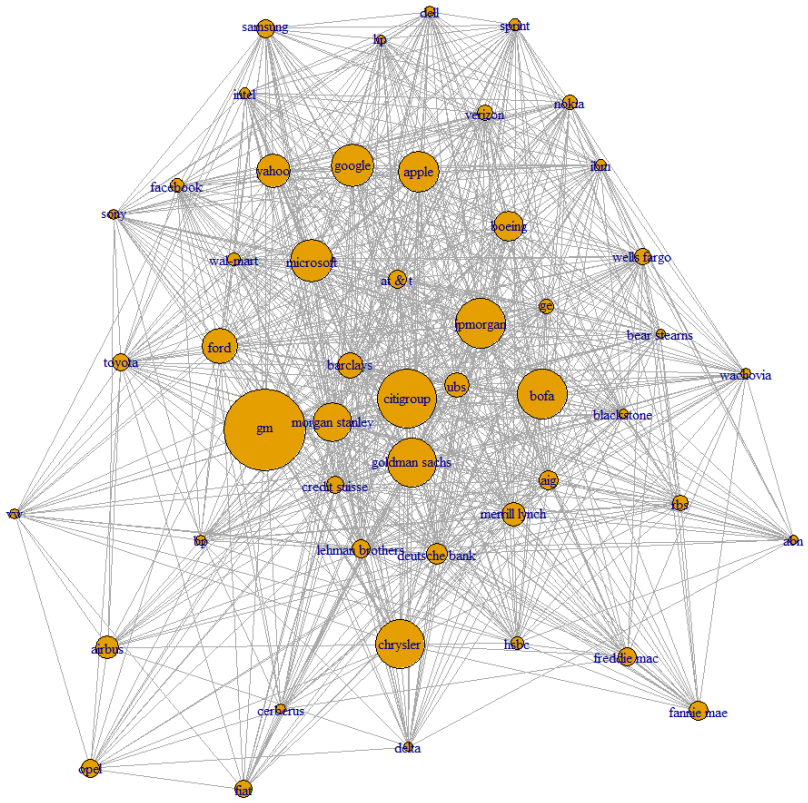


Figure 4: Network of firms implied by the full news data sample covering the years 2006 through 2013. We only plot the largest 50 firm nodes in our network. The size of a node is proportional to the number of times that firm is identified to be connected to another firm in an article in our data. The width of a link between two firms is proportional to the number of times that link is identified in our data.

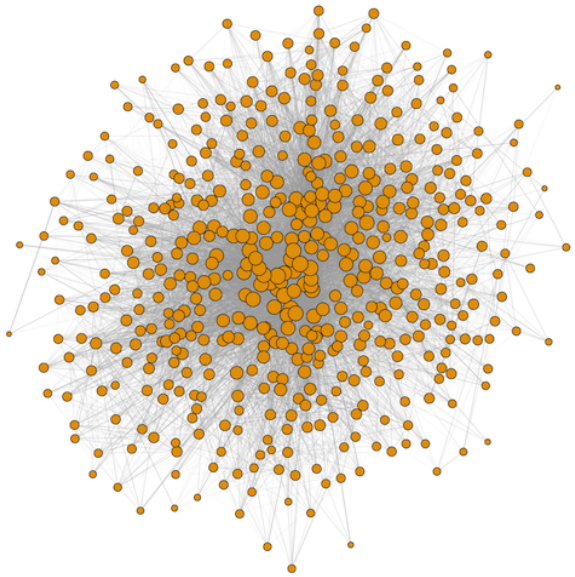


Figure 5: Network of the largest 500 firm nodes implied by our data. The size of a node is proportional to the number of times that firm is identified to be connected to another firm in an article in our data. The width of a link between two firms is proportional to the number of times that link is identified in our data. For clarity, we avoid labeling each node with the corresponding firm name.

