The Network of Firms Implied by the News

G. Schwenkler and H. Zheng[*]

March 20, 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 information about business connections between firms that can be used to infer a precise network of firm interconnections. Links in the news-implied network reflect business relations that facilitate the transmission of risks between firms. Interconnectivity in our network is positively related and also predicts out-of-sample common measures of aggregate risks. The results of this paper 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. JEL codes: E32, E44, L11, G10, C82.

[*] 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, Alireza Tahbaz-Salehi, Andrea Vedolin, Kay Giesecke, Steven Kou, Lucy White, Weiting Zheng, and seminar participants at Boston University.

1 Introduction

Firms are connected to each other through different types of business links: costumer-supplier relationships, strategic relationships, subsidiaries, 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 others). In spite of the demonstrated importance of the network of firm interconnections for the measurement of risks, data access is notoriously limited. Often, only incomplete and lagged data are 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. There are some interbank borrowing and lending data for the financial sector but these data are often proprietary and only available to central bankers. There are also data on costumer-supplier connections but those data do not consider the whole universe of firms and the wide spectrum of business relationships. The Bureau of Economic Analysis reports on input-output links across sectors but these data are only updated once every 5 years and do not provide insights on firm-level connections.

In this paper, we show that financial news articles published in common newspapers report about a wide ranging array of business relationships between firms and that this information can be exploited 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 resemble business networks extracted from alternative data sets when restricted to the same set of firms and business relationships. In contrast to the networks extracted from traditional data sources, however, our news-implied networks are available in high frequency and provide more comprehensive information about interconnections in the whole universe of firms. We formally show that the links reported about in the news correspond to business connections that spread risks across firms, such as parent-subsidiary and interbank relationships. We also show that measures of interconnectivity of the news-implied network positively relate to measures of aggregate risks and predict periods of macroeconomic distress out-of-sample. The results of this paper enable the measurement of otherwise intractable interfirm networks. They also enable the estimation of accurate measures of firm-level and aggregate risks.

We posit several hypotheses about the information contained in the news about relationships between firms. The results of Mullainathan and Shleifer (2005) and García (2018) suggest that financial news outlets have incentives to publish news articles about negative shocks that affect firms in order to target investors' fear of downside risk and attract them as readers. Given the extensive evidence that indicates that business relations serve to spread risks across firms (see Boone and Ivanov (2012), Chakrabarty and Zhang (2012), Fernando et al. (2012), Hertzel et al. (2008), and Jorion and Zhang (2009), among others), we posit that news outlets have incentives to publish articles that report about business connections between firms. We also posit that the news is more likely to report about a business relationship between two firms when one of the firms experiences distress that may contaminate an otherwise healthy counterparty. This is our main hypotheses from which we derive further hypotheses about the relationship between news-implied business links and well-known measures of risk. We conjecture that business links in our news-implied network convey information that is different than the information contained in the covariance matrix of firms' stock returns. This is because links in the news-implied network are focused on downside risk, as posited by our main hypothesis, while stock return correlations consider both the upside and the downside. Furthermore, we conjecture that the news-implied firm network becomes more interconnected in periods of aggregate distress because it is precisely in those periods of time when business connections between firms become active transmittors of risks (see Acemoglu et al. (2012, 2017), Azizpour et al. (2018), and Elliott et al. (2014)). Finally, because business connections between firms are often sticky and difficult to unwind when one of the counterparties experiences distress, we also posit that the degree of interconnectivity in the news-implied network predicts measures of aggregate risks as well as periods of macroeconomic distress out-of-sample.

We validate our hypotheses by exploiting novel machine learning tools known as *natural language processing* (NLP), which allow us 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.[1] NLP is commonly 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. The application of NLP for

[1] 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 partisanship on economic research.

sentiment analysis is 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 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 the same sentence. The stronger the relationship is, the more often should the news report about this relationship in different articles.

We apply our NLP methodology to analyze 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 6,000 firms and 16,000 business links. The network implied by our news data showcases a core consisting of large banks that are strongly interconnected and several smaller banks that are connected to the 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 clusters of non-financial firms surrounding the financial firms, delivering a star architecture for the broader network of U.S. firms similar as in Acemoglu et al. (2012). We find that our news-implied network contains a large fraction of the costumer-supplier links reported in the Compustat Segments data in addition to other types of business connections, such as strategic relationships, investment banking, credit, M&A, and competitive relationships. We also observe that an intersectoral network implied by our news data resembles the input-output networks derived from intersectoral data from the Bureau of Economic Analysis (BEA). However, because the BEA data concerns mostly the use and production of commodities, the networks implied by the BEA data are heavily skewed towards manufacturing sectors while our news-implied network highlights the financial and insurance sectors.

Turning to the dynamic evolution of the network over time, we find that some sectors become more or less prominent over time but the financial sector remains central and strongly connected. Our 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 also find that interconnectivity in our news-implied network is mostly orthogonal to the sentiment of the news articles from which we extract our networks, validating our NLP approach.

Formal tests of our hypotheses reveal that, indeed, the news reports about actual business relationships.[2] We find that the news tends to focus on reporting about competitive, strategic, credit, and M&A relationships. It also heavily reports about interbank links. A logit regression shows that the news is more likely to report about a relationship between two firms when one of the firms experiences negative stock market performance or a credit downgrade, even when controlling for market conditions. These findings validate our main hypothesis and show that the news tends to report about distressed business connections between firms.

We develop a bootstrap test to evaluate whether the network implied by the news data is statistically equivalent to the network implied by the correlation matrix of firms' stock returns. Our test rejects this null hypothesis, suggesting that the information conveyed by news-implied business links is statistically different than the information contained in stock return correlations. In a final step, we evaluate the ability of our news-implied network to explain and predict aggregate risks. We find that the degree of interconnectivity in our news-implied network is persistent and peaks during recessionary periods. We also find the VIX, the BAA-AAA credit spread, and the aggregate default rate are high, and industrial production and consumption growth are low, whenever the news-implied network is highly interconnected. These results confirm that interconnectivity in our news-implied network is positively related to measures of aggregate financial and economic risks. In addition, the yield curve level tends to be low and the slope tends to be large during periods of high interconnectivity. We interpret from these results that the risks reflected in our news-implied networks are expected to be short-lived by market participants. Finally, we find that news-implied interconnectivity predicts the NBER recession indicator as well as the growth rates of industrial production and consumption at the monthly horizon, even when controlling for common explanatory variables.[3] Our results empirically demonstrate that business relationships, such as those reported in the news, 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 paper most closely related to ours is Scherbina and Schlusche (2015). Like us, Scherbina and Schlusche (2015) study news articles to infer business relations. They also show that the

[2] Scherbina and Schlusche (2015) establish a similar result using an alternative approach.

[3] Our findings on the predictability of consumption growth complement recent results by Liu and Matthies (2018), who argue that consumption growth can be predicted with news media content.

news contain information about business relations between firms. However, there are important differences in our focuses, approaches, and results. Scherbina and Schlusche (2015) focus on the diffusion of information in stock markets. They show that news shocks that affect one firm are slowly reflected in the stock prices of the counterparties of that firm. In contrast, we focus on the informational content of the news and show that the news tends to report about distressed business links between firms. Scherbina and Schlusche (2015) do not carry out natural language processing. They obtain news data from Thomson Reuters and exploit the fact that Thomson Reuters tags firm names and topics mentioned in news articles in their data base. Scherbina and Schlusche (2015) utilize the tags provided by Thomson Reuters to identify firms and business relations. We, on the other hand, develop our own NLP methodology to identify firms and business relations from the raw text data. Considering that news data is freely available online and that several of the NLP toolkits we use are also freely available in standard coding progams, such as R, Matlab, and Python, our results are easily replicable by a broad public without requiring costly subscriptions to data providers such as Thomson Reuters. In this regard, our results contribute to the democratization of financial data.

The rest of this paper is organized as follows. Section 2 lays out the hypotheses that we will test in our analysis. Section 3 introduces our data and methodology. Section 4 describes the estimated networks. Section 5 presents our empirical results and Section 6 concludes.

2 Hypotheses

The news have incentives to maximize reader turnout. It is well known that investors are more concerned about downside risk than upside potential; see Kahneman and Tversky (1979), Kuhnen (2015), and others. Because of this, financial news outlets have incentives to publish articles about negative events that may encapsulate risks for investors in order to attract investors as readers. Indeed, García (2018) empirically shows that a negative market return triggers more negative news reporting than a positive market return of equivalent magnitude and that this is primarily driven by reader demand considerations. Mullainathan and Shleifer (2005) show that news outlets tend to fine-tune their reporting towards what readers are interested in reading.

Why would the news report about links between firms? We conjecture that the news is inclined to report about firm connections when these connections serve to spread risk from a distressed to an otherwise healthy firm in order to attract concerned investors as readers. Suppose that Firm A and B share a business link, say, Firm A is supplier to Firm B. In good times, when both Firms A and B are performing well, the relationship between Firm A and B would not be newsworthy because it does not speak to investors' fear of downside risk. In bad times, however, when either Firm A or Firm B is experiencing financial distress, the link between Firms A and B becomes newsworthy because this link can spread risk from the distressed to the healthy firm. That business links between firms serve to spread risk is well established in the literature. There is extensive evidence that costumer-supplier relationships can spread risks across firms that result in equity value losses (Cohen and Frazzini (2008) and Barrot and Sauvagnat (2016)) and increased credit risk (Jorion and Zhang (2009)). Risk spillovers are not limited to costumer-supplier relationships. Banking, financing, competitive relationships, and strategic partnerships can also spread risks across firms as documented in Azizpour et al. (2018), Boone and Ivanov (2012), Fernando et al. (2012), Jorion and Zhang (2007), and Lang and Stulz (1992), among others. Given the evidence in the literature, we conjecture that news outlets are inclined to report about the relationships between firms and that the news tend to report about firm relationships in which one of the firms is experiencing distress. We test the following hypotheses.

Hypothesis 1: The news contain information about relationships between firms.

Hypothesis 2: The news are more likely to report about a link between two firms when one of the two firms is experiencing financial distress.

If the news indeed reports about distressed links between firms, we would expect that the connections reported in the news convey information that is different than the information contained in financial asset correlations.[4] This is because correlations reflect relationships between assets both in the upside and downside, while Hypothesis 2 conjectures that news-implied relationships tend to focus on the downside. We therefore test the following hypothesis.

Hypothesis 3: Relations in the news-implied network capture links between firms that are different from correlations in equity markets.

Hypotheses 1 through 3 focus on firm-level connections. We can obtain a network of firm connections after aggregating these firm-level links. Several theoretical papers argue that the architecture of the firm network is a key driver of aggregate risks. Gabaix (2011) shows that idiosyncratic shocks to large firms can amplify to large aggregate shocks because large firms tend to be highly connected with other firms. These connections facilitate the transmission of

[4] See Diebold and Yılmaz (2014, 2016) and Demirer et al. (2018) for methodologies to extract a network of firm connections from asset return correlations.

local shocks across firms. Acemoglu et al. (2012) show that aggregate volatility is high in a disaggregated economy that is highly interconnected.[5] Acemoglu et al. (2017) extend these findings by showing that tail risk also tends to be large in disaggregated and highly interconnected economies. Based on this evidence and our previous hypotheses, we conjecture that when we aggregate the news-implied firm links to a news-implied firm network, interconnectivity measures in this network should closely relate to aggregate measures of risk.

Hypothesis 4: Measures of interconnectivity in the news-implied network are positively related to measures of aggregate risks.

We extend our approach by also evaluating the predictive power of the information contained in news articles. Business relationships between firms are highly sticky because it is often costly to renegotiate these relationships; see Berger and Udell (1995), Gulati (1995), Joskow (1987), Mayer and Argyres (2004), and Petersen and Rajan (1994), among many others. Furthermore, firms that experience financial distress tend to remain in a distressed state for a prolonged period of time (Bris et al. (2006), Wruck (1990)). If the news indeed reports about interfirm relationships that transmit distress across firms and given that interfirm relationships are sticky, we conjecture that interconnectivity in the news-implied network predicts periods of macroeconomic distress out of sample. We test the following hypothesis.

Hypothesis 5: Measures of interconnectivity in the news-implied network of firms predict adverse macro-economic outcomes.

3 Data & methodology

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 and Panel (a) of Figure 1 provides a sample news article 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 only contain a few sentences sentence. Panel (b) of Figure 1 shows that the number of articles published each year is fairly constant, although we have a much shorter sample for the year 2006.

[5] Here, *"disaggregated"* means that there are several clusters or sectors of firms in the economy while *"interconnected"* means that different clusters share links with each other.

3.1 Identification

We analyze each news article in our data to identify whether an article reports about a relationship between two firms. Intuitively, if the news is reporting about two firms that share some sort of business relation, then these two firms should be mentioned in one article within close proximity from each other. Based on this insight, we identify a business relationship whenever two firms are mentioned in the same sentence of an article.

