Text-Based Linkages and Local Risk Spillovers in Equity Market*

Shuyi Ge †

September 6, 2020

Abstract

One stylized fact of asset returns is that the interconnectedness in idiosyncratic returns is non-negligible even in large dimensional systems. The network architecture of firms is the key to study the transmissions of local shocks. However, such linkage data is usually unavailable for researchers. This paper uses extensive text data to construct firms' links that have not been documented in other sources. Utilizing the novel text-based linkage data, I quantity the strength of local risk spillovers in the equity market by estimating a heterogeneous spatial autoregressive model (HSAR) for the de-factored (idiosyncratic) equity returns. The model outperforms several alternative methods in terms of out-of-sample fit. The estimation results show that after removing the common risk factors and industry risk factors, there is still a considerable degree of local risk spillovers, and with substantial industrial heterogeneity. By constructing spatial-temporal spillover matrix using estimated parameters, we are able to identify the major systemic risk contributors and receivers, which are of the interest to microprudential polices. From a macroprudential perspective, a rolling-window analysis reveals that the strength of local risk spillovers increases during crisis period, when, on the other hand, market factor loses its importance.

Keywords— Networks, text analysis, systemic risk, local risk spillovers, weak and strong cross-sectional dependence, heterogeneous spatial autoregressive model (HSAR)

[*] The author would like to thank Oliver Linton, Hashem Pesaran, Michael Leung, Shaoran Li, Merrick Li, Chen Wang, Weiguang Liu, Chong Shu, Sampad Mohanty, Yimeng Xie, Lidan Tan, Grigory Franguridi, Chris Yoo and all the participants at the University of Cambridge Econometrics workshop and the University of California Econometrics workshop for their useful comments.

[†] The author is at the Faculty of Economics, University of Cambridge. Her email address is sg751@cam.ac.uk

1 Introduction

According to asset pricing theories such as the classical capital asset pricing model (CAPM) developed by Sharpe (1964)[33], and the arbitrage pricing theory (APT) of Ross (1976)[30], asset returns have a common factor representation with a strong pervasive component driven by a few common factors and an idiosyncratic component that is weakly correlated. Many studies (Kou, Peng and Zhong (2018)[22], Baily et al. (2019[5],2020[6]) among others) have found that the APT models with the factors in the existing literature seem to be not sufficient to capture all the significant interdependencies in asset returns. Local risk spillovers may still play a non-negligible role even in large-dimensional systems (see Gabaix (2011)[16], Acemoglue et al.(2012)[1] and Barigozzi and Hallin (2017)[9] for example). The network architecture of firms is the key to study the local spillovers of idiosyncratic risk. However, such linkage data is usually unavailable to researchers, which hinders the studies of local dependencies.

This paper uses extensive text data to construct firms' linkages. LexisNexis Academic news database has a collection of news from a wide range of sources. Company names and tickers that are mentioned in the each piece of news are tagged. By this feature, I identify firms that share business links by common news coverage. The maintained assumption is that two companies share a link if they are the only two that get mentioned in the same piece of news. The estimated full sample network is plotted in Figure 6, which has a core-periphery structure. Big banks including JPMorgan Chase (JPM), Citi (C), Goldman Sachs(GS) and Bank of America (BAC), and big hitech firms including Microsoft (MSFT), Apple (AAPL), Intel (INTC) and Oracle (ORCL), and big manufacturers and conglomerates including General Electric (GE) and Procter & Gamble (PG) are the most connected companies from the $S\&P500$ universe, occupying the center of the graph. The novel dataset complements existing network datasets in several perspectives. While existing network datasets are usually lagged, incomplete, and cover certain types of links for certain types of firms[1], the link mining method complements these information sources by identifying additional types of links that have not been documented elsewhere. In addition to interbank relationship and customer-supplier links, the method also finds strategic partnerships, business lines acquisitions, investment banking relationships, funding relationships, similar legal and regulatory exposures, and $M\&A$ relationships, etc. As a comparison, Figure 8 plots the network among $S\&P500$ firms using Compustat segments data and it is visible that much fewer links are identified. In response to the lack of network data, there has been a strand of literature using pure statistical methods to estimate links from a panel of equity returns/volatilites (see Diebold and Yilmaz (2014)[14], and Hale and Lopez (2019)[18], Barigozzi and Hallin (2017)[9], Barigozzi and Brownlees (2019)[8]), and Demirer et al. (2018)[13]) Figure 9 plots the long-run variance decomposition network (LVDN), long-run Granger causality network (LGCN) and partial correlation network (PCN) among $S\&P500$ companies estimated from the idiosyncratic returns using the high-dimensional methodology of Barigozzi and Hallin (2017)[9]. The links identified are very different from period to period and very few links from the crisis LVDN appear before the crisis. The links that turn out to be important for risk transmissions in the crisis period are like the hidden iceberg that is hard to detect ex-ante and reveal themselves only when large shocks hit the system. Additional sources of information could be fruitful in aiding the link detection, as the text-based network constructed using pre-crisis news outperforms

[1] For example, interbank network data only covers the lending relationships among banks, and they are not even publicly available. The Compustat segment data is available to researchers, but it only contains customer-supplier link.

LVDN estimated using pre-crisis sample in terms of detecting LVDN links from the crisis period.[2] This is due to the fact that our text-based links are much more persistent. On average, 59.32% of the linked pairs identified in a year continue to get identified in later years, showing that the method identifies long-lived economic links among companies. Taken together, it can be seen that the text-based network complements alternative network datasets and can be viewed as a promising alternative to other datasets.

Utilizing the novel text-based linkages data, I quantify the strength of local risk spillovers in the equity market using a heterogeneous spatial autoregressive model (HSAR) studied in Baily, Holly and Pesaran (2016)[4] and Aquaro, Baily and Pesaran (2019)[2]. The model captures temporal dependence as well as spatial-temporal dependence. It is flexible, and individual-specific parameters can be consistently estimated for any N as long as T is large. Since the equity returns comovement reflects both exposures to common risk factors and local risk spillovers, I first remove the strong cross-sectional dependence (CSD) by de-factoring equity returns. I show that after removing the common risk factors and industry risk factors, there is still a considerable degree of local risk spillovers via the links implied by the news. The flexible framework allows us to study the industrial heterogeneity in terms of the intensities of the local risk spillovers. We find a substantial degree of industrial heterogeneity. In particular, manufacturing firms and financial firms are more sensitive to the shocks of their neighbours. It is also worth noticing that the lead-lag effect in the risk spillovers for financial firms is more pronounced as for any lag order, the percentage of significant individual-specific spatial-temporal coefficients are about twice as large as that of other industry groups. The spatial-temporal framework allows us to analyse a complicated diffusion pattern of local shocks over time and space. The decay of shock along spatial dimension is slower than that along time dimension. By constructing spatial-temporal spillover matrix using the estimated parameters, we are able to identify the major systemic risk contributors and receivers, which are of the interest to microprudential policy makers. The firms contribute the most to the systemic risks are the large cap financial institutions and manufacturers. Apart from systemic risk contributors, companies that are particularly sensitive to others' shocks are also found. It is worth noticing that the well-connected systemic risk contributors themselves are not necessarily the major risk receivers. They are the periphery firms that receive a lot of risks from the core. To assess the performance of the proposed method and evaluate the benefits of using this novel data, I compare the in-sample and out-of-sample mean squared error (MSE) of the spatial-temporal model estimated using different adjacency matrices, and an alternative high-dimensional VAR approach that requires no explicit link information from Barigozzi and Hallin (2017)[9], which we refer to as BH-VAR for short. In terms of in-sample fit, BH-VAR has the smallest MSE. This is not surprising, given the method selects the model by minimizing a Bayesian information criterion. However, when we look at out-of-sample fit, which is more important practically, the spatial-temporal model estimated with the text-based network outperform all other specifications.

To examine how the strength of local risk spillovers evolve over time, I consider a rolling window analysis.

[2] Table 13 shows the percentages of crisis period Long-run variance Decomposition network (LVDN) links that get identified using alternative pre-crisis network information. Different hard thresholds are applied to the LVDN given the network implied by LVDN is very dense (the link densities for pre-crisis and crisis sample are 77.5% and 95.3%, respectively). We do not need to apply thresholding to text-based network since it is already very sparse (the link density of the full sample network is 4.5%, and for the short pre-crisis sample the density is even smaller). For any non zero thresholds applied, the text-based networks consistently outperform that of pre-crisis LVDN in terms of detecting out-of-sample links.