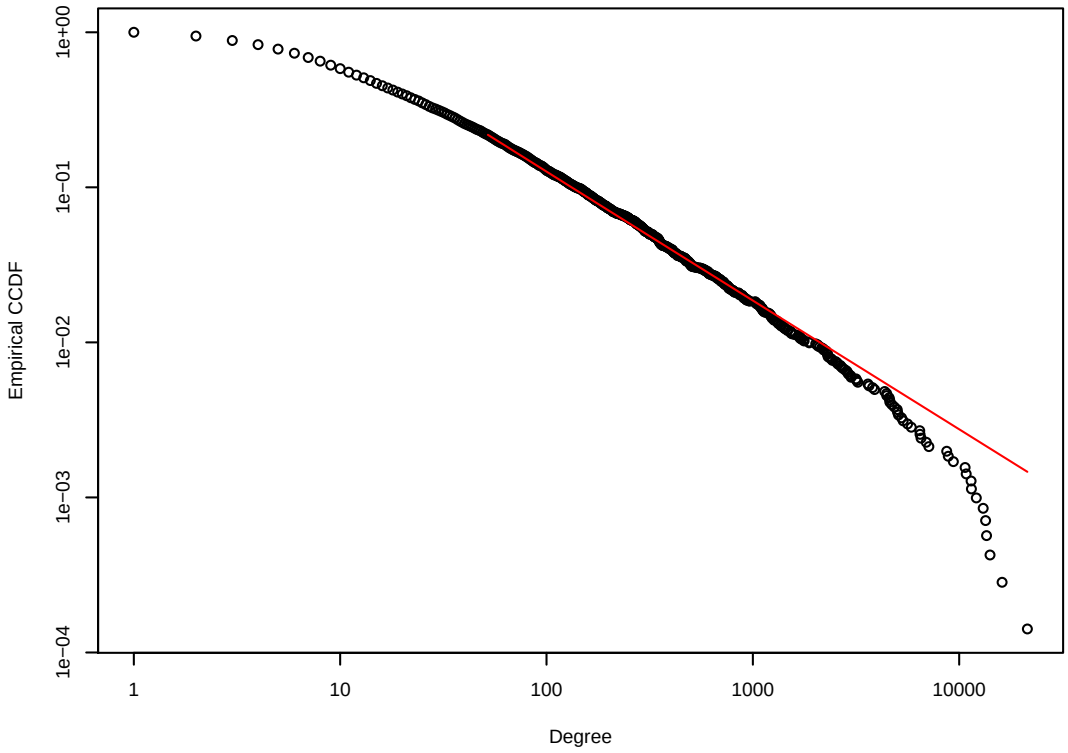
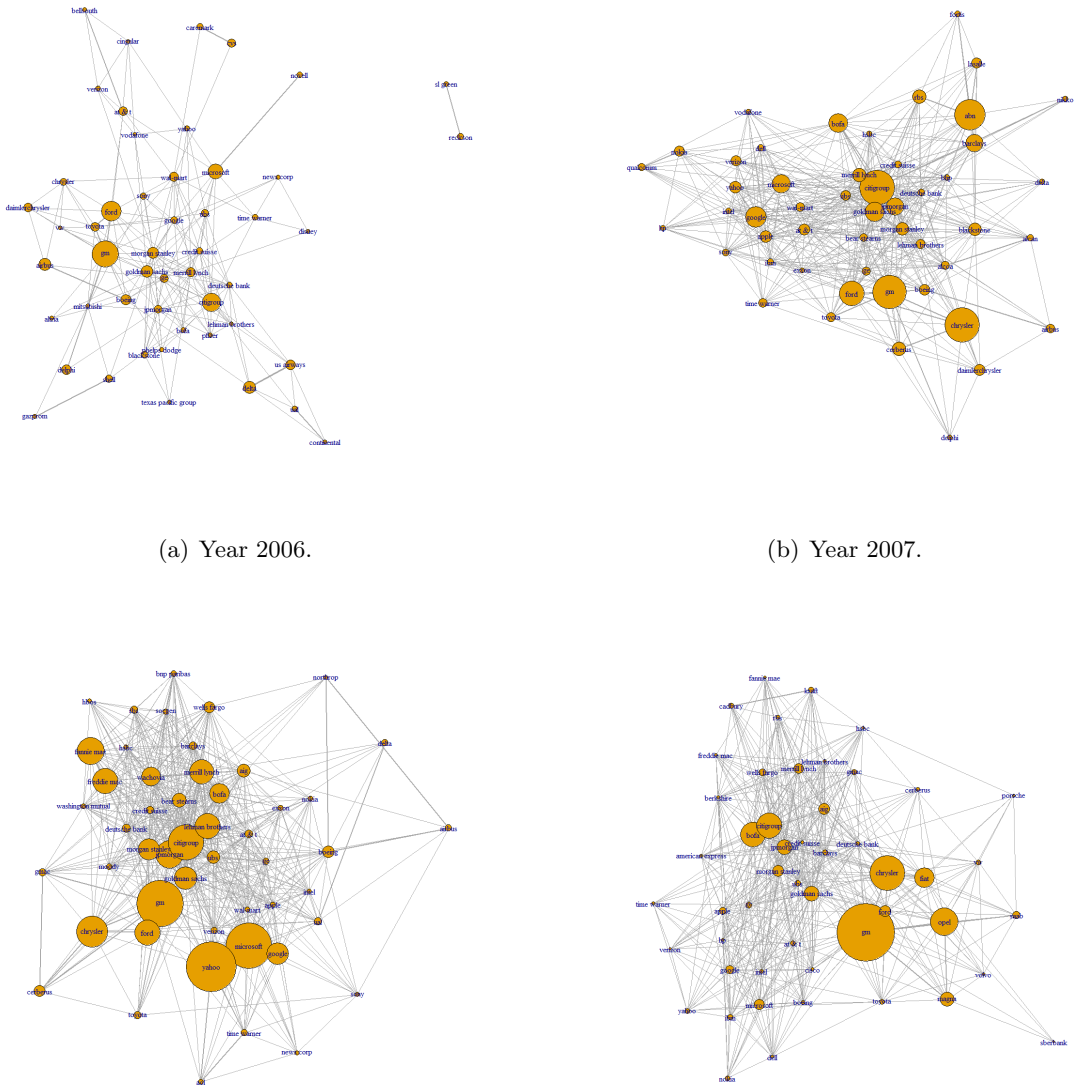