3.2 Methodology

We require a methodology that can identify firms mentioned in each sentence of a news article such as the one in Figure 1. This is not a trivial task. One could use a static list of firm names but, given the dynamic nature of firm birth and failure, a static firm name list may miss some firms. Furthermore, firm names are often abbreviated or replaced with pseudonyms in the news. For example, General Electric Company is often just called GE, Ford Motor Company is often just referred to as Ford, and JPMorgan Chase often goes by JPMorgan, J. P. Morgan, or J. P. Morgan Chase. Keeping track of all possible abbreviations or pseudonyms is computationally costly. Finally, the use of alternative firm identifiers, such as tickers, also presents a series of challenges. Tickers are not always mentioned in news articles. Even when they are, tickers change periodically and this restricts the usefulness of a static list of tickers.

We develop a two-step machine learning methodology to address the challenges with identifying firm names in text data. We summarize the methodology here and provide details in Appendix A. The first step consists of using a natural language processing (NLP) toolkit to identify all nouns mentioned in a news article that could potentially be firm names.[6] For this step we use the Stanford *coreNLP* toolkit available in R (see Manning et al. (2005)).[7] The coreNLP toolkit identifies in text data nouns that refer to entities and classifies these into different categories: named entities (“PERSON”, “LOCATION”, “ORGANIZATION”, “MISC”), numerical entities (“MONEY”, “NUMBER”, “ORDINAL”, “PERCENT”), and temporal en-

[6] Natural language processing (NLP) is a branch of machine learning that focuses on processing and analyzing text data. NLP tools can be used to identify different parts of speech in text data; say, labeling words as verbs, nouns, adjectives, and so on. Gentzkow et al. (2019) provide a detailed overview of how NLP tools are currently being used for financial and economic research.

[7] The Stanford coreNLP toolkit is one of the most popular NLP resources used by 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.

tities ("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."* Figure 2 shows the output of the coreNLP algorithms applied to this sentence. The coreNLP algorithms recognize the following entities in the sentence: (GM, ORGANIZATION), (Ford, ORGANIZATION), (Chrysler, ORGANIZATION), and (Tuesday, DATE). Even though coreNLP does not recognize United Auto Workers union as an entity, it performs well at recognizing all three corporations mentioned in the sentence. The coreNLP toolkit has been demonstrated to be highly accurate in identifying named entities, with accuracy rates in the order of 80% (see Costa et al. (2017) and Dlugolinsky et al. (2013)).

In the second step, we take all the entities classified by the coreNLP toolkit as organizations and run an algorithm developed by us to determine which of these organizations are indeed corporations. Our algorithm works as follows (details can be found in Appendix A). We first remove all organizations whose names contain words that signal government agencies or nonprofit institutions, such as the words "agency", "cooperation", "federal", "foundation", or "university". For the remaining organizations, we remove from their names all numbers, special symbols, adoptions, determiners, adverbs, and unreasonable postfixes. We also remove all words that indicate business types (like "Co.," "Inc.," and "Ltd."). We assume that every organization that survives these steps is a firm. Still, there may be instances in which one firm goes by several names. We run additional steps to determine a unique name for each firm. We begin by creating clusters of firms with common words in their names and consider the most frequently mentioned name in a cluster as the name stem. Consider the following example. Suppose there is a cluster consisting of 6 firms that go by the names "Toyota," "Toyota USA," "Toyota Motor," "Toyota Motor Credit," "Toyota Motor," and "Toyota Motor". In this cluster, the most common name is "Toyota Motor" so we designate "Toyota Motor" as the name stem for the cluster. Then, for each one of the firms in the cluster we check whether the name of the firm is fully contained in the stem or viceversa. If so, we update the stem to be either the name of the firm or the prevalent stem, whichever is shorter. If not, we remove the firm from the original cluster. We proceed iteratively until no more improvements of the name stem can be made. All firms that remain in the cluster are considered to be the same firm and we assign the name stem as the name of this firm. In our example, we would iterate through the firms named "Toyota," "Toyota

USA," and "Toyota Motor Credit." Given that "Toyota" is the shortest name fully contained in the original stem, we would update "Toyota" to be the new firm name 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.

Step 1 (coreNLP) and Step 2 (firm identification algorithm) introduced above deliver a list of firm mentions in our news data. When running steps 1 and 2, we also keep track of the article in which a firm is mentioned, the sentence within an article where the firm mention was found, and the publishing date of the article. We then establish that two firms share a connection whenever the firms are identified in the same sentence of an article.

3.3 Output of methodology

Our methodology finds 15,618 firm mentions in our data. Figure 3 shows the number of recognized firm mentions in each year. Except for the year 2006 for which we have a shorter data sample, we see that our algorithm recognizes around 2,000 firm mentions in any given year. We also see that the number of firms recognized in any given year is fairly constant. Of course, not every mention corresponds to a different firm. In total, our algorithm identifies 6,440 different firms during the time span covered by the data. The five most frequently mentioned firms are General Motors, Citigroup, Chrysler, Bank of America, and JPMorgan Chase.

Table 1 indicates that an average article mentions a large number of firms (about 3.40). It also shows that our methodology recognizes a large number of business connections: On average, we identify 2.90 firm connections per article with a standard deviation of 6.16 connections per article. Over the whole data sample, we identify 308,512 links between firms.

4 Estimated news-implied networks

4.1 Full data sample

We plot in Figure 4 the network of firms implied by all of the news articles in our data sample. Each node represents a firm in our data. The size of a node is proportional to the number of times that firm is found in the data while the width of a link is proportional to the number of times that link is identified in the data. For the sake of clarity, in Figure 4 we only show the largest 50 nodes that correspond to the most frequently mentioned firms in the sample.

We observe several interesting features. We first see that the big banks – Citigroup, Goldman Sachs, JPMorgan Chase, Bank of America, and Morgan Stanley – 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, Lehman Brothers, Credit Suisse, Barclays, Merrill Lynch, Wells Fargo, RBS, and ABN Amro. 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 because they share a link with Goldman Sachs. To the extent that the links identified from the news data indeed represent risky business connections, the network in Figure 4 provide empirical support for the financial accelerator model of Bernanke et al. (1999).

We also observe in Figure 4 that there are several large nodes 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. An alternative data source for business relationships between non-financial firms is the Compustat Segments data, which reports about costumer-supplier links. In Panel (a) of Figure 5, we plot the network implied by the Compustat Segments data covering the same time span as our data. We focus on the largest 50 nodes in the Compustat Segments data as measured by sales. We also plot in Panel (b) of Figure 5 the network implied by our news-based approach with the restriction that we only include the same 50 firms showcased in the Compustat Segments network. To facilitate the comparison across networks, we color in red any link in the Compustat Segments network that is missing in our news-implied network. We also color in green any link in our news-implied network that is also available in the Compustat Segments network.

We see that our news-implied network includes about a third of the links in the Compustat Segments data. A large chunk of the links that are missing concern the costumer-supplier network surrounding the large pharmaceutical firms Bristol-Myers Squibb (ticker "BMY") and Pfizer (ticker "PFE") as well as the pharmaceutical distribution firms AmerisourceBergen (ticker "ABC") and McKesson Corporation (ticker "MCK"). By the nature of the pharmaceutical industry, the costumer-supplier links between these firms may be associated with large sales and are therefore prominent in the Compustat Segments data. However, if Hypothesis 2 is valid, then these links may not be newsworthy because they do not transfer significant risks and, as a result, are not showcased in our news data. In spite of these shortcomings, Figure 5 shows that our news-implied network includes many more links that go beyond the costumer-supplier relationships included in the Compustat Segments data. As we show in Section 5.1, the news often report about competitive, investment banking, credit, and M&A relationships. Our approach highlights the rich set of diverse business connections between firms.

Figure 4 also exhibits several sector-based clusters. On the bottom left corner, we find a cluster of firms associated with transportation sectors. In the bottom right corner there is a technology cluster. Above it we find a communications cluster. The top part of Figure 4 is dominated by financial firms. 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. For a full comparison, we obtain from the merged CRSP / Compustat database data on the NAICS sector codes for firms in our network.[8] Figure 6 shows our news-implied network aggregated by two-digit NAICS codes. We also display in Figure 6 the networks implied by the 2012 BEA industry-by-industry total requirement table.

We see that the news-implied intersectoral network in Panel (a) of Figure 6 exhibits a similar star structure as highlighted in Acemoglu et al. (2012) and also showcased in the BEA

[8] We lose about two-third of the firms in our sample when we restrict ourselves to firms that have a match in the merged CRSP / Compustat data set. We match with the firm name across data sets.

input and output networks in Panels (b) and (c) of Figure 6. Similar as in the BEA input network, the most prominent sector in our intersectoral network is the manufacturing sector (NAICS code "33"). This sector includes computer, electrical, furniture, machinery, metal, and transportation manufacturing firms which heavily dominate the production of final goods in the U.S. economy. In contrast, the BEA output network highlights sectors that produce raw goods, such as agriculture (NAICS code "11") and mining (NAICS code "21"), which are often used as inputs in other sectors. Because the BEA data mostly measures the use and production of commodities, the BEA input-output networks diminish the importance of the insurance and financial sectors (NAICS codes "51" and "52", respectively). In contrast, those sectors are highly central and prominent in our news-implied network, in consistence with the theoretical models of Bernanke et al. (1999) and Carvalho and Gabaix (2013) that put the financial industry at the center of the U.S. economy. Intersectoral input-output linkages are mostly determined by costumer-supplier relationships between firms. Costumer-supplier relationships are only a subset of the interfirm relationships we are able to identify in the data (see Figure 5 and Section 5.1). Because of this, the news-implied intersectoral network in Panel (a) of Figure 6 is denser than the equivalent networks implied by the BEA input-output data.

BEA intersectoral input-output linkages are only updated every 5 years. In contrast, our news-implied network is available in high frequencies because the news is published online in a continuous fashion. Our approach can therefore provide more immediate and granular information about intersectoral linkages and how they affect the macroeconomy. These kinds of insights are inaccessible from the Bureau of Economic Analysis data.

To wrap up our analysis of the full network, we study the distribution of the degree of the firms in our networks.[9] Figure 7 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).

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

4.2 Time series of networks

We plot yearly time series of networks implied by news articles in our data sample in Figures 8 and 9.[10] For each year between 2006 and 2013, we use the methodology of Section 3 to extract all business connections implied by news articles published in that year. For clarity, we only plot the connections between the largest 50 firms in every year. We do this just for yearly times series for the sake of simplicity. However, similar plots can be constructed for arbitrary frequencies – as frequently as daily or hourly and as infrequently as quarterly or annually.

The time series of news-implied networks yields 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 the news-implied networks in 2006 and 2007 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. The automobile sector became overly dominant in the news-implied network for the year 2009, consistent with the prevailing crisis in that sector. After 2010 when the great recession ended, the news-implied 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.[11] 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.[12] 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.[13] 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 introduced by Acemoglu

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

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

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

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

et al. (2012). 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. 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 10.[14] We see that there is significant time variation in the interconnectivity measures. The time series of interconnectivity measures appear to be persistent. The regression results in Table 2 confirm these visual insights by showing that the monthly AR(1)-coefficients of 0.302 for first-order interconnectivity and 0.311 for second-order interconnectivity are significantly large.

We evaluate the relationship between interconnectivity and sentiment in our data. Figure 11 plots the time series of the average article sentiment in our sample. If Hypothesis 2 is correct and the news tends to report about risky business links between firms, then one may be concerned that the interconnectivity measures for our news-implied networks are just picking up on negative sentiment in the news articles. 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 controlling for the persistence of the interconnectivity measures. Table 2 summarizes the results of our regressions. We find that interconnectivity is negatively related to sentiment but this relationship is not statistically strong. We therefore reject the notion that spikes in our interconnectivity measures are only driven by sentiment. These results 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.

5 Hypothesis tests

We run several tests to assess the validity of the hypotheses posited in Section 2. In these test, we will often use data on several financial and macroeconomic factors. Figures 12 and 13 plot the time series of our interconnectivity measures in conjunction with financial and macroeconomic factors that we include in our analysis. Table 3 provides summary statistics of our macro and financial factors.

[14] For this, we compute analogous networks as those in Figures 8–9 but on a monthly basis and then compute the resulting interconnectivity measures.

5.1 Hypothesis 1

The first hypothesis states that the news contains information about relationships between firms. To validate this hypothesis, we begin by showcasing several sample sentences for the 10 strongest links in the full-data network of Figure 4 (i.e., the 10 most frequently identified relationships in the data). The 10 strongest links are (in decreasing order of strength): (GM, Chrysler), (Microsoft, Yahoo), (Fannie Mae, Freddie Mac), (Boeing, Airbus), (GM, Ford), (GM, Opel), (Chrysler, Fiat), (Apple, Samsung), and (Goldman Sachs, Morgan Stanley). Table 4 shows sample sentences in our news data for these links. We see that several of the sentences point towards competitive relations. These competitive relationships can be strategic partnerships (such as the joint operations of Fannie Mae and Freddie Mac) or destructive (as in the case of Apple fighting Samsung). We also see that some sentences point to joint investment banking solutions provided by big banks, such as in the context of Goldman and Morgan Stanley sponsoring the Alibaba IPO. There are some sentences that point to parent-subsidiary relationships, like Opel belonging to the GM group. And there other sentences that point to M&A activity such as Fiat acquiring Chrysler. The sample sentences of Table 4 hint that the news contains information about different types of business relationships between firms.