The estimation results reveal that the local dependencies intensify during periods of financial crisis and turmoils. The surge in local risk spillovers could be a signal of rising systemic risk, which is useful for macroprudential purposes. Previous studies have documented that asset returns depart from fundamentals during times of financial crisis, and stocks dis-connect from the market factor (Baily et al.(2019[5],2020[6])). Our analysis tracks the evolution of strong cross-sectional dependence (CSD) and weak cross-sectional dependence (CWD) at the same time and it documents an interesting fact: the local risk spillovers intensifies when then market factor loses its importance during the financial crisis and turmoils, which is evidence for market decoupling.

This paper contributes to three strands of literature. The first strand of literature that it relates to is textual analysis and its application in the financial market. In particular, how to quantify the soft information contained in news articles. Text analysis has been a useful tool to construct novel datasets. It fills the gaps in data availability induced by limited disclosure and slow update, thus complement traditional economic datasets. For example, there has been an exploding number of researches on sentiment analysis (for example, Garćia (2013)[17], Price et al. (2012)[29], among others. For a survey of textual analysis in accounting and finance, see Loughran and Mcdonald (2016)[26]). Sentiment, unlike other traditional economic variables, is hard to measure. Thanks to the text analysis techniques, it has become available to the hand of researchers. Similar to the sentiment index, there has been an economic policy uncertainty index (EPU) developed by Baker et al. (2016)[7], which is based on newspaper coverage frequency of political words. Text analysis has also been used for link mining. Hoberg and Phillips (2016)[20] construct peer links by applying text analysis on firm 10K Product description. Scherbina and Schlusche (2015)[31], Schwenkler and Zhang (2019)[32] both identify firm links from business news. The second strand of literature this paper is related to is the local risk spillovers in the equity returns. Local shocks transmit among economically-linked firms. Cohen and Frazzini (2008)[12] found evidence of return predictability across firms linked by supply chains. Scherbina and Schlusche (2015)[31] found that there is cross-predictability in returns between firms linked via various types of business relationships. Equity returns comovement reflects both exposure to common risk factors and local risk spillovers. While exposure to common factors gives rise to strong cross-sectional dependence (CSD), local risk spillovers represent weak cross-sectional dependence (CWD), and the latter form of interdependence receives much less attention compared with the former one. A key reason for the lack of empirical work is the lack of network information. Using the text-based network, the paper documents the existence of 'excess-comovement' in linked stocks beyond what is predicted by standard asset pricing models. Compared with the high-dimensional VAR approaches which shrink, select and estimate high-dimensional network used by Barigozzi and Hallin (2017[9], 2019[8]) and Demirer et al. (2018)[13] when there are no explicit links observed, our approach outperforms in terms of out-of-sample fit. I also compare the performances of the spatial-temporal model estimated using alternative publicly networks, and the text-based network outperforms both in terms of in-sample and out-of-sample mean squared errors (MSE). Utilizing the novel dataset, the paper proposes a promising new method of addressing local risk spillovers in the equity returns. In the end, it also contributes to the studies of the network structure in equity market and has important implications from both microprudential and macroprudential perspectives.

The rest of the paper is organized as follows: section 2 describes the data and link identification strategy and shows some key properties of the estimated linkages. Section 3 talks about the modeling of strong and weak cross-sectional dependence using a factor plus spatial two-stage procedure. The main focus in on local risk spillovers (CWD) among linked stocks. Section 4 provides full sample estimation results and the construction of spatial-temporal spillover matrices using estimated parameters. And this section also presents the model comparison results. Section 5 provides a rolling window analysis and characterizes the evolution of local risk spillovers over time. Section 6 shows the robustness check and placebo test. Section 7 gives some concluding remarks.

2 Data and Link Identification

All the stock market related data are from the Center for Research in Security Prices (CRSP). Since our econometric framework requires large T for consistent estimation, I use the daily stock file. Industry classification is based on Standard Industrial Classification (SIC) code from the CRSP/Compustat Merged database and the modified classification criteria provided on the Kenneth French's home page. As I will elaborate more in section 4, to obtain mean group (MG) estimates of each industry group's parameters, one need the number of stocks within each group to be big enough. Due to that consideration, I build the industry classification on top of FF5 industry definitions where they classify all stocks according to their SIC code into 5 broad groups: 'Consumer', 'Health', 'Hitech', 'Manufacturing' and 'Others'. For the first four categories, I keep the same definitions as Fama and French. Since there are a large proportion of financial companies in the $S\&P500$ universe and our sample period covers the financial crisis, it would be interesting to separate financial firms from the 'Others' category. Among the stocks that fall into 'Others', I categorize the stocks with SIC in the range $6000 - 6799$ as 'Finance' and make the rest of the stocks stay in the 'Others' category. Daily Fama-French factor returns and industry portfolios' returns are taken from Kenneth French's home page.

As for the text data, I download all the full-text business news from Business Wire that tagged $S\&P500$[3] companies from January 2006 to December 2013 on LexisNexis Academic[4]. A news contains a title, date, body and classification. A typical business news in the dataset is in shown in Figure 5. This example news reports the strategic partnership between American Express and Regis Corporation. The main subject of the news are summarized by some key words. And in the classification section, the relevant companies are tagged with their tickers listed. There are $345,880$ distinct business news that tagged sample companies during the whole sample period, and each sample month has around $3,200$ distinct business news. This section will mainly focus on the identification of links from our text database and some key properties of those links.

2.1 Identification of Links

Common news coverage reveals information about linkages among companies. In this paper, links are identified by common business news coverage. The identification assumption is that if a piece business news reports two companies together, then the two firms have some sort of business relationship/link. Although news that mention multiple companies together may carry potential information about links, they are more noisy (for

[3] The composition of $S\&P500$ index changes over time. All stocks that have stayed on the list and have no missing return observations for more than one year during the sample period are considered.

[4] LexisNexis Academic is a database of full text online news, legal cases and company from information. News from hundreds of source are available. After entering the company names and narrowing down the subject to 'business news', Business Wire is always among the top sources list. To maximize the number of relevant business news during sample period and to avoid duplications, only the news from Business Wire is used. The python code of data scraping is available upon request

example, analyst recommendations, ratings changes, index movements might stack multiple companies together when they actually don't have real links). Due to that concern, I discard news that tag more than two firms.

I use a $N \times N$ adjacency matrix $W = (w_{ij})$ to store all the links identified in the sample news. N is the number of sample companies and a typical entry w_{ij} is the number of times i and j are co-mentioned in different news. The link estimation procedure is as follows. For each piece of distinct news in the sample, (1) firstly we extract the tickers tagged; (2) keep the news if only two distinct[5] tickers are tagged; (3) match the tagged with sample companies; (4) if both tagged companies are successfully matched, say if the they are matched to the companies correspond to the ith and jth row/column, then we add one to both w_{ij} and w_{ji}. Process (1)-(4) is repeated for every piece of distinct news in the sample.

2.2 Estimated Links

The links identified using all the business news from Business Wire from 2006 to 2013 is plotted in Figure 6. Only companies with links are plotted on the figure. Given a long sample period and huge amount of business news, very few sample companies never got co-mentioned with others[6]. In the figure, nodes represent companies and two nodes are connected by an edge if there is a link between them. The size of a node is proportional to the number of neighbours it has (i.e., its degree) and the color of a node indicates which industry it is in.

The estimated full sample network has a core-periphery structure. The most connected companies in the network graph include big banks, big hitech companies and big manufactures. Big banks including JPMorgan Chase (JPM), Citi (C), Goldman Sachs(GS) and Bank of America (BAC) and big hitech firms including Microsoft (MSFT), Apple (AAPL), Intel (INTC) and Oracle (ORCL) and big manufacturers and conglomerates including General Electric and Procter & Gamble (PG). And they occupy the center of the graph. Table 8 provides the link validation results for the most frequently mentioned pairs. Big banks engage in a variety of business relationships with other companies including financing, joint venture, strategic partnerships, joint investment banking, acquisition of business lines and competition. Hitech giants are very well connected with each other to form strategic partnerships and develop new products together. Supplier-customer relationship and business lines acquisitions are found among Big manufacturers and conglomerates. Companies within the same industry appear as clusters, indicating there are dense intra-industry linkages. Most of the hitech firms lie on the third quadrant (bottom left corner) of the graph, while most of the health and consumer companies show up in first quadrant (top left corner) of the graph. Manufacturing and financial companies companies are more scattered.