Figure 6: Empirical counter-cumulative distribution of the degree of connectivity of the nodes in our network. The degree of a node in our network is equal to the number of connections that node has. The y-axis shows the probability we can find a node with at least as many nodes as indicated on the x axis. Both axes are represented in log-scale. The red line shows the maximum likelihood fit of a power law with exponent equal to 1.83.



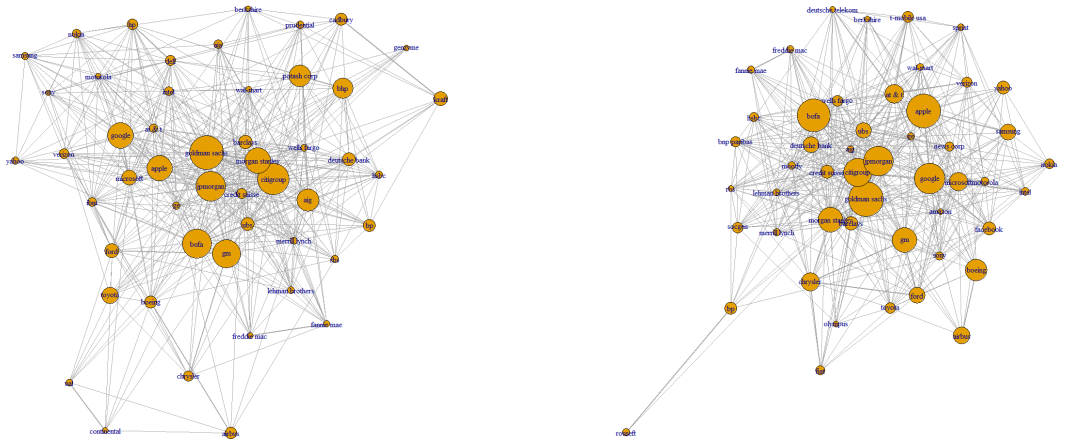
(a) Year 2006.