We carry out a formal textual analysis of the sentences in which we identify relationships between firms to assess whether links in our network correspond to known business relationships. For each of the strongest 100 links in Figure 4, we collect all sentences in which the link was identified and extract the 10 most frequently mentioned nouns in those sentences. We then apply a clustering algorithm that clusters the different links according how similar their most frequently mentioned nouns are.[15] Our clustering algorithm finds that it is optimal to cluster the top 100 links into 14 different clusters. Table 5 shows the 3 most frequent nouns in each cluster, together with the 3 most frequent verbs, the number of allocated links, and a representative link for each cluster. It also shows the relationship that we interpret to be associated with the cluster based on the most frequent nouns and verbs. We see that several of the clusters are associated with competitive relationships between firms in the same industry (automotive, banking, technology, etc.). There are clear interbank links, be it because of banks' common participation in credit markets or because they partner up for investment banking purposes. We also find among the

[15] We use a k-mediod clustering algorithm implemented in R under the function "pamk" of the package "fpc". This algorithm selects an optimal number of clusters via subsampling (CLARA method; see Kaufman and Rousseeuw (1990, Ch. 3)).

top 100 links several links that are associated with M&A activity, such as the merger of Daimler and Chrysler and the acquisition of Cadbury by Kraft. Finally, we find links that are associated with strategic partnerships between firms, such as the partnership between BP and Rosneft during the sample period.

All in one, the results of this analysis show that the news contains information about actual business relationships between firms. They validate Hypothesis 1.

5.2 Hypothesis 2

Hypothesis 2 posits that the news are more likely to report about a business link when one of the two business partners experiences distress. We begin our evaluation of Hypothesis 2 with a visual inspection of the degree of connectivity of individual firms in our news-implied network. Figure 14 plots the time series of the degree of connectivity for Citigroup, Ford, and Microsoft.[16] We also mark in the plots events experienced by these firms that were frequently reported in the news. Consistent with Hypothesis 2, we see that the connectivity measures for these firms tend to spike up when a firm experiences adverse events. For example, we see that the connectivity measure of Ford spikes up around the automobile crisis when Ford refused to be bailed out by the government. We also see that the connectivity measure of Citigroup peaked right before the financial crisis when it announced large write-downs related to subprime mortgages. In addition, we also see that the connectivity measure of a firm may spike up around times of M&A activity. For example, the connectivity measure for Microsoft spiked up when Microsoft was considering acquiring Yahoo and when Microsoft and Nokia announced an alliance.

We proceed to formally test Hypothesis 2. For this, we create monthly time series of link-level dummy variables that indicate whether in a given month we identified a link between two firms. We then run a logit regression of the link dummy variables on the monthly return of the two firms, their credit ratings, the change in their credit ratings relative to the previous month, and aggregate market and macro controls.[17] Table 6 summarizes our findings.

Consistent with Hypothesis 2, we find that it is more likely to observe a link between two firms when one of the firms experiences negative monthly stock returns or a credit downgrade.

[16] Connectivity is proportional to the number of firms a firm is connected with in the network of the 200 most connected firms in our data.

[17] We obtain returns and credit ratings data from CRSP / Compustat. We restrict the analysis to links between the 50 firms that are most frequently identified by our algorithms and for which we also have a match in the merged CRSP / Compustat database.

These result hold even when controlling for aggregate market conditions and time and link fixed effects. The results of Table 6 validate Hypothesis 2. The news is more likely to report about a link between two firms when one of the firms experiences financial distress.

5.3 Hypothesis 3

We test Hypothesis 3 that posits that links in our news-implied network convey different information than stock return correlations. To test this hypothesis, we match firms in our data sample with the CRSP / Compustat merged data set to obtain ticker and stock return information. We use our news-based methodology to determine the 50 most connected firms among the firms that have a match in CRSP / Compustat and evaluate the correlation matrix for those 50 firms. Panel (a) of Figure 15 plots the network implied by the resulting correlation matrix. In this figure, all nodes are of the same size and the width of the link is proportional to the absolute value of the correlation coefficient (we truncate correlations of absolute value less than 0.1 for simplicity). For comparison, we also show in Panel (b) of Figure 15 the network implied by our news data for the 50 most connected firm for which we also have stock return data. Visually, we see that our news-implied network is sparser than the network implied by the correlation matrix. Even though both the news-implied and the correlation-based network have a star structure as in Acemoglu et al. (2012), the correlation-based network appears to have a more densely populated core.

We develop a bootstrap approach to test the null hypothesis that the links in the news-implied and the correlation-based networks are equivalent. We focus on the product-moment correlation between the two networks as the test statistic. The product-moment correlation of two networks is the correlation coefficient between the link widths in both networks.[18] Under the null hypothesis of equivalent links, the product-moment correlation between both networks should take on a value of close to 1. However, for the two plots in Figure 15, the product-moment correlation is 0.031. To assess whether the measured product-moment correlation is significantly different than 1, we generate 1,000 bootstrap samples of the correlation-based network and, for each bootstrap sample, we evaluate the product-moment correlation with the correlation-based

[18] Formally, if $A^n = [a^n_{i,j}]_{i,j=1,\dots,N}$ for $n \in \{1,2\}$ is the adjacency matrix of network n, where N is the number of nodes and $a^n_{i,j}$ corresponds to the width of link (i,j) in network n, then the product moment correlation is given by $\frac{Cov(1,2)}{\sqrt{Cov(1,1)Cov(2,2)}}$. Here, $Cov(n,m) = \frac{1}{N^2}\sum_{i,j=1}^{N}(a^n_{i,j} - \mu^n)(a^m_{i,j} - \mu^m)1_{\{i \neq j\}}$ for μ^n equal to the average link width in network n. We use the function "gcor" in the R package "sna" to compute the product-moment correlation.

network in Panel (a) of Figure 15. Here, we assume that the stock return correlations used to construct the network in Panel (a) of Figure 15 were measured with error, where the error is normally distributed around the measured correlation with a standard deviation equal to the standard error of the correlation estimate. We determine an empirical p-value for the measured product-moment correlation between the two plots in Figure 15 using the bootstrap samples. Figure 16 summarizes the results of this analysis and shows that we reject the null hypothesis of equivalent links.

Putting everything together, the results of this section validate Hypothesis 3. They show that the informational content of our news-implied network is different than the information contained in the correlation matrix of stock returns.

5.4 Hypothesis 4

Hypothesis 4 conjectures that interconnectivity in our news-implied network is positively related to measures of aggregate risks. We run several regressions to test this hypothesis. We begin by regressing measures of financial risks on interconnectivity and other controls. Table 7 shows that the first and second-order interconnectivity measure are positively related to the VIX, a forward looking risk measure for the aggregate equity market. This holds even when controlling for the leverage effect, sentiment, and the state of the macroeconomy. We also regress measures of corporate credit risk on interconnectivity. Table 8 summarizes the results of regressions of the BAA-AAA corporate bond yield spread and the aggregate default rate in the U.S. on our interconnectivity measures and several controls. We find that both the corporate yield spread and the aggregate default rate are highly correlated with the first-order interconnectivity measure of our news-implied networks. The results of Tables 7 and 8 imply that interconnectivity is positively associated with measures of aggregate financial risks.

We further evaluate the relation between interconnectivity and financial risks by studying the government bond market. In Table 9, we regress the level and slope of the Treasury yield curve on our measures of interconnectivity and several controls. We find a negative relationship between interconnectivity and yield curve level and a positive relationship between interconnectivity and yield curve slope. These results indicate that periods of high interconnectivity are characterized by expensive Treasury bonds and high long-term yields. Together with the results of Tables 7 and 8, our findings suggest that the interfirm links reported in the news reflect equity and corporate bond risks that lead investors to flea to the Treasury bonds. The fact that the yield curve tends to be positively steep in periods of high interconnectivity suggests that the financial risks reflected in our news-implied network are expected to be short-lived by market participants.

Next, we evaluate the relationship between interconnectivity and measures of macroeconomic risks. We begin by regressing our interconnectivity measure on a recession indicator and lagged values of themselves. Table 10 summarizes our findings. We find that 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. These finding suggest that interconnectivity is high in recessionary periods.

We extend our analysis by assessing the relationship between interconnectivity and industrial production and consumption growth. Table 11 reports the results of a regression of industrial production and consumption growth on interconnectivity, sentiment, and lagged values of themselves. We find that there is a strong negative relationship between industrial production and first-order interconnectivity. There is only a weak negative relationship between consumption growth and second-order interconnectivity. Together with the regression results of Table 10, these findings confirm that interconnectivity in our news-implied network spikes up during periods of macroeconomic distress.

All in one, the results of this section validate Hypothesis 4: Interconnectivity is positively related to measures of aggregate risks. Our results provide model-free empirical evidence in support of the theoretical models of Acemoglu et al. (2012, 2017) that highlight that aggregate risks are high in periods in which disaggregated economic networks become highly interconnected.

5.5 Hypothesis 5

Finally, we test Hypothesis 5 that states that interconnectivity predicts periods of macroeconomic distress. We begin by running predictive regressions for industrial production and consumption growth based on lagged values of themselves and of our interconnectivity measures. Table 12 summarizes our findings. We see that the interconnectivity measures significantly predict industrial production and consumption growth at the monthly frequency even after controlling for their lagged realizations.

We also run predictive probit regressions for the NBER recession indicator. Table 13 shows that our interconnectivity measures positively predict the recession indicator at the monthly time horizon. This holds both when controlling for lagged values of the recession indicator, which are generally not available at the time of prediction because the NBER announces recession periods ex-post, and for lagged values of the industrial production and consumption growth rates that are always available. Overall, the results of Table 12 and 13 validate Hypothesis 5: Measures of interconnectivity in our news-implied network predict adverse macroeconomic outcomes.

6 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. We show that links in the news-implied firm network correspond to business relationships that spread risks across firms. We also find that interconnectivity in our news-implied network is positively related to measures of aggregate risks and predicts periods of macroeconomic distress out-of-sample. The results of our paper enable the development of measures of risks that accurately reflect the network interconnections of firms.

A Methodology

We lay out the steps that go into our methodology to take news articles as input and output a network of firm connections implied by the news.

A.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 17 illustrates the output delivered by coreNLP. It extracts words from the news article and it assigns a universal part-of-speech tag (*"upos"*).[19] 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 the token within the article (*"tid"*). We refer to Arnold (2017) for a detailed overview of the coreNLP package and its output.

A.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 Panel (b) of Figure 1 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

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

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.[20] 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" 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 all organizations whose names contain flags for government agencies or non-for-profit businesses such as "federal", "ministry", "association", "commission", "senate", "parliament", "parliamentary", "congress", "congressional", "cooperation", "university", "foundation", "republicans", "democrats", "council", "league", "policy", "institute", "embassy", "agency", "federation", "airport", "charity", "charities", "institute", "court", and "school". Next, we 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

[20] 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.

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). We do not believe that this occurs too often in our data set given that the results of Section 5.1 indicate that we identify several parent-subsidary relationships. 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). We manually adjust the most mentioned abbreviations like GM, GE, and others.

A.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 A.2. While our rule to identify connected firms may seem restrictive at first, the results of our paper show that our approach delivers reasonable networks of business connections that closely relate to macro and financial factors.

A.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.

B Data

References

Acemoglu, Daron, Asuman Ozdaglar and Alireza Tahbaz-Salehi (2017), 'Microeconomic origins of macroeconomic tail risks', *American Economic Review* **107**(1), 54–108.

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.

Barrot, Jean-Noël and Julien Sauvagnat (2016), ' Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks', *The Quarterly Journal of Economics* **131**(3), 1543–1592.

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.

Berger, Allen N. and Gregory F. Udell (1995), 'Relationship lending and lines of credit in small firm finance', *The Journal of Business* **68**(3), 351–381.

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.

Boone, Audra L. and Vladimir I. Ivanov (2012), 'Bankruptcy spillover effects on strategic alliance partners', *Journal of Financial Economics* **103**(3), 551–569.

Bris, Arturo, Ivo Welch and Ning Zhu (2006), 'The costs of bankruptcy: Chapter 7 liquidation versus chapter 11 reorganization', *The Journal of Finance* **61**(3), 1253–1303.

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.

Chakrabarty, Bidisha and Gaiyan Zhang (2012), 'Credit contagion channels: market microstructure evidence from Lehman Brothers' bankruptcy', *Financial Management* **41**(2), 320–343.

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.

Cohen, Lauren and Andrea Frazzini (2008), 'Economic links and predictable returns', *The Journal of Finance* **63**(4), 1977–2011.

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.

Diebold, Francis X. and Kamil Yılmaz (2016), 'Trans-atlantic equity volatility connectedness: U.s. and european financial institutions, 2004–2014', *Journal of Financial Econometrics* **14**(1), 81–127.