Table 9 gives a summary statistics of the news-based links estimated using full sample. In total, 40185

[5] A same company listed on different stock exchanges may have different tickers. For example, Citi used to list on both the New York Stock Exchange (NYSE) and the Tokyo Stock Exchange (TSE) at the same ticker with different ticker names: C(NYSE) and 8710 (TSE). To avoid double counts of a same company, only tickers associated with the New York Stock Exchange (NYSE), National Association of Securities Dealers Automated Quotation System (Nasdaq) and American Stock Exchange (AMEX) are kept.

[6] For the balanced panel of 413 sample companies, only 5 out of 413 never got co-mentioned with others. There are 546 firms that have stayed on the $S\&P500$ list and had no missing return observations for at least a year, and they are included in the one year rolling sample. Among 546 firms that are included in the rolling sample, only 26 of them never got co-mentioned with others.

links are identified in the full sample period, and among them there are 6742 unique pairs of companies that share links. The former number is much larger than the latter one since most pairs of firms are mentioned together multiple times in different articles. The link density of the full sample network 4.5%. Those links are discovered over time and the yearly network plotted in Figure 7 are sparser. Over the full sample period, each company are connected to around 24 other companies in the $S\&P500$ universe on average. The network shows a substantial level of degree heterogeneity with a small number of firms being highly-connected, which is shown by the $90th$ percentile of degree. And this feature is consistent with the core-periphery structure of most empirical networks. To have a more detailed understanding of the features of the links, I further break down the links to intra-industry and inter-industry links. Our method identifies a lot of inter-industry links as well as intra-industry links. Although the full sample network is sparse, the intra-industry link densities are high. For example, the intra-industry link density for hitech companies reaches 16%. This feature is obvious from Figure 6, where companies within the same industry appear as clusters. Most of the hitech firms lie on the third quadrant (bottom left corner) of the graph, while most of the health and consumer companies show up in first quadrant (top left corner) of the graph. Manufacturing and financial companies companies, on the other hand, are more scattered. Compared with the peer link mining method in the literature (see Hoberg and Phillips (2016)[20], Lee, Ma and Wang (2015)[23], the method used in this paper has the advantage of discovering not only peer links but also inter-industry links. Over the full sample, the number of distinct cross-sector pairs identified is larger than the number of distinct intra-sector pairs identified. And this is still true even if we look at different industries separately.

The full sample period is long and different links are identified in different years. The network graphs and summary statistics for each year from $2006 - 2013$ are given in Figure 7 and Table 10, respectively. For the links identified using only one year's news, the statistics are much smaller than that from Table 9. This implies that new links get identified over time and they carry timely information about the interconnectedness among companies. To roughly gauge the percentages of 'new' and 'stale' links, I calculate the percentage of linked pairs identified in one year that were identified in the previous year. On average, 37.82% of the linked pairs identified in a year are 'stale' links and 62.18% are 'new' links. However, the high percentage of 'stale' information is not necessarily a bad thing, as it implies the news-based links are persistent. I calculate the percentage of linked pairs identified in one year that continue to get identified in later years. On average, 59.32% of the linked pairs identified in a year continue to get identified in later years, showing that the method identifies long-lived economic links among companies.

2.3 Comparison with other networks

The novel dataset complements existing network datasets in several perspectives. While existing network datasets are usually lagged, incomplete, and cover certain types of links for certain types of firm. The link mining method complements these information sources by identifying additional types of links that have not been documented elsewhere. Figure 8 plots the network of $S\&P500$ firms using Compustat data. Consumer companies such as Walmart, McKesson are well-connected given they have a wide range of suppliers and customers. Apart from several consumer companies, there is no apparent star in the network. Very few links of financial firms are uncovered. On the other hand, the link mining approach applied in this paper uncovers a huge amount of intra-industry as well as inter-industry links for financial firms.

Instead of turning to existing limited network dataset, there has been a strand of literature using pure statistical methods to estimate links from a high-dimensional time series (Barigozzi and Hallin (2017)[9], Barigozzi and Brownlees (2019)[8], Demirer et al. (2018)[13]). Figure 9 plots the long-run variance decomposition network (LVDN), long-run Granger causality network (LGCN) and partial correlation network (PCN) among $S\&P500$ companies estimated from the our sample of idiosyncratic returns using the high-dimensional methods from Barigozzi and Hallin (2017)[9]. Although we are using idiosyncratic returns while they use idiosyncratic volatilities, two prominent features remain true. The first feature is that the Financial Crisis has blown up the interconnectedness in the system. From figure 9, it is clear that for all 3 types of networks considered, the network in the Crisis period is much denser than that of others[7]. The second feature is that the links identified are very different from period to period. Table 13 shows the percentages of thresholded crisis period LVDN links that also appear in the thresholded pre-crisis LVDN. Expect the results for no thresholding, where the link densities for pre-crisis and crisis sample are 77.5% and 95.3% respectively, for other thresholds applied, very few links from the crisis LVDN appear before the crisis. From those two features, the links that turn out to be important for risk transmissions in the crisis period are like the hidden iceberg that is hard to detect ex-ante and reveal themselves only when large shocks hit the system. As a result, such high-dimensional link estimation method alone is not so useful for policymakers to monitor systemic risk. Additional sources of information, could be fruitful in aiding the link detection. For example, if we apply 5% hard threshold to the both the pre-crisis and crisis LVDN, then only 4% of the thresholded LVDN links identified from the crisis period were also identified from the pre-crisis period. On the other hand, our text-based links identified from the same pre-crisis period reveals the 34% of the thresholded LVDN links identified from the crisis. For other non-zero thresholds, the text-based links consistently outperform that of pre-crisis LVDN. This is due to the fact that our text-based links are much more persistent. On average, 59.32% of the linked pairs identified in a year continue to get identified in later years, showing that the method identifies long-lived economic links among companies. Taken together, it can be seen that the text-based network complements alternative network datasets and can be veiw as a promising alternative to other datasets.

3 Local Risk Spillovers Among Linked Stocks

Equity returns comovement reflects both exposure to common risk factors and local risk spillovers and the latter source of comovement receives much less attention compared with the former one. However, the models that focus on strong cross-sectional dependence such as CAPM and ATP fail to capture all the cross-sectional dependence in the equity returns. There are many work show that the local dependence in the idiosyncratic component is non-negligible (Gabaix (2011)[16], Acemoglue et al.(2012)[1], Barigozzi and Hallin (2017)[9], Kou et al. (2018)[22] among others), thus it is important to examine the role played by local interactions. Adopting the econometrics framework in Baily, Holly and Pesaran (2016)[4], I remove the strong cross-sectional dependence using a factor approach and then use spatial models to examine the local risk spillovers (weak cross-sectional dependence) remaining in the idiosyncratic returns. Unlike spatial interactions in the geographical

[7] Table 11 shows the number of links from the thresholded LVDN for pre-crisis, crisis and full sample periods (different thresholds applied). Table 11 shows the number of links from the LGCN and PCN forpre-crisis, crisis and full sample periods

systems, where there exist natural network structure, for a panel of equity returns there is no natural network structure. The text analysis approach in this paper helps to identity the business links among listed firms and thus allows us to construct the channels of which local shocks transmit. It is found that there is significant local dependence among linked firms' idiosyncratic returns.

3.1 De-factoring Equity Returns

To disentangle weakly correlated idiosyncratic return from the strongly correlated returns driven by pervasive factors, one could use factor models. To be specific, I apply the below hierachical factor model:

$$r_{it} - r_{ft} = \alpha_i + \mathbf{b}_i'\mathbf{f}_t + \gamma_i'\mathbf{f}_{g,t} + \epsilon_{it} \tag{1}$$

where r_{it} denote the return of stock i at time t and subtracting risk free rate r_{ft} gives the excess return. $\mathbf{f}_t$ is the K_1 vector of common risk factors that affect every stock in the market. Since a large proportion of the links identified are intra-industry links, to avoid suprious found spillovers that are actually caused by industry common factors, we add the K_2 industry risk factor $\mathbf{f}_{gt}$ that affect members of industry g but not others. $\mathbf{b}_i$ and $\gamma_i$are the loadings of common risk factors and industry risk factors, respectively. For the choice of factors, we can either use observed factors like Fama-French factors or statistical factors extracted using principal component analysis.

Our analysis need the number of members to be large within each industry group g, so we consider 6 broad industry categories that I will elaborate in details next. For the choice of $\mathbf{f}_t$, I consider 5 factors proposed by Fama-French (2015)[15] plus the momentum factor. And as for the industry factors $\mathbf{f}_{g,t}$, I use the within group cross-sectional averages. As an alternative to observed market and industry factors, one could use unobserved factors, and I will use it as a robustness check.