(b) Year 2007.

(c) Year 2008.

(d) Year 2009.

Figure 7: Time series of news-implied networks in our data sample for the years 2006 through 2009. For any given year, we aggregate the tupels $E = (F_1, F_2, T)$ recognized by the methodology of Section 2 for which T falls within that year. We aggregate these tupels into a matrix of interconnections and plot the resulting connections in R using the package "igraph."



(a) Year 2010.

(b) Year 2011.

(c) Year 2012.

(d) Year 2013.

Figure 8: Time series of news-implied networks in our data sample for the years 2010 through 2013. For any given year, we aggregate the tupels $E = (F_1, F_2, T)$ recognized by the methodology of Section 2 for which T falls within that year. We aggregate these tupels into a matrix of interconnections and plot the resulting connections in R using the package "igraph."

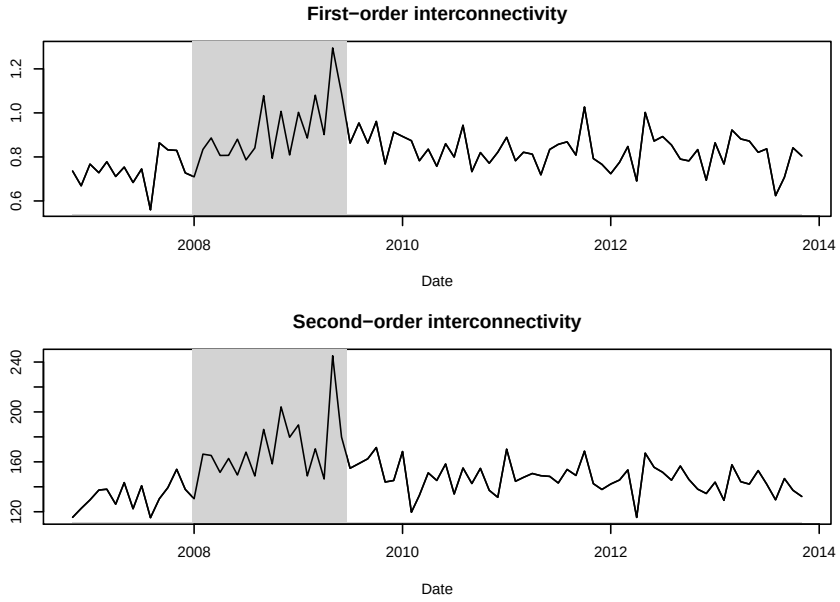


Figure 9: Time series of the first and second-order interconnectivity measures for the networks implied by our news data. For any given month in our data sample, we collect all news article published in that month and extract business connections using the methodology of Section 2. Given the network for each month, we then proceed to evaluate the interconnectivity measures. The first-order interconnectivity measure is given by $\frac{1}{\bar{d}_t}\left(\frac{1}{N-1}\sum_{n=1}^{N_t}(d_t^n-\bar{d}_t)^2\right)^{1/2}$, where N_t is the number of nodes in the network of date t, $d_t^n=\sum_{j=1}^{N} w_t^{j,n}$ is the degree of node n that sums up the number of links n has with other nodes on date t ($w_t^{j,n}=1$ iff nodes j and n are connected on date t), and $\bar{d}_t=\frac{1}{N}\sum_{n=1}^{N} d_t^n$ is the average degree of a node in the network of date t. The second-order interconnectivity measure is given by $\sum_{n=1}^{N}\sum_{j\neq n}\sum_{k\neq j,n} d_t^n w_t^{n,k} w_t^{k,j} d_t^j$. We compute these measures only for the networks containing the largest 100 nodes at a given point of time, where size is measured by the number of links a node has. The grey shaded areas correspond to NBER recession dates.