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.

García, Diego (2018), The kinks of financial journalism. Working Paper, CU Boulder.

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.

Gulati, Ranjay (1995), 'Does familiarity breed trust? the implications of repeated ties for contractual choice in alliances', *The Academy of Management Journal* **38**(1), 85–112.

Hertzel, Michael G., Zhi Li, Micah S. Officer and Kimberly J. Rodgers (2008), 'Inter-firm linkages and the wealth effects of financial distress along the supply chain', *Journal of Financial Economics* **87**(2), 374–387.

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 (2007), 'Good and bad credit contagion: evidence from credit default swaps', *Journal of Financial Economics* **84**(3), 860–883.

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

Joskow, Paul L (1987), 'Contract Duration and Relationship-Specific Investments: Empirical Evidence from Coal Markets', *American Economic Review* **77**(1), 168–185.

Kahneman, Daniel and Amos Tversky (1979), 'Prospect theory: An analysis of decision under risk', *Econometrica* **47**(2), 263–291.

Kaufman, Leonard and Peter J. Rousseeuw (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, Series in Probability and Statistics, Wiley.

Kuhnen, Camelia M. (2015), 'Asymmetric learning from financial information', *The Journal of Finance* **70**(5), 2029–2062.

Lang, Larry and Rene Stulz (1992), 'Contagion and competitive intra-industry effects of bankruptcy announcements', *Journal of Financial Economics* **32**, 45–60.

Liu, Yukun and Ben Matthies (2018), Long run risk: Is it there? Working Paper.

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.

Mayer, Kyle J. and Nicholas S. Argyres (2004), 'Learning to contract: Evidence from the personal computer industry', *Organization Science* **15**(4), 394–410.

Mullainathan, S. and A. Shleifer (2005), 'The market for news', *American Economic Review* **95**(1), 1031–1053.

Petersen, Mitchell A. and Raghuram G. Rajan (1994), 'The benefits of lending relationships: Evidence from small business data', *The Journal of Finance* **49**(1), 3–37.

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.

Scherbina, Anna and Bernd Schlusche (2015), Economic linkages inferred from news stories and the predictability of stock returns. Working Paper.

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.

Wruck, Karen Hopper (1990), ‘Financial distress, reorganization, and organizational efficiency’, *Journal of Financial Economics* **27**(2), 419 – 444.

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 firms per article	3.41	3.22	0	41
Number of connections identified in an article	2.90	6.16	0	305

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 “US”.

	First-order	First-order	Second-order	Second-order
Intercept	*** 0.593	*** 0.603	*** 104.757	*** 109.758
	(6.638)	(6.696)	(6.586)	(6.853)
Lagged interconnectivity	** 0.302	** 0.289	** 0.311	*** 0.277
	(2.893)	(2.745)	(3.007)	(2.653)
Average article sentiment		−0.108		−38.849
		(−0.930)		(−1.672)
Number of observations	84	84	84	84
Adjusted R^2	0.082	0.080	0.088	0.108

Table 2: Regressions of the first and second-order interconnectivity measures on lagged values of themselves and the average article sentiment. The time series are monthly. 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
VIX	22.736	10.324	10.820	62.640
Moody's BAA-AAA yield spread	1.279	0.588	0.750	3.380
Aggregate default rate	0.109	0.103	0.000	0.433
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 3: Summary statistics of our financial and macroeconomic factors. All factors are sampled at the monthly frequency. The VIX in a given month is evaluated as the average VIX observed during that month. The Moody's BAA-AAA yield spread corresponds to the difference between the yields of BAA and AAA rated corporate bonds as rated by Moody's. 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. We obtain a monthly GDP growth rate time series by interpolating with the most recent quarterly observation. 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). 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. Inflation is the month-to-month change rate in the personal consumption expenditures index excluding food and energy (BEA account code DPCCRG). Data on the above factors are obtained from the St. Louis Fed's FRED database. We compute a nonparametric measure of the aggregate default rate in the U.S. economy as the fraction of days in a month with at least one default observation. We use the same historical default timing data as in Azizpour et al. (2018), which is obtained from Moody's Default Risk Service and covers the years 1970 through 2012.

Link	Representative sentence
(GM, Chrysler)	*"Larger rival General Motors Corp. reported 6 percent sales growth, while Chrysler Group posted a 3 percent rise, breaking a nine-month losing streak."*
(Microsoft, Yahoo)	*"Microsoft Corp's $44.6 billion bid to acquire Yahoo Inc marked a coup for the companies' advisers."*
(Fannie Mae, Freddie Mac)	*"Freddie Mac chief executive Richard Syron said he was wary of calls for Freddie Mac and fellow mortgage finance company Fannie Mae to buy loans and securities no longer favored by private investors."*
(Boeing, Airbus)	*"In 2007, Boeing's top rival Airbus, a unit of EADS, named Saudi Prince Alwaleed bin Talal as the first private buyer of an A380 superjumbo."*
(GM, Ford)	*"Chrysler's larger rivals Ford Motor Co and General Motors Co are both exploring ways to soften their respective pension risks, which analysts say have taken a toll on their stock price."*
(GM, Opel)	*"Germany's Opel is part of the GM group."*
(Chrysler, Fiat)	*"Less than a month after it filed Chapter 11, Chrysler is seeking approval to sell its stronger operations to a "New Chrysler" owned by Fiat, labor unions and the U.S. and Canadian governments."*
(Apple, Samsung)	*"Many commentators see Apple's fight against Samsung as a proxy for fighting Android itself."*
(Goldman Sachs, Morgan Stanley)	*"Alibaba's IPO is being sponsored by Goldman Sachs and Morgan Stanley, with Rothschild an adviser."*

Table 4: Sample sentences in which our algorithms recognize relationships between two firms.

Cluster	Sample link	# links	Nouns	Verbs	Implied rel.
1	(GM, Chrysler)	15	government, sale, automaker	tell, give, hold	Competitors
2	(Google, Microsoft)	6	advertising, business, internet	see, compete, fall	Competitors
3	(Fannie Mae, Freddie Mac)	12	bank, loan, mortgage	fall, report, rise	Interbank / credit
4	(Chrysler, Daimler)	6	firm, stake, automaker	hold, own, remain	Parent / subsidiary
5	(Ford, Toyota)	6	automaker, car, sale	fall, rise, see	Competitors
6	(Apple, Samsung)	8	patent, phone, maker	use, compete, lose	Competitors
7	(Goldman Sachs, Morgan Stanley)	14	bank, investment, source	advise, lead, decline	Partners
8	(Kraft, Cadbury)	9	offer, bid, holder	agree, raise, offer	Acquisition
9	(Verizon, Vodafone)	5	analyst, carrier, customer	pay, add, agree	M&A
10	(JPMorgan, Bear Stearns)	4	bank, investment, credit	agree, announce, come	Interbank
11	(United, Continental)	4	parent, airline, bankruptcy	file, merge, plan	Merger
12	(BP, Rosneft)	2	bp, oil, barrel	agree, become, block	Partners
13	(UBS, Credit Suisse)	6	bank, investment, asset	advise, cut, report	Advisory
14	(IBM, HP)	3	computer, analyst, business	fall, rise, compare	Competitors

Table 5: Cluster analysis for the most frequently mentioned nouns in sentences in which we identify links in our news data. For each of 100 most frequently identified links in our data, we determine the 10 most frequently mentioned nouns (we exclude proper nouns) among all sentences in which that links is identified. We then cluster the top 100 links according to the similarities between their most frequently mentioned nouns. We use the function “pamk” of the R package “fpc” for this purpose. The optimal number of clusters is 14 and it is determined by the algorithm using a CLARA subsampling method (see Kaufman and Rousseeuw (1990, Ch. 3))). For each of the identified clusters, we report a sample link in the cluster, the number of links allocated to that cluster, and the three most frequently mentioned nouns and verbs in the sentences allocated to the cluster. The column “Implied rel.” states the relation that we interpret to be associated with the cluster based on its most common nouns and verbs.

	(1)	(2)	(3)	(4)
Intercept	*** −1.954	*** −1.840	*** −1.967	0.161
	(−33.355)	(−26.724)	(−17.577)	(0.150)
Return of Firm A	· −0.001	−0.001	** −0.002	0.000
	(−1.932)	(−1.632)	(−2.924)	(0.003)
Return of Firm B	** −0.002	** −0.001	*** −0.002	0.000
	(−3.042)	(−2.636)	(−3.589)	(−0.133)
Rating of Firm A	*** 0.027	*** 0.028	*** 0.027	*** −0.035
	(14.237)	(14.444)	(14.154)	(−3.617)
Rating of Firm B	*** 0.018	*** 0.019	*** 0.019	−0.009
	(9.316)	(9.665)	(9.266)	(−0.947)
Rating change of Firm A	*** −0.104	*** −0.100	*** −0.116	−0.016
	(−4.713)	(−4.422)	(−5.143)	(−0.496)
Rating change of Firm B	** −0.065	** −0.059	** −0.063	* −0.058
	(−3.250)	(−2.922)	(−3.287)	(−2.201)
S&P 500 return		** −0.006		*** −0.018
		(−2.637)		(−5.971)
VIX		** −0.004		*** −0.007
		(−3.270)		(−3.802)
GDP growth		* −0.032		· −0.036
		(−2.128)		(−1.698)
Ind. prod. growth		−0.004		−0.011
		(−0.415)		(−0.804)
Yield curve level		*** −0.028		*** −0.064
		(−6.184)		(−7.654)
Time fixed effect	N	N	Y	N
Link fixed effect	N	N	N	Y
Number of observations	47028	47028	47028	47028

Table 6: Regressions of link dummy variables on returns, credit ratings, and credit rating changes for the two linked firms, and other controls. The link dummy variable for the link between firms A and B indicates whether in a given month we identified at least one link between A and B among news articles published in that month. We obtain monthly total returns and S&P credit ratings for long maturity debt from the merged CRSP / Compustat database. We rank the ratings according to credit quality in descending order (26 through 1 for rating ranging between AAA and D). We then compute rating changes as the different between the numerical values of the ratings in consecutive months. The yield curve level is the yield of the 3-Month Treasury Bill on the secondary market. Data on S&P 500 returns, the VIX, GDP growth, industrial production growth ('Ind. prod. growth") and the yield curve level is obtained from the St. Louis Fed FRED database; Table 3 provides summary statistics. The values in parentheses give t-statistics. ***, **, *, and · denote significance on the 99.9%, 99%, 95%, and 90% confidence levels, respectively.

	(1)	(2)	(3)	(4)	(5)
Intercept	−8.696	−10.118	−8.696	−8.389	* 12.936
	(−1.082)	(−1.472)	(−1.117)	(−1.321)	(2.140)
1st-order IC	*** 37.065		*** 37.065	*** 37.302	* 17.366
	(3.942)		(4.072)	(5.025)	(2.626)
2nd-order IC		*** 0.217	* 0.202	* 0.159	0.033
		(4.830)	(2.558)	(2.453)	(0.612)
S&P 500 returns				*** −1.221	*** −0.706
				(−6.504)	(−4.207)
GDP growth					*** −6.642
					(−5.610)
Article sentiment					−12.282
					(−1.657)
Number of observations	85	85	85	85	85
Adjusted R^2	0.148	0.210	0.201	0.469	0.667

Table 7: 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. VIX data is obtained from the St. Louis Fed FRED database; Table 3 provides summary statistics. The values in parentheses give *t*-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	BAA-AAA yield spread				Aggregate default rate			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Intercept	* −0.900	** −0.982	* −0.900	*** 1.431	** −0.262	*** −0.265	*** −0.262	−0.105
	(−2.060)	(−2.690)	(−2.177)	(4.131)	(−3.391)	(−4.039)	(−3.528)	(−1.157)
1st-order IC	*** 2.569		*** 2.569	* 0.850	*** 0.437		*** 0.437	** 0.275
	(5.030)		(5.315)	(2.428)	(4.844)		(5.041)	(2.999)
2nd-order IC		*** 0.217	** 0.014	0.003		*** 0.002	** 0.002	0.001
		(4.830)	(3.263)	(1.229)		(5.763)	(2.806)	(1.496)
Yield curve level				*** −0.177				−0.002
				(−3.808)				(−0.131)
Yield curve slope				** −0.209				0.002
				(−3.229)				(0.133)
S&P 500 return				0.006				* 0.006
				(0.750)				(2.602)
GDP growth				*** −0.432				* −0.037
				(−6.879)				(−2.238)
Article sentiment				*** −1.940				** −0.343
				(−4.070)				(−2.749)
Observations	85	85	85	85	85	85	85	85
Adjusted R^2	0.224	0.312	0.305	0.753	0.211	0.277	0.271	0.447