3.2 Local Risk Spillovers: a Heterogeneous Coefficient Spatial-temporal Model

3.2.1 Heterogeneous Coefficient Spatial-temporal Model

After removing the strongly pervasive component driven by common risk factors, the cross-sectional dependence in the remaining part is weak (local). Spatial econometrics methods are natural tools to address the weak (local) cross-sectional dependence in the idiosyncratic component, where entities interact locally. Conventional homogeneous spatial models restrict the spatial response parameter to be the same across all units. While such restriction is necessary for small T panels, it need not to be imposed when T is large. For a panel data set with sufficiently large T, one can exploit the data along the time dimension to estimate individual-specific parameters for all N units.

One might reasonably suspect that the sensitivity to neighbours' risks is different from firm to firm. Since stock market data set usually covers long time period, we can utilize this nice feature to explore the heterogeneity in the strength of local dependency. The local risk spillovers in the idiosyncratic component is modelled using a heterogeneous coefficient spatial-temporal model ((Bailey et.al 2016[4], LeSage and Chih 2016[25] and Aquaro et.al 2019[2]) that written as follows:

$$\boldsymbol{\epsilon}_t = \mathbf{a}_\epsilon + \underbrace{\sum_{k=1}^{L_1} \boldsymbol{\Lambda_k \epsilon}_{t-k}}_{\text{temporal dependence}} + \underbrace{\sum_{k=0}^{L_2} \boldsymbol{\Psi_k} W \boldsymbol{\epsilon}_{t-k}}_{\text{spatial temporal dependence}} + \boldsymbol{\upsilon}_t \tag{2}$$

where $\boldsymbol{\epsilon_t}$ is the $N \times 1$ de-factored returns and $\mathbf{a}_\epsilon = (\alpha_{\epsilon,1}, \ldots, \alpha_{\epsilon,N})$ is the $N \times 1$ vector of intercepts. $\boldsymbol{\lambda}_k = diag(\lambda_{k,1}, \ldots, \lambda_{k,N})$ is autoregressive parameters of the kth lag for $k = 1, \ldots, L_1$, $\boldsymbol{\Psi}_0 = diag(\psi_{0,1}, \ldots, \psi_{0,N})$ is the contemporaneous spatial coefficients and $\Psi_k = diag(\psi_{k,1}, \ldots, \psi_{k,N})$ is the spatial-temporal parameters of the of the kth lag for $k = 1, \ldots, L_2$. Notice that for the individual specific spatial coefficients $\boldsymbol{\psi}_i = (\psi_{0,i}, \ldots, \psi_{L_2,i})'$ to be identifiable, company i has to have non-zero number of neighbours. For unconnected i, we need to restrict their spatial related coefficients $\boldsymbol{\psi}_i = \mathbf{0}$. The error variance $\boldsymbol{\sigma}_{\upsilon^2} = var(\upsilon_{it})$ are allowed to differ for differnt i. W is the $N \times N$ adjacency matrix that specifies the channels from which shocks transmits. As a convention in spatial econometrics, the diagonal elements are set to zero ($w_{ii} = 0$ for all $i = 1, \ldots, N$), all other entries are assumed to be non-negative ($w_{ij} >= 0$) and the weights are row-normalized so that $\sum_j^N w_{ij} = 1$ for all $i = 1, \ldots, N$.

The model can be consistently estimated using the QML procedure proposed in Bailey et.al 2016[4] and Aquaro et.al 2019[2]. We collect all the parameters in the $N{*}(L_1{+}L_2{+}3)$ by 1 vector $\boldsymbol{\theta} = (\mathbf{a}_\epsilon{}', \boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_{L_1}, \boldsymbol{\Psi}_0', \ldots, \boldsymbol{\Psi}_{L_2}, \boldsymbol{\sigma}'_{\upsilon^2}$ and the log-likelihood function of (2) is written as follows:

$$\mathcal{L}_T(\boldsymbol{\theta}) = -\frac{NT}{2} ln(2\pi) - \frac{T}{2} \sum_i^N ln(\sigma_i^2) + \frac{T}{2} ln \mid \boldsymbol{S}'(\boldsymbol{\psi_0}) \boldsymbol{S}(\boldsymbol{\psi_0}) \mid - \frac{1}{2} \sum_{t=1}^T [\boldsymbol{S}(\boldsymbol{\psi_0}) \boldsymbol{y_t} - \boldsymbol{B x_t}]' \Sigma^{-1} [\boldsymbol{S}(\boldsymbol{\psi_0}) \boldsymbol{y_t} - \boldsymbol{B x_t}] \tag{3}$$

where $\boldsymbol{S}(\boldsymbol{\psi_0}) = I_N - \boldsymbol{\Psi}_0 W$, $\boldsymbol{y_t} = (y_{1t}, \ldots, y_{Nt})$. We stack the constant and all weakly exogeneous variables for i at t in $x_{it} = (1, \epsilon_{i,t-1}, \ldots, \epsilon_{i,t-L_1}, W\epsilon_{i,t-1}, \ldots, W\epsilon_{i,t-L_2})$ and $\boldsymbol{x_t} = (x_{1t}', \ldots, x_{Nt}')'$ is the $(1{+}L_1{+}L_2)N$ by 1 vector. $\boldsymbol{B}$ is the N by $(1{+}L_1{+}L_2)N$ block diagonal matrix with elements $\boldsymbol{\beta}_i' = (a_{\epsilon_i}, \lambda_{1,i}, \ldots, \lambda_{L_1,i}, \psi_{1,i}, \ldots, \psi_{L_2,i})'$ on the main diagonal and zeros elsewhere. Finally, $Var(\boldsymbol{\upsilon}) = \Sigma$.

The quasi maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{QMLE}$ maximizes (3). The error terms need not to be Gaussian, but when it is, $\hat{\boldsymbol{\theta}}_{QMLE}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$. For further details of computationally cheaper estimation procedure and inference, one could read Aquaro et.al 2019[2].

3.2.2 Spatial-temporal Responses to Local Risk

The spatial-temporal framework allows us to analyse a complicated diffusion pattern of local shocks over time and space. The parameter estimates of equation (2) only shows a part of the picture. To fully understand how $\epsilon_{i,t}$, a local shock arising from firm i at time t affects $\epsilon_{j,t+h}$, one need to trace the time profile of shocks over time and space. To examine the dependence across time and space implied by (2), we first rewrite it in a vector autoregression (VAR) form that we are familiar with:

$$\boldsymbol{\epsilon_t} = \sum_{\tau=1}^{max\{L_1, L_2\}} \Phi_\tau \boldsymbol{\epsilon_{t-\tau}} + R\boldsymbol{\upsilon_t} \tag{4}$$

where $R = (I_N - \Psi_0 W)^{-1}$, $\Phi_\tau = R\Lambda_\tau + R\Psi_\tau W$ and the lag order depends on the maximum of AR lag order and spatial-temporal lag order. Under the assumptions that $E(\boldsymbol{e_t}) = \mathbf{0}, E(\boldsymbol{\upsilon_t \upsilon_t}') = \Sigma_\upsilon = \{\sigma_{ij}, i, j = 1, \ldots, N\}$, which is a positive definite matrix, $E(\boldsymbol{v_t}\boldsymbol{v_\tau}') = \mathbf{0}$ for $t \neq \tau$, and the stability of the process, the VAR can be as a vector moving average (VMA) process,

$$\boldsymbol{\epsilon_t} = \sum_{p=1}^{\infty} A_p \boldsymbol{\eta}_{t-p} \quad \text{for } t = 1, \ldots, T \tag{5}$$

where $\boldsymbol{\eta}_t = R\boldsymbol{v_t}$ and A_p can be obtained recursively by

$$A_p = \Phi_1 A_{p-1} + \Phi_2 A_{p-2} + \cdots + \Phi_m A_{p-m} \tag{6}$$

where $m = max\{L_1, L_2\}$ and $A_0 = I_N$. Given that $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ is a hypothetical primitive shock hitting the economy at t, the generalized impulse response function (Pesaran and Shin 1996[27], Koop et al. 1996[21]) at horizon h is written as

$$GI(h, \boldsymbol{\delta}, \Omega_{t-1}) = E(\boldsymbol{\epsilon}_{t+h} \mid \boldsymbol{v_t} = \boldsymbol{\delta}, \Omega_{t-1}) - E(\boldsymbol{\epsilon}_{t+h} \mid \Omega_{t-1}) = A_h R\boldsymbol{\delta} \tag{7}$$

The primitive idiosyncratic shock $\boldsymbol{v_t}$ is allowed to be correlated. To look at the effect of a shock to one firm's (say the kth firm) effect on the whole system, we integrate out the effects of all other primitive shocks using the historically observed distribution of $\boldsymbol{v_t}$. The generalized impulse response function of the effect of a primitive shock to firm k at time t on the system h period in the future is given by

$$GI(h, \delta_k, \Omega_{t-1}) = E(\boldsymbol{\epsilon}_{t+h} \mid v_{k,t} = \delta_k, \Omega_{t-1}) - E(\boldsymbol{\epsilon}_{t+h} \mid \Omega_{t-1}) = \delta_k A_h R(\frac{\Sigma_v \boldsymbol{e_k}}{\sigma_{kk}}) \tag{8}$$

where $\boldsymbol{e_k}$ is a $N \times 1$ selection vector with 1 as its kth element and zeros elsewhere. $\frac{\Sigma_v \boldsymbol{e_k}}{\sigma_{kk}}$ is the adjustment due to potentially correlated primitive shocks $\boldsymbol{v_t}$. When Σ_v is a diagnoal matrix, $\frac{\Sigma_v \boldsymbol{e_k}}{\sigma_{kk}} = \boldsymbol{e_k}$, and $GI(h, \delta_k, \Omega_{t-1}) = \delta_k A_h R\boldsymbol{e_k}$.