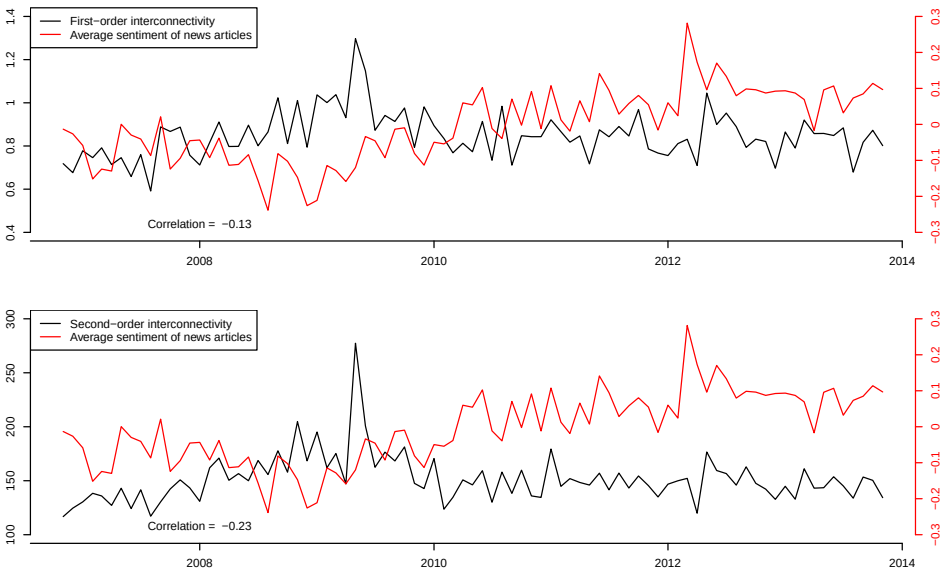


Figure 10: Time series of the interconnectivity measures and the average article sentiment. We construct a recession indicator from the NBER recession dates. We construct our measure of sentiment article by article. For each article, we use the sentiment annotator in the coreNLP toolkit to evaluate the sentiment of each sentence and then take the average across the sentiment of all sentences in an article. For each month, we compute an average article sentiment measure as the average sentiment across all articles in that month. We standardize the monthly average measures using the full same mean and standard deviation.