Table 8: Regressions of the BAA-AAA yield spread and the aggregate default rate on contemporaneous values of the interconnectivity measures and other 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. We obtain yields for BAA and AAA-rated corporate bonds by Moody’s from the St. Louis Fed FRED database. The aggregate default rate is the fraction of days of a month with at least one default observation as recorded in the historical default timing data collected by by Moody’s Default Risk Service (see Azizpour et al. (2018)). Table 3 provides summary statistics for these factors. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	Yield curve level				Yield curve slope			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Intercept	*** 5.979	*** 4.870	*** 5.979	** 4.371	−1.129	−0.643	−1.129	−0.631
	(4.664)	(4.189)	(4.640)	(4.369)	(−1.414)	(−0.898)	(−1.412)	(−0.671)
1st-order IC	*** −5.883		*** −5.883	*** −5.155	*** 3.621		*** 3.621	** 3.472
	(−3.924)		(−3.904)	(−4.075)	(3.878)		(3.872)	(3.398)
2nd-order IC		** −0.026	−0.005	−0.010		*** 0.017	0.007	0.009
		(−3.373)	(−0.394)	(−1.025)		(3.651)	(0.878)	(1.097)
GDP growth				* 0.568				−0.122
				(−1.290)				(−0.646)
Inflation				* 4.123				−2.056
				(2.866)				(−1.547)
S&P 500 return				−0.053				0.023
				(−1.011)				(0.878)
Article sentiment				*** −9.345				* 2.758
				(−5.195)				(2.404)
Number of observations	85	85	85	85	85	85	85	85
Adjusted R^2	0.146	0.110	0.138	0.394	0.143	0.128	0.141	0.190

Table 9: Regressions of the level and slope of the Treasury yield curve on contemporaneous values of the interconnectivity measures and several 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. We obtain yield curve, inflation, S&P 500 returns, and GDP growth data from the St. Louis Fed FRED data base. Summary statistics can be found in Table 3. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	First-order	Second-order
Intercept	*** 0.646	*** 127.24
	(7.404)	(8.210)
Lagged interconnectivity	* 0.220	0.130
	(2.119)	(1.252)
Recession indicator	*** 0.083	*** 23.500
	(2.969)	(4.161)
Number of observations	84	84
Adjusted R^2	0.162	0.240
F-test	** 8.815	*** 17.315

Table 10: 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.

	Industrial production growth				Consumption growth			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Intercept	*** 2.379	*** 2.282	*** 2.379	** 2.033	0.389	* 0.466	0.389	· 0.420
	(3.471)	(3.795)	(3.504)	(4.369)	(1.647)	(2.245)	(1.656)	(1.723)
1st-order IC	*** −2.789		*** −2.789	** −2.364	−0.359		−0.359	−0.377
	(−3.479)		(−3.512)	(−4.075)	(−1.300)		(−1.306)	(−1.332)
2nd-order IC		*** −0.015	−0.011	−0.006		· −0.003	−0.003	−0.003
		(−3.812)	(−1.607)	(−1.025)		(−1.857)	(−1.356)	(−1.196)
Lagged value				· 0.188				−0.134
				(−1.290)				(−1.131)
Article sentiment				* 2.100				· 0.562
				(−5.195)				(1.701)
Number of observations	85	85	85	84	85	85	85	84
Adjusted R^2	0.117	0.139	0.133	0.244	0.008	0.028	0.018	0.037

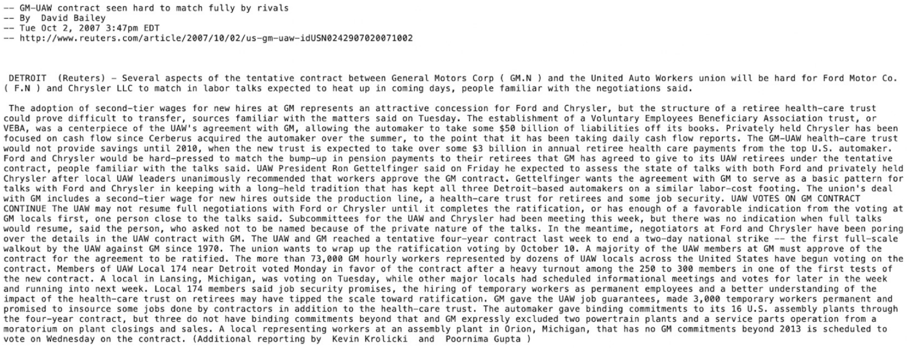
Table 11: Regressions of the first and second-order interconnectivity measures on the recession indicator and their lagged values. The time series are monthly. We construct a recession indicator from the NBER recession dates. The values in parentheses give t-statistics. ***, **, *, and · denote significance on the 99.9%, 99%, 95%, and 90% confidence levels, respectively.

	Industrial production		Consumption	
	(1)	(2)	(1)	(2)
Intercept	0.012	0.029	0.087	*** 0.807
	(0.132)	(0.041)	(2.683)	(3.457)
Lagged value	** 0.343	0.299	−0.013	−0.059
	(3.305)	(2.731)	(−0.114)	(−0.541)
Lagged 1st-order IC		−0.018		** −0.845
		(−0.021)		(−3.114)
Lagged 2nd-order IC		* −0.018		0.000
		(−2.608)		(−0.022)
Number of observations	84	84	84	84
Adjusted R^2	0.107	0.156	−0.012	0.075
F-test		* 3.413		* 4.849

Table 12: Predictive regressions for the growth rates of industrial production and consumption based on lagged values of themselves and the interconnectivity measures. “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 sampling horizon is monthly. The values in parentheses give t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	NBER recession indicator		
	(1)	(2)	(3)
Intercept	*** −2.166	* 0.275	0.013
	(−5.503)	(2.034)	(0.045)
Lagged value	*** 3.759	*** 0.950	
	(6.045)	(20.553)	
Lagged 1st-order IC		· −0.311	0.264
		(−1.918)	(0.817)
Lagged 2nd-order IC		0.001	** 0.008
		(0.762)	(2.925)
Lagged ind. prod. growth			*** −0.233
			(−5.203)
Lagged consumption growth			· −0.255
			(−1.958)
Number of observations	84	84	84

Table 13: Predictive probit regressions for the NBER recession indicator based on lagged values of itself, the interconnectivity measures, and other 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 sampling horizon is monthly. The values in parentheses give t-statistics. ***, **, *, and · denote significance on the 99.9%, 99%, 95%, and 90% confidence levels, respectively.

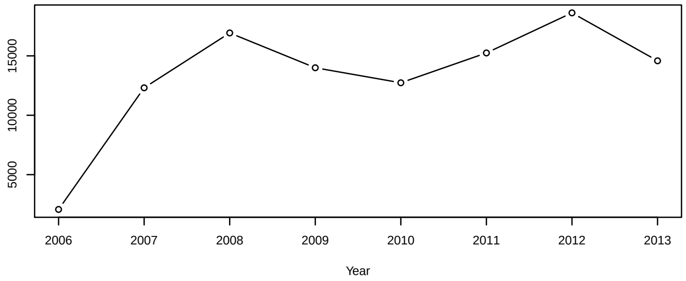


-- GM-UAW contract seen hard to match fully by rivals
-- By David Bailey
-- Tue Oct 2, 2007 3:47pm EDT
-- http://www.reuters.com/article/2007/10/02/us-gm-uaw-idUSN0242907020071002

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.

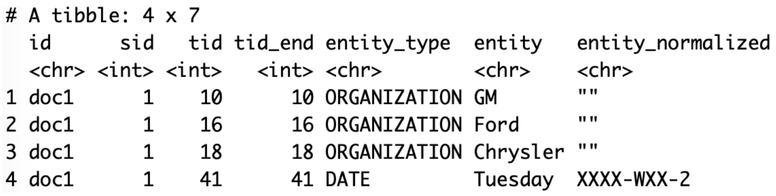
The adoption of second-tier wages for new hires at GM represents an attractive concession for Ford and Chrysler, but the structure of a retiree health-care trust could prove difficult to transfer, sources familiar with the matters said on Tuesday. The establishment of a Voluntary Employees Beneficiary 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.



(b) Number of articles per year.