4 Full Sample Estimation

In this section, I estimate the local risk spillovers in the weakly correlated idiosyncratic returns using the heterogeneous spatial-temporal model (2) discussed above. Using the estimated parameters, then I compute the spatial-temporal responses and construct the spatial-temporal spillover matrix D_h for each horizon h. Based on the spatial-temporal spillover matrices, we are able to find important systemic risk contributors and receivers. In the end, to assess the performance of the proposed method, I compare the in-sample and out-of-sample mean squared error (MSE) of the spatial-temporal model (2) estimated using alternative W and the high-dimensional vector autoregressive (VAR) model from Barigozzi and Hallin (2017).

4.1 De-factored (Idiosyncratic) Returns

Our full sample spans from 03/01/2006 to 31/12/2013 ($T = 2014$ days). To obtain a balanced panel, we end up with $N = 413$ stocks. We first estimate the hierachical factor model (1) by running time series regression for each company $i = 1, \ldots, N$. I consider 5 factors proposed by Fama-French (2015)[15] plus the momentum factor. And as for the industry factors $\mathbf{f}_{g,t}$, I use the Fama-French 5 industry porfolios. To make sure the number of members is large within each industry group g (for the construction of industry factors and the consistency of industrial mean group (MG) estimator, which we will elaborate more in later section), we consider 6 broad industry categories. The industry classification is built on top of Fama and French 5 industry portfolios where they classify all stocks into 5 groups 'Consumer', 'Health', 'Hitech', 'Manufacturing' and 'Others' based on Standard Industrial Classification (SIC) code. For the first four categories, I keep the same definitions as Fama and French. To address the importance of financial sectors, I categorize the companies with SIC code in the range $6000 - 6799$ as 'Finance'. For the firms previously in the 'Others' group with SIC code outside this range, I keep them in the 'Others' group. The SIC code is sourced from Compustat[8]. Statistical factor model (the hierarchical PCA) is used as a robustness check.

Table 1 summarizes the share of variance explained by the factors (regression R^2) for N cross sections. The R^2 varies from 13.2% to as high as 77.2%. On average, these factors explain 49.1% of the variation of the excess returns of $S\&P500$ stocks and the R^2 is higher than 40% for three-fourths of the stocks.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Hierarchical factor model	0.132	0.401	0.498	0.491	0.586	0.772

Table 1: Summary statistics for corss-sectional regression R^2 for the hierarchical factor model

The de-factored (idiosyncratic) returns can be obtained by estimating the above factor model (1) and collecting the residuals $\hat{\epsilon}_{it}$. Below I list some stylized features of the de-factored returns, which motivates our choice of the spatial-temporal modelling approach. First of all, $\hat{\epsilon}_{it}$ is serially correlated for around half of the sample $S\&P500$ stocks. Table 2 shows the summary statistics of the Q_m statistics of the Ljung-Box test and the corresponding P-value for the sample stocks. our lag orders are considered. $m = 1, 5, 10, 22$ corresponds to the number of trading days in a day, a week, two weeks and a month, respectively. If we consider the significance level $\alpha = 0.05$, then we reject the hypothesis of white noise for half of the stocks in the sample for all the three lag orders except $m = 1$, as the P-value for the median is smaller than 0.05 for $m = 5, 10, 22$. This results shows that there are predictability in terms of estimated idiosyncratic returns. However, examining the estimates of the correlation coefficients (I will not report here) shows that correlations are in general very small economically, rendering the predictability unprofitable given the trading cost.

[8] Guenther and Rosman(1994) and Kahle and Walkling(1996) found that the two-digit SIC codes between CRSP data and Compustat differ on average 38% of the time, and using Compustat SIC codes yields higher returns.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
m=1	Q_m	0.001	0.451	2.512	5.806	6.050	104.088
	P-value	0.000	0.014	0.113	0.272	0.502	0.985
m=5	Q_m	1.000	6.501	11.183	15.864	18.014	130.446
	P-value	0.000	0.003	0.048	0.174	0.260	0.962
m=10	Q_m	2.951	13.122	19.308	25.598	27.570	189.538
	P-value	0.000	0.002	0.037	0.151	0.217	0.983
m=22	Q_m	11.48	28.19	37.67	47.81	52.30	273.00
	P-value	0.000	0.000	0.019	0.126	0.169	0.967

Table 2: Summary statistics of the Q_m statistics of the Ljung-Box test and the corresponding P-value for the sample stocks. Note: m is the lag order of the test. $Q_m = T(T-1)\sum_{j=1}^{m}\frac{1}{T-j}\hat{\rho_j}^2 \sim \chi^2_m$.

To choose the lag order L_1 for the autoregressive term, one could apply information criterion such as Akaike information criterion (AIC), Bayesian information criterion (BIC), etc. In this paper, since the estimation of a large heterogeneous spatial temporal model is time consuming, and applying model selection techniques on the full model (2) would be theoretically possible but computationally burdensome. Thus, I pre-select the lag order L_1 of the model by examining the maximum number of lags included in the autoregressive (AR) model for each individual stock. I select the optimal number of autoregressive lags for each stock i using BIC criterion since AIC criterion usually selects a bigger model than BIC, and we hope to keep the model parsimonious given that the number of parameters need to be estimated [9] is $N * (L_1 + L_2 + 3)$. Among all sample stocks, 95% of them have optimal lag order smaller or equal to 5, and according to that, I pre-specify $L_1 = 5$, that is, the number of trading days in a week. The spatial temporal part is specified to have the same lag order $L_2 = 5$, and according to the estimation results they are sufficient to capture the spatial-temporal relationships.

In addition to the temporal correlation, the cross-sectional dependencies in the idiosyncratic returns are of major interest. The defactoring process removes the strong cross sectional dependence by reduceing the average pairwise correlation coefficient from as large as $\hat{\bar{\rho_N}} = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\hat{\rho}_{ij}$ to as small as $\hat{\bar{\rho}}_{N,r} = 0.4308$ to $\hat{\bar{\rho}}_{N,\epsilon} = 0.008$. Then I go on to test the null of cross-sectionally uncorrelated idiosyncratic returns $H_0 = E(\epsilon_{it}, \epsilon_{jt}) = 0$ for all t and $i \neq j$. I compute a scaled version of the Breusch and Pagan (1980)[10] LM test statistics, which has asymptotically standard normal distribution when N and T are both large.

$$CD_{LM} = \sqrt{\frac{1}{N(N-1)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(T\hat{\rho}_{ij}^2 - 1) \tag{9}$$

Using $\hat{\epsilon}_{it}$ estimated from equation (1), $CD_{LM} = 1653.40$, which strongly reject the null that idiosyncratic returns are cross-sectionally uncorrelated. Consistent with arbitrage pricing theory (APT), the interconnectedness in stock idiosyncratic returns, although being weak, is non-negligible and needs to be accounted for. Spatial models are natural tools for addressing local dependencies among neighbouring units, and with the novel business links constructed using text analysis, we can model the channel from which local risks transmits and quantity its strength.

[9] For each i, there are $L1$ AR parameters, $(L2 + 1)$ spatial temporal parameters, 1 intercept parameter and 1 scale parameter.