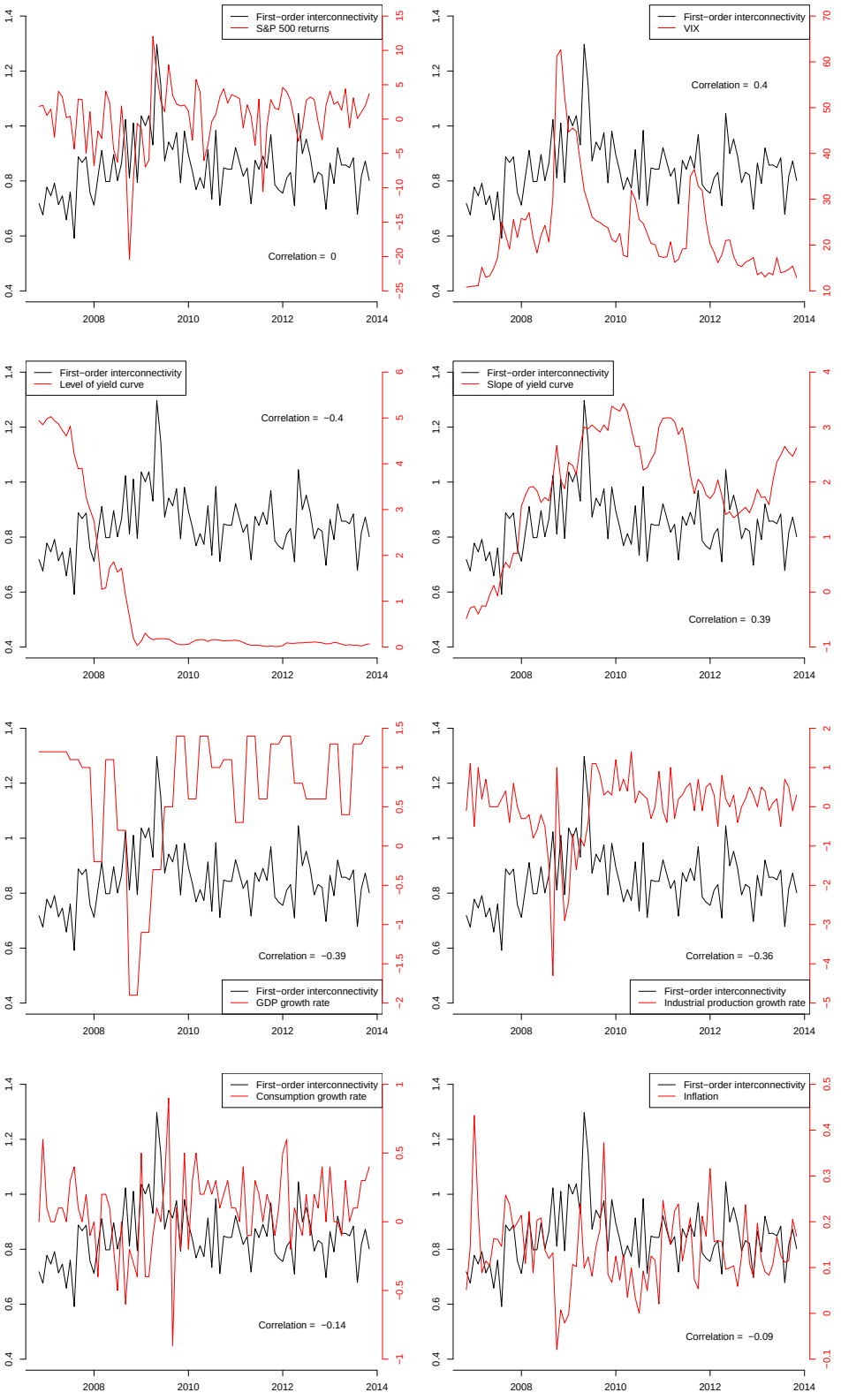


Figure 11: Time series of the first-order interconnectivity measures and financial and macroeconomic factors. See Table 4 for summary statistics of the financial and macro factors.