Figure 1: Sample of a news article in our data together with the time series of the number of articles per year.



```
# 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.

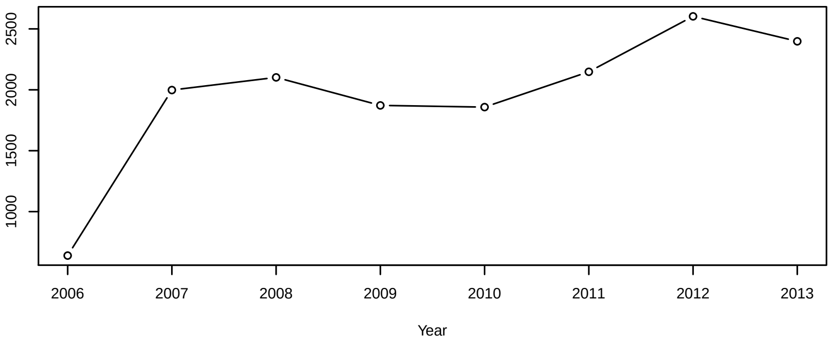


Figure 3: Time series of the number of recognized firms among articles published in a given year.

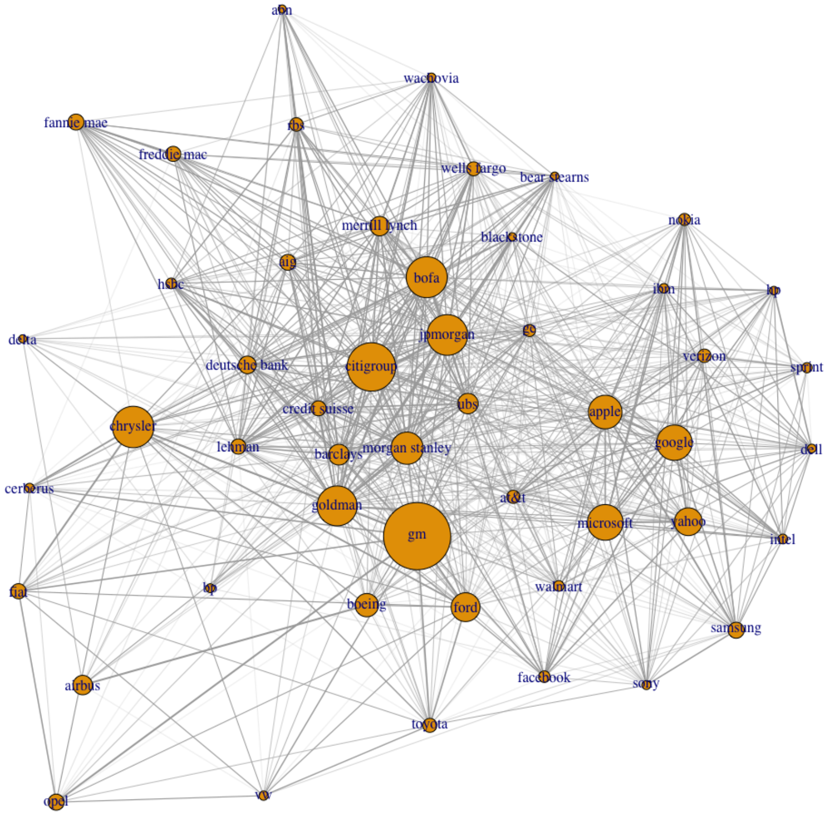
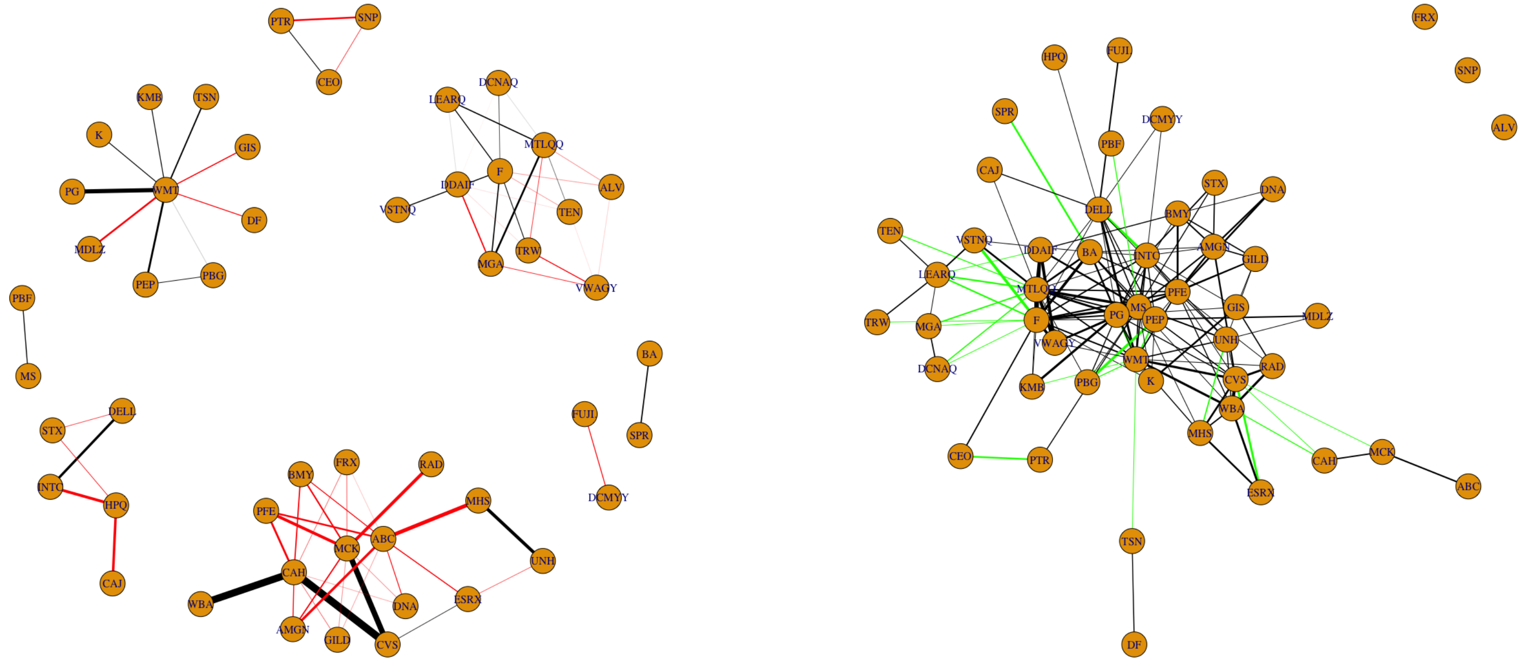


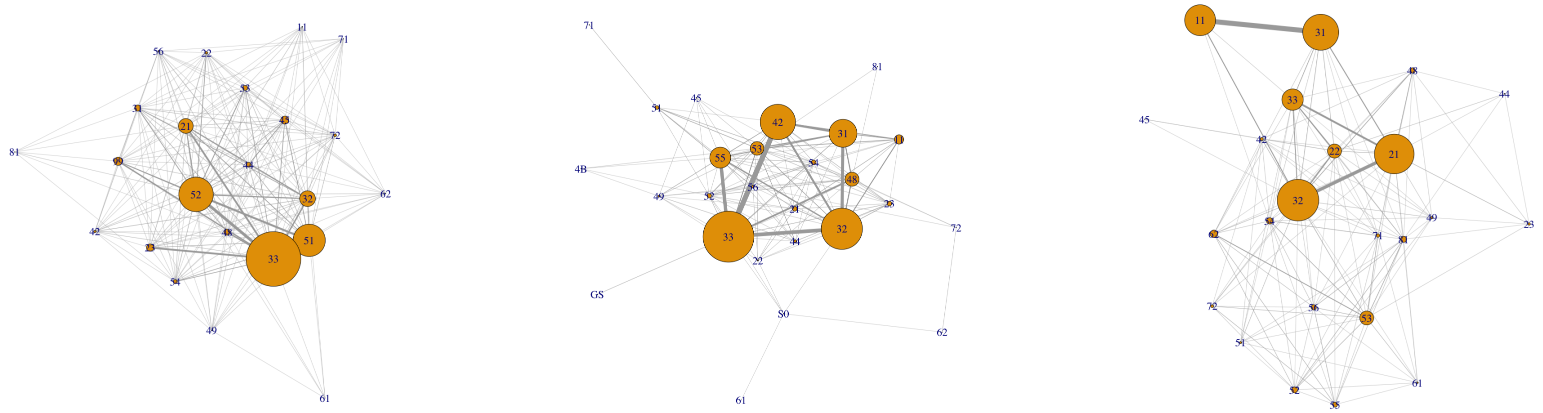
Figure 4: Network of firms implied by the full news data sample covering the years 2006 through 2013. We only plot the largest 50 firm nodes in our network. The size of a node is proportional to the number of times that firm is identified to be connected to another firm in an article in our data. The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



(a) Compustat Segments network.

(b) News-implied network.

Figure 5: Networks implied by the Compustat Segments data and our news data. We obtain data on costumer-supplier relationships covering the same time span as our data from the Compustat Segments database. In Panel (a), we plot that largest 50 nodes as implied by sales in the Compustat Segments data together with all the costumer-supplier links between these top 50 firms. The width of a link is proportional to the value of the sales associated with that costumer-supplier relationship. For the same 50 nodes displayed in Panel (a), in Panel (b) we show the network implied by our news data.



(a) News-implied network.

(b) BEA input network.

(c) BEA output network.

Figure 6: Intersectoral networks implied by our news data and by the 2012 BEA input-output matrix. We obtain data on NAICS codes from the CRSP / Compustat database. We then aggregate firms by the first two-digits of their NAICS sector codes. The size of a node is proportional to the number of times that sector is identified to be connected to another sector in the data. The width of a link between two sectors is proportional to the number of times that intersectoral link is identified in the data. Note that not all firms identified in the news data also have a match in the CRSP / Compustat data. We are able to match 2,386 firms with CRSP / Compustat. We build the BEA input (output) network from the upper (lower) triangular matrix of the BEA data. We use the 2012 Industry by Industry/After Redefinitions/Producer Value Total Requirements table. The size of a node in the BEA input network is equal to the net value of the input from other sectors required to produce one dollar of output. The width of a link is proportional to the net input shared across sectors. In the BEA output network, the size of a node is equal to the net value of the output of that sector which is used by other sectors as input. The width of a link is proportional to the net value of the output shared between two sectors.