4.2 Adjacency Matrix

For the full sample estimation, W contains all the links that are identified within the sample period. Here is how we calculate W: (1) we firstly add up all the monthly observed adjacency matrix $W_1 + \cdots + W_t + \cdots + W_T$ to get a non-normalized adjacency matrix W_{raw}. Given there are T_m months in the sample, for the t th $(1 \leq t \leq T_m)$ month the sample, $W_t = (w_{ij,t})$ with $w_{ij,t}$ being a 1/0 dummy indicating whether company i and j are co-mentioned in the news published in this month. (2) we row-normalize W_{raw} and get W, as a convention in spatial econometrics. Notice that W_t is a unweighted adjacency matrix, while on the other hand W is weighted. This is because news tend to report the development of one issue for consecutive days and we may thus observe two companies get co-mentioned several times within that period. This 'multiple co-mentions within a short period of time' does not imply the relationship between two companies is stronger. However, if two companies get co-mentioned consistently in different monthly windows, there is reason to believe their links are stronger or the public are more aware/pay more attention to their links. That is why we add up unweighted monthly W_t and then apply row normalization to get weighted W. Alternative specifications of adjacency matrix are considered in later sections.

4.3 Spatial-temporal Model Estimation Results

4.3.1 Parameter Estimates

Equation (2) is estimated using quasi maximum likelihood (QML) and it is assumed that $e_{it} \sim IID(0, \sigma_i^2)$, for $i = 1, \ldots, N$. Since it is a heterogeneous coefficient model, we can only identify the spatial coefficients of those units with at least one link given T is large enough. We need to restrict the spatial parameters of the companies without any links to be zero. If we apply the full sample adjacency matrix W discussed above, only 5 out of $N = 413$ companies don't have any links.

Given the huge amount of parameters in the model, here I only report some summary statistics of the estimates in Table 3. Full estimation results can be requested from the author. Given a heterogeneous coefficient panel model, what is often of the interest to empirical researchers is the average estimates across all entities (or all entities within a sub-group). Assuming individual specific coefficients are randomly distributed around their common means as follows:

$$\lambda_{k1,i} = \lambda_{k1,0} + \zeta_{k1,i}, \psi_{k2,i} = \psi_{k2,0} + \varsigma_{k2,i} \text{ for } k1 = 1, \ldots, L_1,\ k2 = 1, \ldots, L_2 \text{ and } i = 1, \ldots, N \tag{10}$$

The common mean parameters $\lambda_{k1,0}$ and $\psi_{k2,0}$ for $k1 = 1, \ldots, L_1$, $k2 = 1, \ldots, L_2$ are the the objects of interest and they can be consistently[10] estimated with the following mean group (MG) estimator given N and T are large enough. The mean group (MG) estimates are provided in Table 3 with standard errors in the parenthesis.

$$\hat{\lambda}_{k1,0}^{MG} = \frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}_{k1,i} \ \ and \ \ \hat{\psi}_{k2,0}^{MG} = \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{k2,i} \tag{11}$$

[10] see Pesaran and Smith (1995)[28] for proofs the consistency when individual specific coefficients are independently distributed and the recent development by Chudik and Pesaran (2019)[11] who prove the consistency under weakly correlated individual specific estimators. In both cases, T and N are required to be big enough. Intuitively, big T is required for the consistent estimation of individual specific coefficients and N needs to be big enough for the consistent estimation of the means. To see how the MG estimators behave in the context of heterogeneous spatial-temporal model, see Aquaro (2019)[2].

	(1)AR terms					**(2) spatial-temporal terms**						**(3) σ**
	λ_1	λ_2	λ_3	λ_4	λ_5	ψ_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	σ
Median	-0.028	-0.012	-0.013	-0.008	-0.004	0.275	0.026	0.011	-0.005	0.009	0.005	1.426
MG estimates	-0.027	-0.013	-0.015	-0.010	-0.006	0.307	0.037	0.009	-0.008	0.008	0.011	1.517
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.021)	(0.007)	(0.006)	(0.005)	(0.005)	(0.006)	(0.027)
% sig (at 5%)	39.7%	23.0%	19.9%	19.1%	19.6%	81.4%	21.6%	20.3%	16.2%	17.2%	14.7%	-
non-zero coef.	413	413	413	413	413	408	408	408	408	408	408	413

Table 3: **QML estimation results of heterogeneous spatial-temporal model (2) using full sample.** Note: The median and mean group (MG) estimates are computed using unrestricted parameters only. The standard error of the MG estimates are in the parenthesis. The second last row show the percentage of significant parameters at 5% out of the unrestricted parameters and the last row of the table shows the number of unrestricted entities out of N. Panel (1), (2), (3) report the results of autoregressive parameters, spatial-temporal parameters and standard deviation of error, respectively.

From Table 3, we can see that the mean group (MG) estimates of contemporaneous spatial coefficients ($\psi_{0,0}$) and the first spatial-temporal coefficents ($\psi_{1,0}$) are both highly significant at the 5% level. And $\hat{\psi}_0^{MG} = 0.307(0.021)$ shows the strength of local dependence is big. Some general conclusions can be drawn here. After removing the common risk factors and industry risk factors, there is still a considerable degree of local spatial-temporal risk spillover among $S\&P500$ firms.

It is reasonable to suspect that the sensitivities to local risk spillovers are different for different industry groups. Given that the consistency of mean group (MG) estimator requires large N, one consideration when doing industry classification is that the number of members of each industry group need to be sufficiently big. Thus I adopt broad the industry classification scheme described in section 4.1, which guarantees large N condition to be satisfied for each industry group. Table 4 presents the estimation results grouped by industry and it reveals the considerable level of heterogeneity among different industry groups. In particular, the size of mean contemporaneous spatial effect ($\psi_{0,0}$) for manufacturing firms is largest, with the MG estimates $\hat{\psi}_{0,manufacturing,0}^{MG} = 0.446(0.033)$. Manufacturing firms are highly connected with other firms via supplier-customer linkages, and it is well documented (see Cohen and Frazzini (2008)[12]) that any shock to one firm has sizeable effect on its linked partner along the supply chain. Financial firms are also exposed to quite large mean contemporaneous spatial effect with the MG estimates $\hat{\psi}_{0,finance,0}^{MG} = 0.345(0.039)$. Apart from the large contemporaneous spatial coefficient, it is also worth noticing that the lead-lag effect in the risk spillovers for the financial firms is more pronounced as the the percentage of significant spatial-temporal coefficients $\hat{\psi}_{k_2,finance,i}$ is about twice as large as that of other industry groups for any lag order $k2 = 1, \ldots, 5$. We need to interpret the mean group estimates of these spatial-temporal parameters with care. The individual parameters $\hat{\psi}_{k_2,finance,i}$ are quite dispersed, with some firms having significantly positive spatial temporal terms and some having significantly negative ones. That is why the mean group estimates $\hat{\psi}_{k2,finance,0}$ don't look very significant although individually they are pretty significant —there are too much heterogeneity! Firms from consumer industry and hitech industry are also significantly exposed to their economic neighbours' local risks, although with slightly smaller sensitivities. Health firms are least sensitive to shocks elsewhere and the mean group estimate $\hat{\psi}_{0,health,0}^{MG} = 0.061(0.061)$ is not statistically significant. However, we should interpret that result with care since the number of companies from health industry is relatively small thus the mean group estimate is likely to be imprecise.