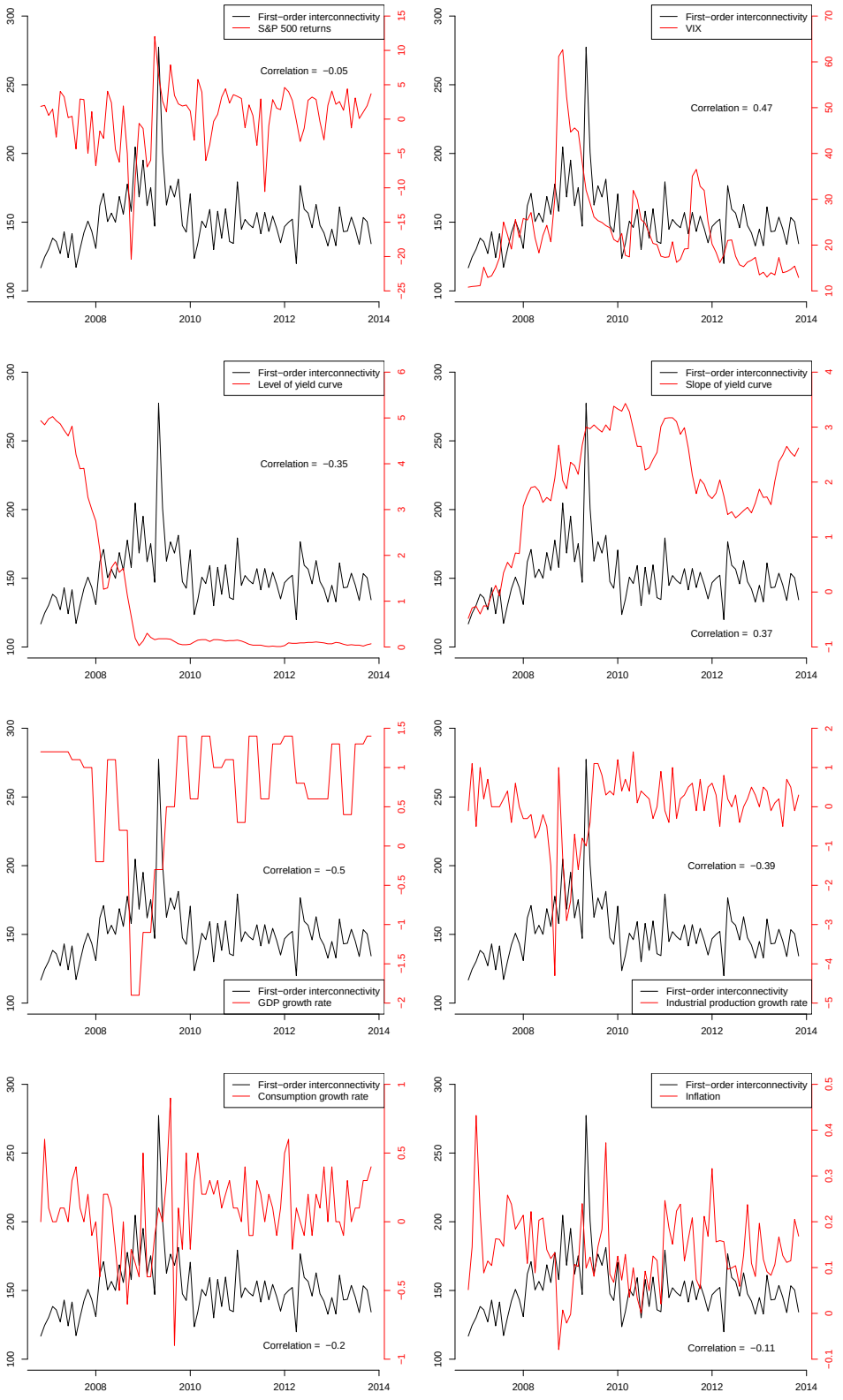


Figure 12: Time series of the second-order interconnectivity measures and financial and macroeconomic factors. See Table 4 for summary statistics of the financial and macro factors.

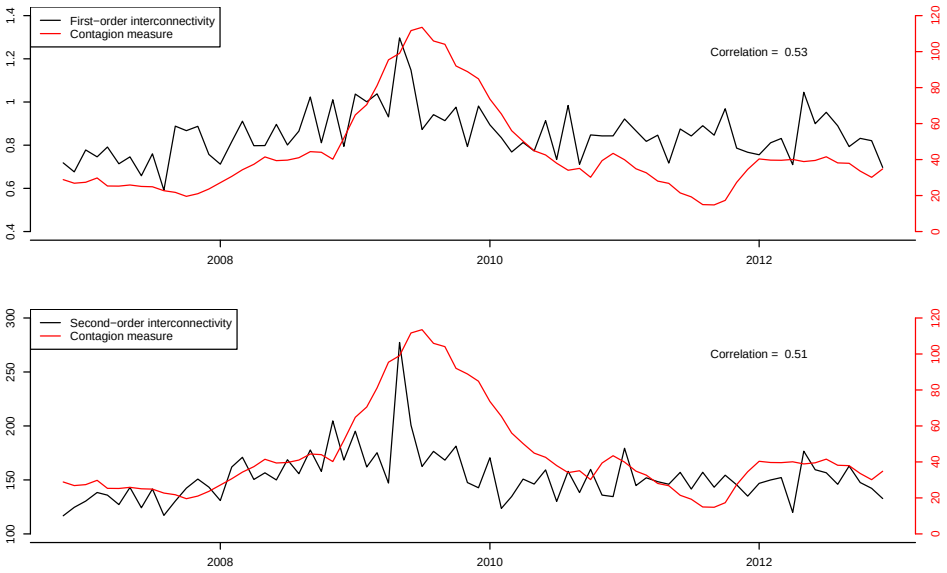


Figure 13: Time series of the interconnectivity measures and a measure of contagion. We obtain daily data on the contagion measure of Azizpour et al. (2018) and aggregate this data to a monthly time series by taking monthly averages. The contagion measure of Azizpour et al. (2018) is only available through December 31, 2012.