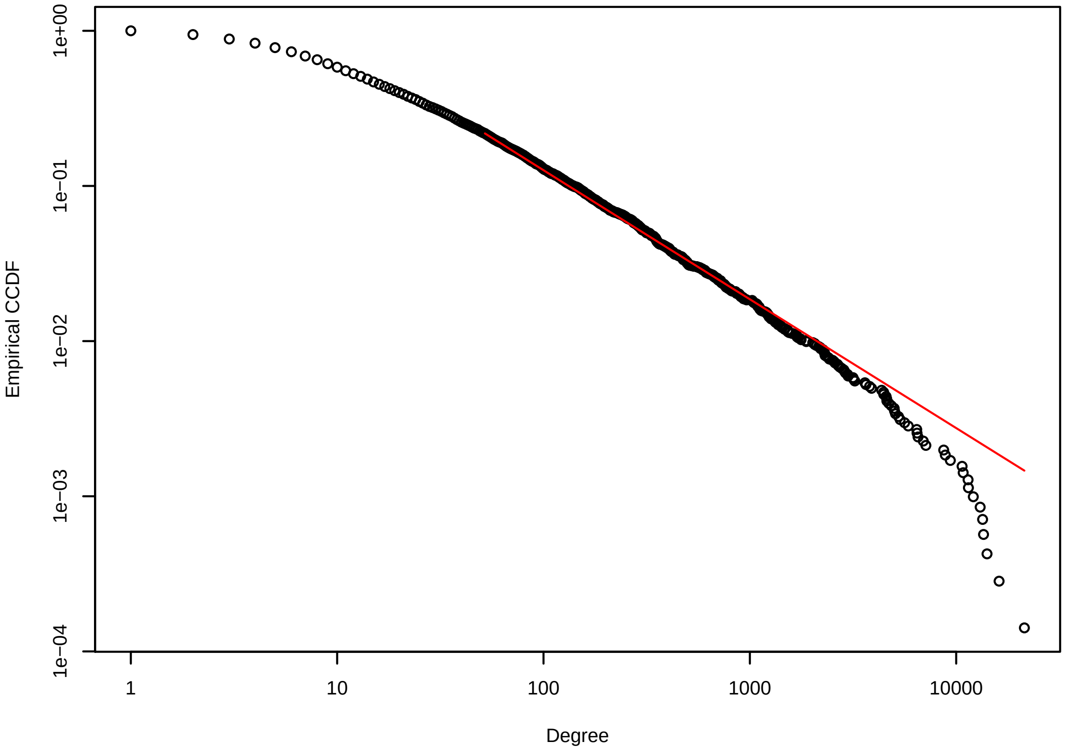
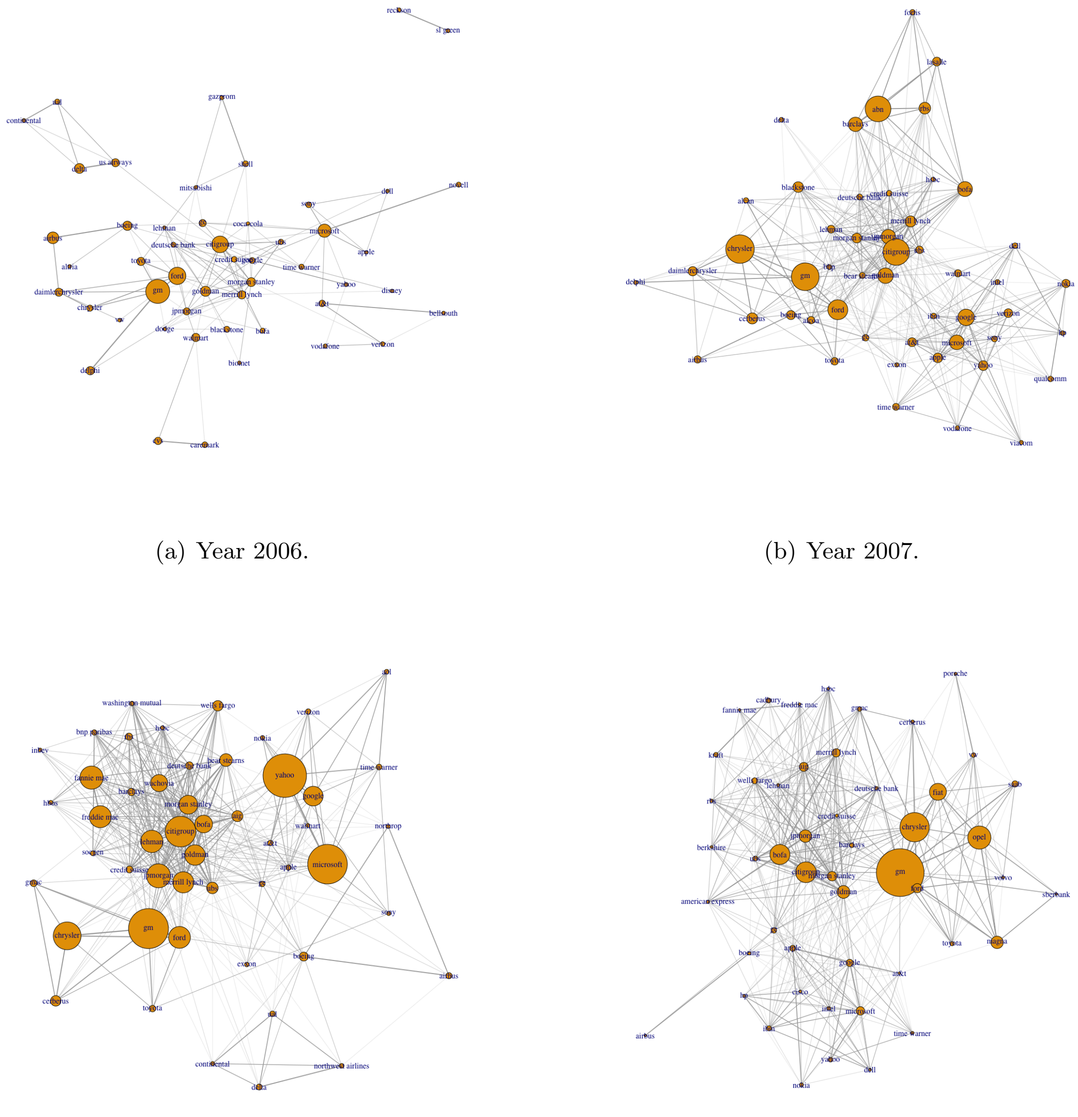


Figure 7: Empirical counter-cumulative distribution of the degree of connectivity of the nodes in our network. The degree of a node in our network is equal to the number of connections that node has. The y-axis shows the probability we can find a node with at least as many nodes as indicated on the x 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.



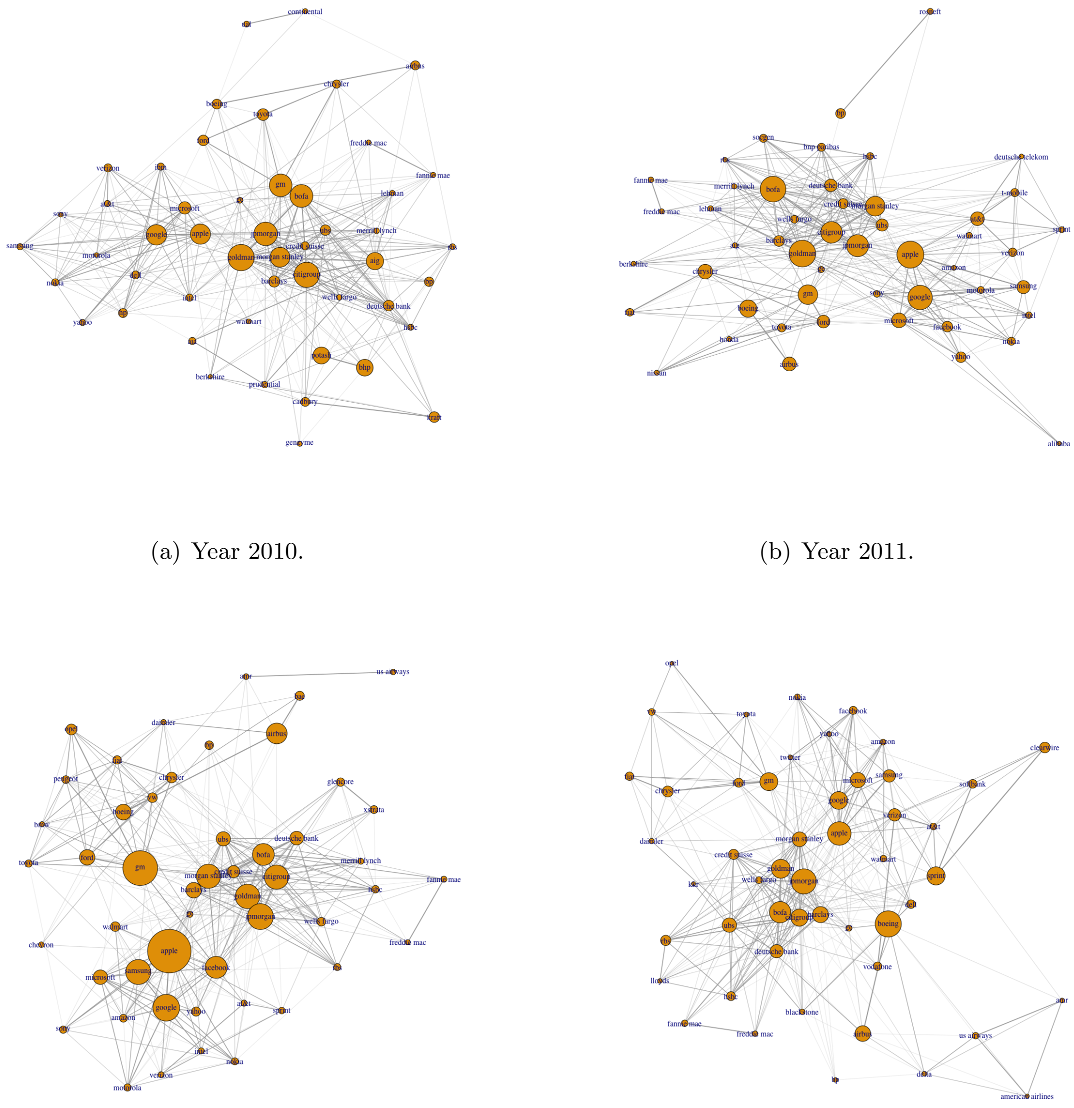
(a) Year 2006.

(b) Year 2007.

(c) Year 2008.

(d) Year 2009.

Figure 8: Time series of news-implied networks in our data sample for the years 2006 through 2009. For any given year, we aggregate the links recognized by the methodology of Section 3 that fall within that year. We aggregate these links into a network and plot the network in R using the package "igraph."



(a) Year 2010.

(b) Year 2011.

(c) Year 2012.

(d) Year 2013.

Figure 9: Time series of news-implied networks in our data sample for the years 2010 through 2013. For any given year, we aggregate the links recognized by the methodology of Section 3 that fall within that year. We aggregate these links into a network and plot the network in R using the package "igraph."

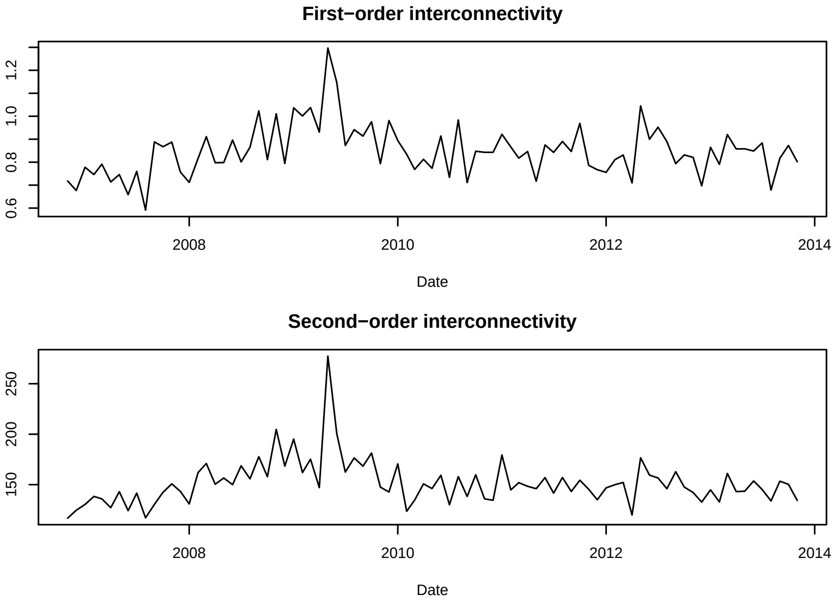


Figure 10: Time series of the first and second-order interconnectivity measures for the networks implied by our news data. For any given month in our data sample, we collect all news article published in that month and extract business connections using the methodology of Section 3. Given the network 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.

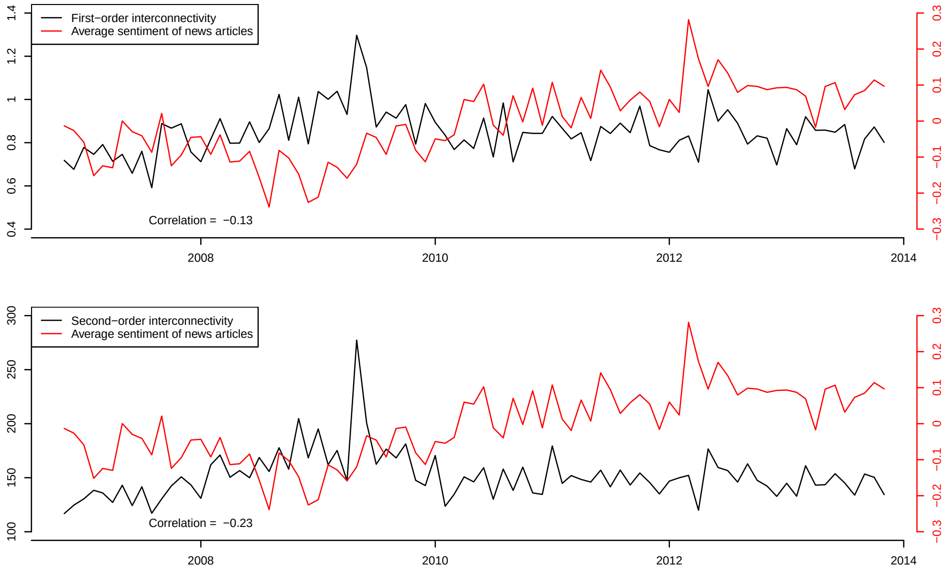


Figure 11: Time series of the interconnectivity measures and the average article sentiment. We construct our measure of sentiment article by article. For each article, we use the sentiment annotator in the coreNLP toolkit to evaluate the sentiment of each sentence and then take the average across the 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.

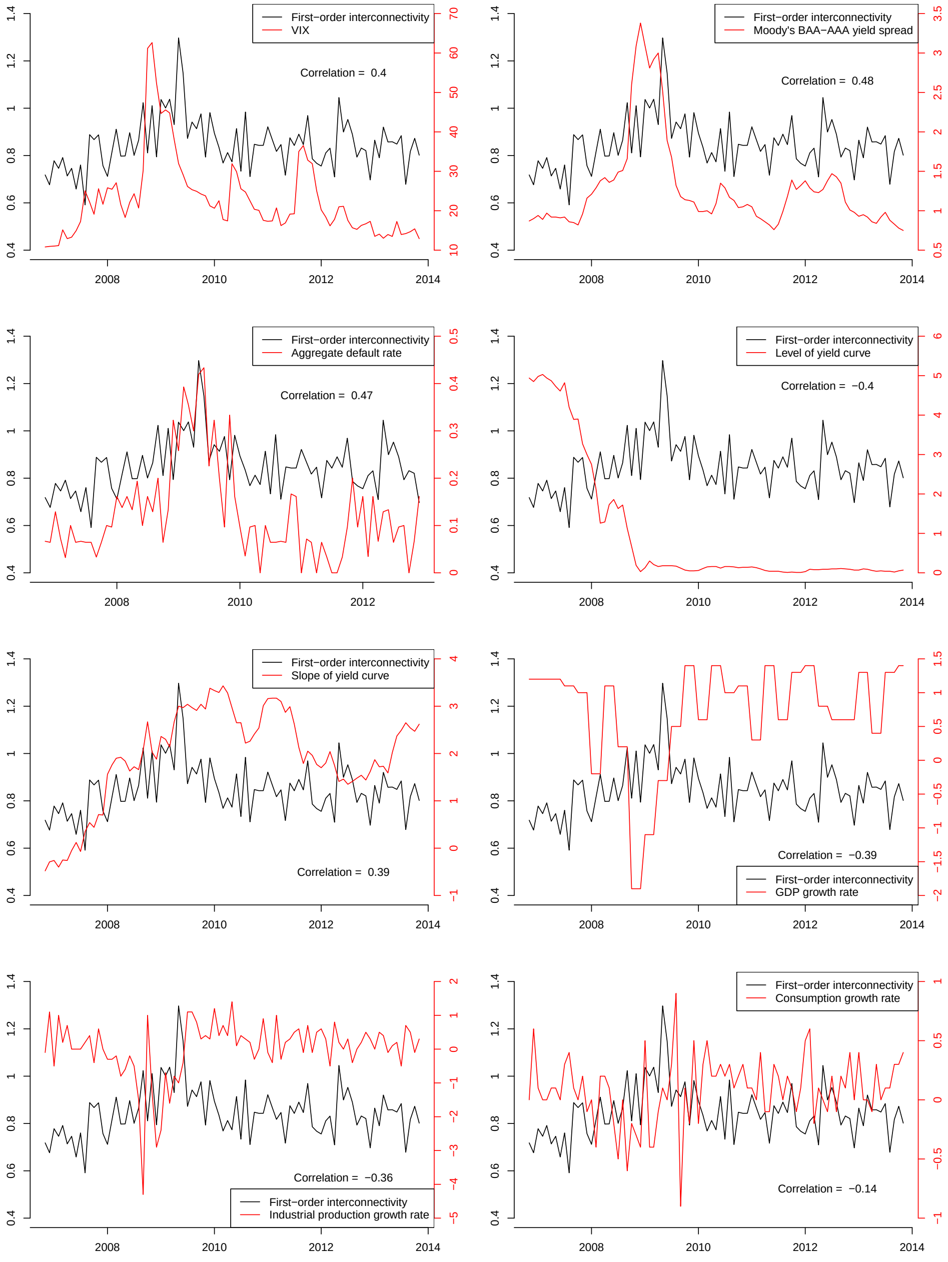


Figure 12: Time series of the first-order interconnectivity measures and financial and macroeconomic factors. See Table 3 for summary statistics of the financial and macro factors.

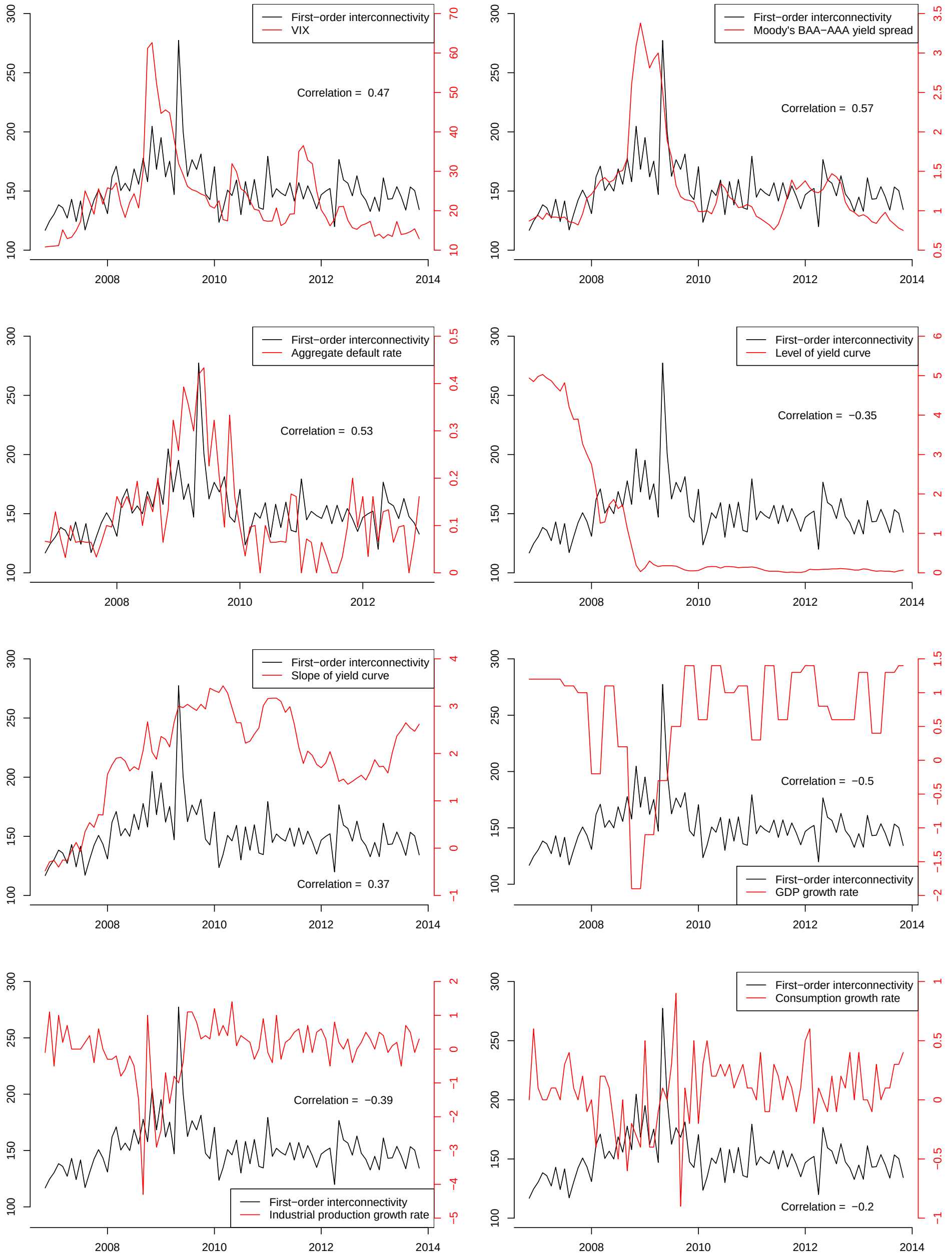
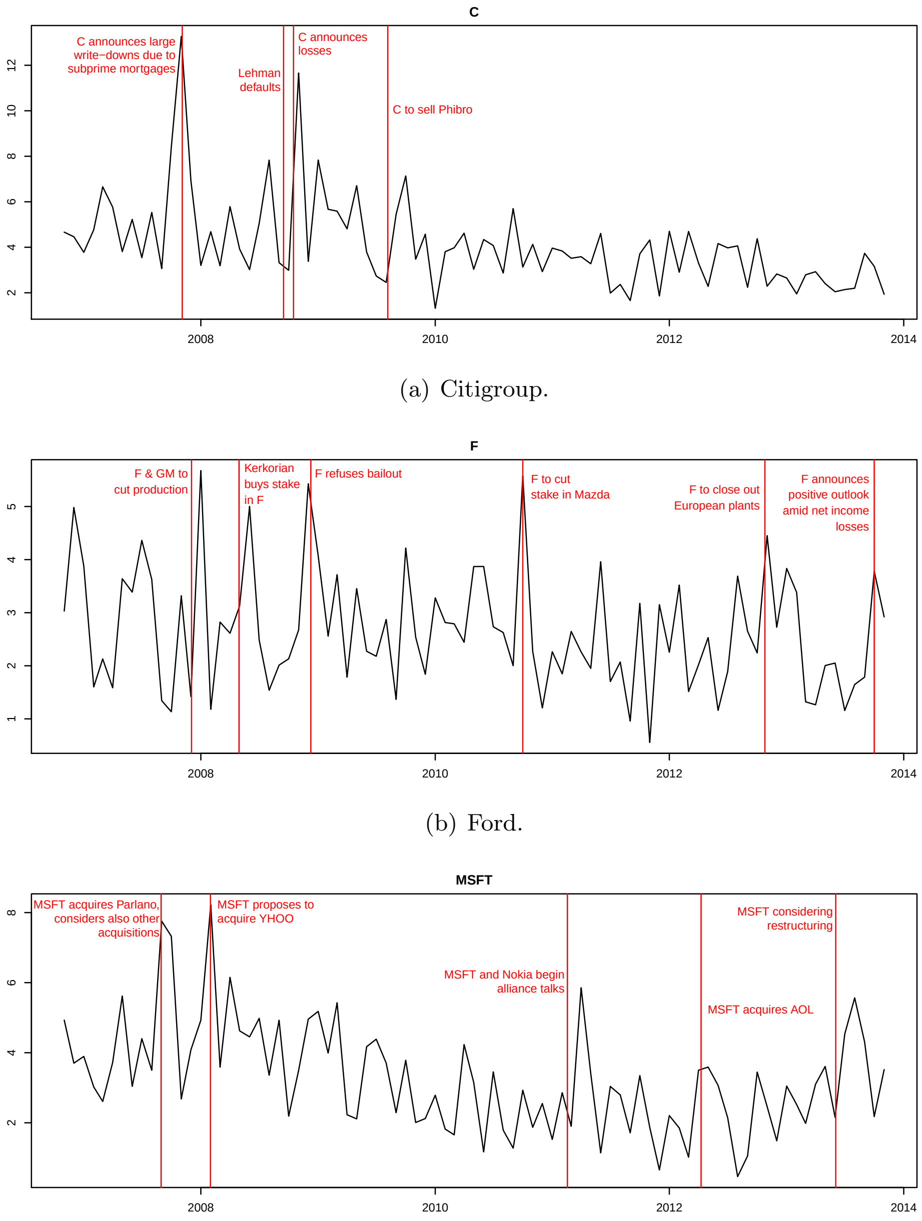


Figure 13: Time series of the second-order interconnectivity measures and financial and macroeconomic factors. See Table 3 for summary statistics of the financial and macro factors.

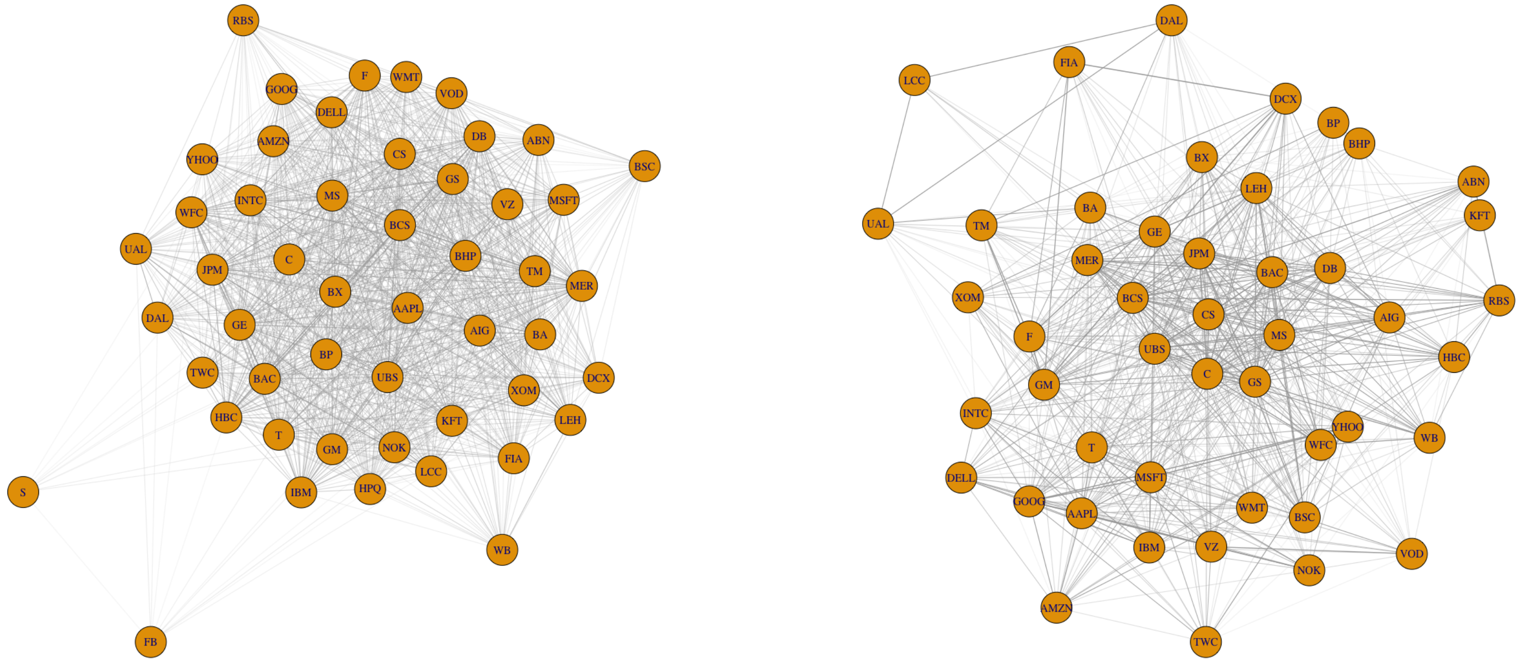


(a) Citigroup.

(b) Ford.

(c) Microsoft.

Figure 14: Time series of connectivity for Citigroup, Ford, and Microsoft. Connectivity is given by the sum of the entries of the column that correspond to a node in the normalized adjancency matrix of the largest 200 nodes in our data. Our connectivity measure corresponds to the weighted outdegree defined in Acemoglu et al. (2012, p. 1985) and is proportional to the number of firms a firm is connected with.



(a) Network implied by stock return correlations over the whole data sample.

(b) Network implied by news articles over the whole data sample.

Figure 15: Networks implied by equity correlations and our news data. We match firms in our data sample with the CRSP / Compustat merged data set to obtain ticker and monthly stock return information. Only one-quarter of all firms in our data sample have a match in CRSP / Compustat. Out of this restricted data sample, we extract the 50 most connected firms as implied by the methodology of Section 3. We compute the correlation matrix for the monthly stock returns of these 50 firms over the whole data sample. The figure in Panel (a) shows the network implied by that correlation matrix. All nodes size are of equal size. The width of a link is proportional to the absolute value of the correlation coefficient between the monthly stock return of the two connected firms. Panel (b) shows the news-implied network for the 50 most connected firms in the restricted sample of firms that have a match in the CRSP / Compustat data.

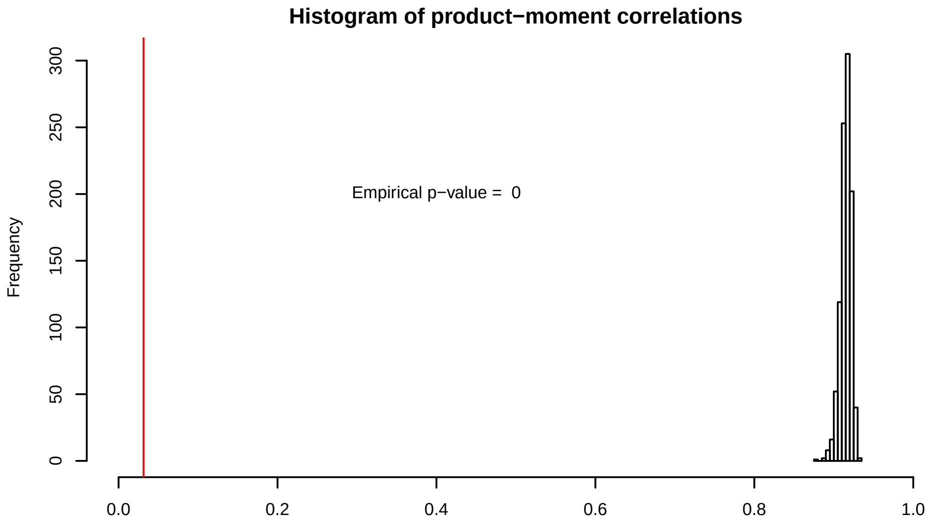
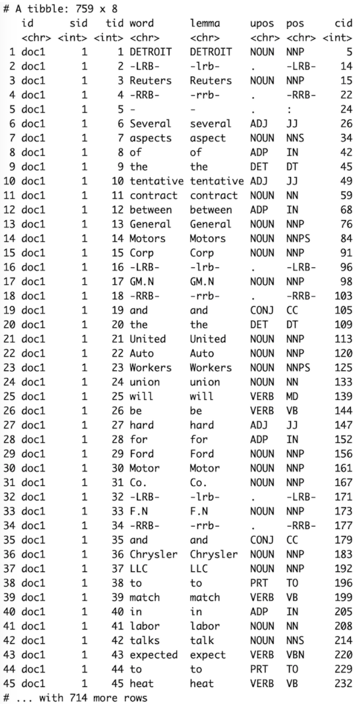


Figure 16: Histogram of the product-moment correlations of the correlation-based network in Panel (a) of Figure 15 and bootstrap samples of the same plot. We generate 1,000 bootstrap samples under the assumption that the stock return correlations used to construct the network in Panel (a) of Figure 15 were measured with error, where the error is normally distributed around the measured correlation with a standard deviation equal to the standard error of the correlation estimate. The product-moment correlation of two networks is the correlation coefficient between the link widths in both networks. The red line indicates the product-moment correlation between the two networks of Figure 15.



```
# 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
```

Figure 17: Output of the coreNLP toolkit.