	(1)AR terms					(2) spatial-temporal terms						(3) σ
	λ_1	λ_2	λ_3	λ_4	λ_5	ψ_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	σ
Panel A: Consumer												
Median	-0.020	-0.015	-0.009	-0.009	-0.008	0.236	0.035	0.013	-0.004	0.001	0.001	1.373
MG Estimates	-0.025	-0.015	-0.011	-0.012	-0.008	0.232	0.033	0.026	-0.001	-0.002	0.005	1.456
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.039)	(0.009)	(0.011)	(0.010)	(0.010)	(0.012)	(0.054)
% Sig(at 5%)	29.9%	15.6%	16.9%	18.2%	18.2%	79.2%	15.6%	11.7%	14.3%	11.7%	7.8%	-
Non-zero coef.	77	77	77	77	77	77	77	77	77	77	77	77
Panel B: Finance												
Median	-0.037	-0.013	-0.017	-0.014	-0.005	0.350	0.017	0.000	-0.018	0.001	0.033	1.616
MG Estimates	-0.039	-0.020	-0.021	-0.024	-0.005	0.345	0.056	-0.010	-0.018	0.023	0.050	1.785
	(0.008)	(0.006)	(0.005)	(0.005)	(0.005)	(0.057)	(0.026)	(0.019)	(0.020)	(0.017)	(0.017)	(0.073)
% Sig(at 5%)	57.3%	38.7%	36.0%	32.0%	33.3%	82.7%	32.0%	34.7%	30.7%	30.7%	29.7%	-
Non-zero coef.	75	75	75	75	75	74	74	74	74	74	74	75
Panel C: Health												
Median	-0.007	-0.010	-0.004	0.001	0.008	0.119	0.024	-0.004	0.014	0.027	0.004	1.368
MG Estimates	-0.014	-0.005	-0.010	-0.005	0.006	0.061	0.020	0.001	-0.001	0.029	0.041	1.479
	(0.006)	(0.005)	(0.005)	(0.005)	(0.004)	(0.061)	(0.016)	(0.015)	(0.013)	(0.016)	(0.020)	(0.105)
% Sig(at 5%)	25.7%	20.0%	14.3%	14.3%	8.6%	68.6%	14.3%	11.4%	8.6%	5.7%	14.7%	-
Non-zero coef.	35	35	35	35	35	34	34	34	34	34	34	35
Panel D: Hitech												
Median	-0.036	-0.019	-0.014	-0.009	-0.004	0.212	0.016	-0.004	0.006	0.009	-0.013	1.459
MG Estimates	-0.032	-0.018	-0.012	-0.010	-0.007	0.229	0.018	-0.004	-0.001	0.004	-0.014	1.576
	(0.005)	(0.003)	(0.003)	(0.003)	(0.003)	(0.048)	(0.011)	(0.013)	(0.009)	(0.010)	(0.014)	(0.062)
% Sig(at 5%)	49.3%	19.2%	8.2%	13.7%	16.4%	72.6%	11.0%	13.7%	6.8%	12.3%	11.0%	-
Non-zero coef.	73	73	73	73	73	73	73	73	73	73	73	73
Panel E: Manufacturing												
Median	-0.011	-0.004	-0.018	-0.001	-0.005	0.468	0.022	0.028	-0.011	0.000	0.005	1.249
MG Estimates	-0.019	-0.005	-0.017	-0.002	-0.010	0.446	0.032	0.018	-0.008	0.004	0.005	1.303
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.033)	(0.009)	(0.008)	(0.008)	(0.007)	(0.007)	(0.041)
% Sig(at 5%)	30.0%	20.9%	17.3%	20.9%	20.0%	85.5%	18.2%	18.2%	13.6%	16.4%	13.0%	-
Non-zero coef.	110	110	110	110	110	108	108	108	108	108	108	110
Panel F: Other												
Median	-0.036	-0.015	-0.013	-0.007	-0.002	0.227	0.045	-0.017	-0.005	0.013	-0.017	1.488
MG Estimates	-0.031	-0.016	-0.020	-0.007	-0.001	0.315	0.072	0.010	-0.019	-0.002	-0.007	1.635
	(0.007)	(0.005)	(0.005)	(0.004)	(0.004)	(0.075)	(0.024)	(0.019)	(0.016)	(0.017)	(0.017)	(0.083)
% Sig(at 5%)	46.5%	23.3%	27.9%	7.0%	11.6%	76.7%	32.6%	20.9%	9.3%	9.3%	11.9%	-
Non-zero coef.	43	43	43	43	43	42	42	42	42	42	42	43

Table 4: **QML estimation results of heterogeneous spatial-temporal model (2) using full sample, parameters summarized by industry.**

4.3.2 Spatial-temporal Responses to Local Shocks

For any horizon h, we can summarize the own response and cross-response implied by equation (8) in a similar way as how Lesage and Pace (2009)[24] and LeSage and Chih (2016)[25] summarize direct and indirect partial effects of a change in the kth explanatory variable. For illustration, consider a simple example where Σ_v is diagonal, and firm k receives a unit shock at time t, equation (8) can be simplified as $GI(h, \delta_k = 1, \Omega_{t-1}) = A_h R \boldsymbol{e_k}$. $A_h R$ is a $N \times N$ matrix with N own responses and $N(N-1)$ cross-responses at horizon h on the diagonal and off-diagonal, respectively. For $h = 0, A_0 = I_N$,

$$A_h R = R = (I_N - \Psi_0 W)^{-1} = I_N + \Psi_0 W + \Psi_0^2 W^2 + \Psi_0^3 W^3 + \dots \tag{12}$$

R is an infinite series expansion that adds the own effect I_N, first order neighbour effect $\Psi_0 W$, second order neighbour effect $\Psi_0^2 W^2$ and so on. Ψ_0 is a diagonal matrix that every entry is upper-bounded by 1 in absolute value, so that higher powers of Ψ_0 assigns smaller impact to higher order neighbours. The main diagonal elements of R gives the own responses to a unit shock, which is in general different from 1 since they are the sums of own effects and feed backs from others. The off-diagonal elements of R, on the other hand, are the sums of neighbour effect of different orders. For $h >= 1, A_h R$ gives the combination effects spatial dependence and temporal dependence. In general, when Σ_v is not diagonal, we need to adjust for correlated $\boldsymbol{v_t}$ using equation (8).

For each horizon h, I compute the $N \times 1$ vector $GI(h, \delta_k = 1, \Omega_{t-1})$ for each $k = 1, \ldots, N$ using the estimated parameters. For the diagonal matrices $\Lambda_k, k = 1, \ldots, 5$, the i's diagonal element is $\lambda_{k,i}$ if it is statistically significant at 5% level, otherwise it is replaced by zero. The same is true for the construction of $\Psi_k, k = 1, \ldots, 5$. We denote the spatial-temporal spillover matrix at h as D_h, where $GI(h, \delta_k = 1, \Omega_{t-1})$ is the kth column of it. $D_h = [d_{ij}^h]$ gives the pairwise directional spillovers at horizon h.

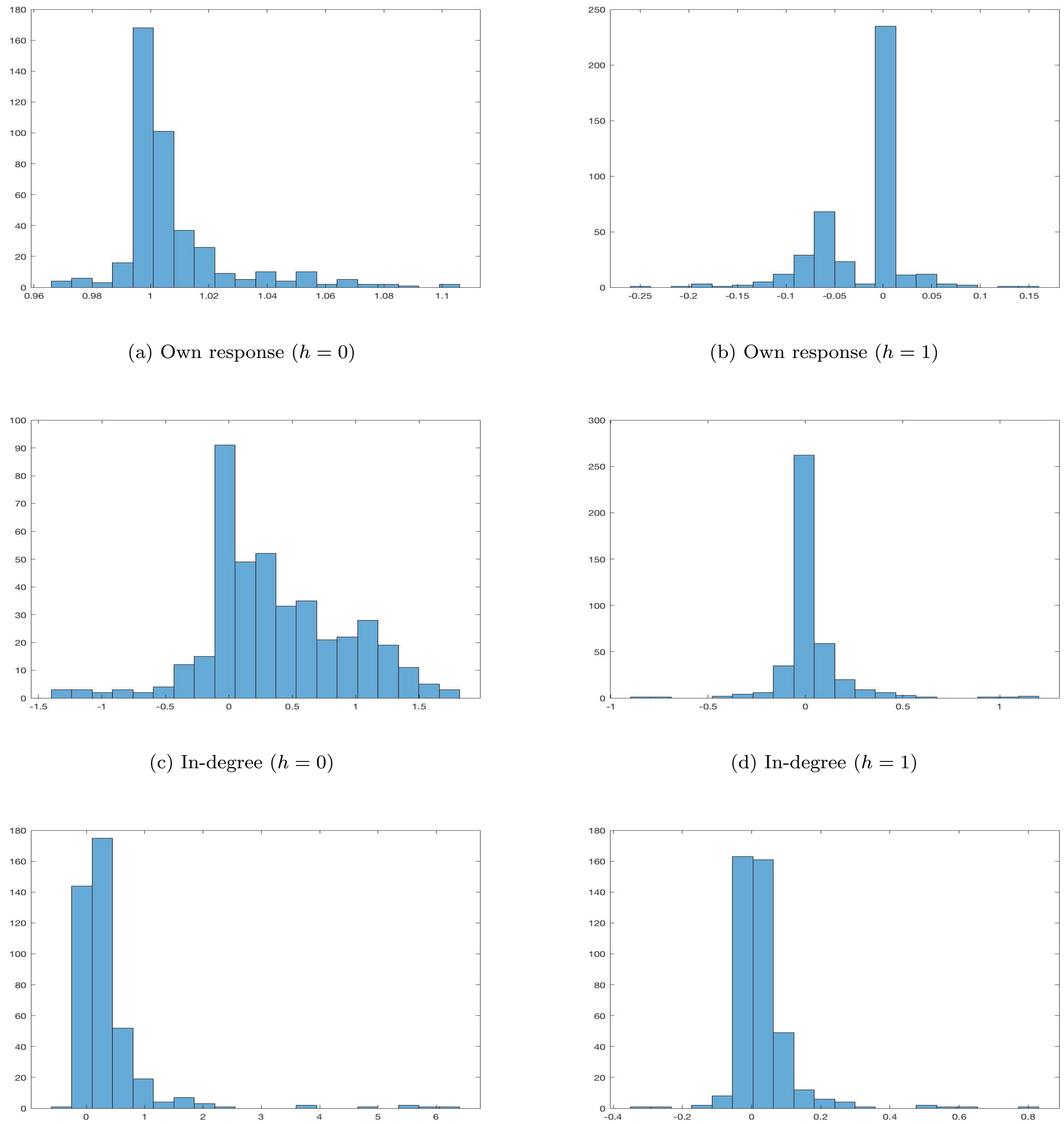
Since N is large in our analysis, it is infeasible to report spillovers at pairwise level, I adopt the scalar summary measure used in Lesage and Pace (2009)[24] and LeSage and Chih (2016)[25]. For each horizon h, I derive individual level own response, which is the diagonal elements of D_h. As for the individual level indirect effect, two measures are used, which are in-degree (C_{in}^h) and out-degree(C_{out}^h). They are defined as follows:

$$C_{i,in}^h = \sum_{j \neq i}^N d_{ij}^h \tag{13}$$

$$C_{j,out}^h = \sum_{i \neq j}^N d_{ij}^h \tag{14}$$

The in-degree measures the shocks a firm receives from other firms, and the out-degree, on the other hand, measures the shocks a firm spreads to others.

Figure 1 plots the histogram for own response, in-degree and out-degree at horizon $h = 0, 1$. The figures for further horizon are at the Figure 10. The two sub-figures at the first row correspond to the contemporaneous responds. When a firm receives one unit primitive shock at t, its contemporaneous own response it not necessarily 1 as the result of the complicated feedback relationships. There are stark differences between two indirect effect measures if we compare the two graphs on the second row with the two graphs on the third row. For $h = 0$, while there are a non-negligible proportion of firms respond negatively to neighbours' shocks, almost all firms are positive spreader of risks (in graph (e), there is only a tiny bin with negative out-degree). Also, it is worth noticing that the out-degree has heavy right tail, with some companies contributing a lot of risk to the system. The right column with $h >= 1$ corresponds to the dynamic responses, which combine the effects of both temporal and spatial dependencies. From Table 3 and Table 4, we can see the estimates of dynamic parameters (both the pure temporal and the spatial-temporal) are small relative to the contemporaneous spatial effect parameter, and this is reflected in Figure 1. Local shocks travel over time and space with decays. One interesting feature is that the decay along the spatial dimension is slower than that along the time dimension. Notice that the current analysis focus on how a unit shock to one firm affects the system, and that explains why the shocks die out quickly. This does contradict with financial crisis scenario where a larger number of firms receive negative shock jointly, which could result in a much slower shock decay.



(a) Own response ($h = 0$)

(b) Own response ($h = 1$)

(c) In-degree ($h = 0$)

(d) In-degree ($h = 1$)

(e) Out-degree ($h = 0$)

(f) Out-degree ($h = 1$)

Figure 1: Histogram for own response, in-degree and out-degree at horizon $h = 0, 1$.

Firms with high in-degree are vulnerable as they are particularly sensitive to shocks elsewhere and firms with high out-degree are dangerous since their 'own' primitive shocks are widespread. Therefore, it is of interest from a microprudential (firm-specific) perspective to identify these two types of firms. Table 5 shows the 20 firms with highest in-degree and out-degree for $h = 0, 1$. Higher order results are not shown in the main text since the shocks decay along time dimension quickly. The firms contribute the most to the systemic risks are the large cap financial institutions and manufacturers, and the findings are highly in line with the systemic risk contributors found in others including Hautsch et al. (2015)[19], Barigozzi and Hallin (2017)[9]. Apart from systemic risk contributors, companies that are particularly sensitive to others' shocks are also found. It is worth noticing that the well-connected systemic risk contributors themselves are not necessarily the major risk receivers. They are the periphery firms that receive a lot of risks from the core.

		Company Ticker
h=0	In-degree	LEN, DUK, EIX, PCG, GD, RTN, STI, ETR, NOC, RIG, HBAN, DHI, CVX, CI, LRCX, CSX, SO, UNH, APA, VLO, FITB
	Out-degree	GE, JPM, MSFT, C, GS, BAC, WFC, PG, XOM, BA, LMT, DUK, INTC, CVX, KO, HPQ, EXC, COP, BK, ORCL
h=1	In-degree	GNW, HBAN, AIG, GE, CI, WY, COF, C, STT, LNC, GT, ATI, TIF, HUM, PG, PBI, JPM, UNH, PRU, SWKS
	Out-degree	BAC, C, JPM, GS, GE, MSFT, APPL, DUK, BK, USB, INTC, PFE, JNJ, UNH, LNC, VZ, BA, AET, LM, FITB

Table 5: The 20 firms with highest in-degree and out-degree for $h = 0, 1$. Note: For $h = 1$, we rank the firms according to their absolute values of the in-degree and out-degree, given the individual spatial-temporal coefficients are very dispersed with some having significantly positive coefficients and having significantly negative ones.

4.4 Comparison with Alternative methods

To assess the performance, I compare the in-sample and out-of-sample mean squared error (MSE) of the spatial-temporal model (2) estimated using different adjacency matrix W and the high-dimensional vector autoregressive (VAR) model from Barigozzi and Hallin (2017). The first column is the benchmark Naive estimator where the predicted de-factored returns are zero all the time. The second column present the results of the high-dimensional vector autoregressive (VAR) model from Barigozzi and Hallin (2017), and we refer it as BH-VAR for short. Column three to column six present the results of the spatial-temporal model (2) estimated using 4 alternative adjacency matrices W. The first candidate W is the empty adjacency matrix where there is no links. The second candidate W is the industry network where within each industry, companies are completely connected, and there are no inter-industry links. The third candidate W is the compustat customer-supplier network. The fourth W is the news-based networks. The spatial-temporal model (2) allows cross-sections to have heterogeneous coefficients. While the highly flexible model promises better in-sample fit, some might suspect the model does not guarantee a better out of sample fit as a result of potential over-fitting. To examine the above issue, for each candidate W, I compute the in-sample and out-of-sample MSE using three alternative specifications, given by row (1)-(3) of each panel. The results of the Naive estimator and BH-VAR are shown in the row (4) of each panel. The training sample spans from 03/01/2006 to 31/12/2013 (2014 days) and the testing sample spans from 03/01/2014 to 31/12/2014 (252 days).

	Naive	BH-VAR	W_{empty}	$W_{industry}$	$W_{compustat}$	W_{news}
In Sample MSE						
(1)Heterogeneous coef	-	-	2.785	**2.759**	2.783	**2.685**
(2)Industrial-heterogeneous coef	-	-	2.806	2.768	2.806	2.764
(3)Homogeneous coef	-	-	2.804	2.771	2.804	2.766
(4)	2.812	**2.146**	-	-	-	-
Out-of-Sample MSE						
(1)Heterogeneous coef	-	-	1.331	1.322	1.330	**1.277**
(2)Industrial-heterogeneous coef	-	-	1.326	1.308	1.326	**1.290**
(3)Homogeneous coef	-	-	1.327	1.312	1.327	**1.298**
(4)	1.326	1.397	-	-	-	-

Table 6: In-sample and out-of-sample MSE (in basis point) of alternative models. Note: for each panel, the best 3 (smallest MSE) cases are in bold.

In terms of in-sample fit, BH-VAR has the smallest MSE. This is not surprising, given the method selects the model by minimizing a Bayesian information criterion. The heterogeneous coefficient spatial-temporal model with news-based network and industry network rank second and third, respectively. However, when we look at out-of-sample fit, BH-VAR loses its advantage with its MSE being even larger than that of the Naive predictor. The spatial-temporal model with news-based network, under any three parameter heterogeneity assumption, outperforms the rest of the specifications.

The strength of local risk spillovers via news-based linkages exhibits high level of heterogeneity. As a result, the heterogeneous coefficient specification improves not only in-sample fit but also out-of-sample fit. Although the spatial-temporal model with news-based network underperforms BH-VAR model in term of in-sample fit, but it beats BH-VAR when we compare out-of-sample fit, which is more important practically. It is also worth noticing that W_{news} beats all other alternative Ws in terms of both in-sample and out-of-sample fit. The out-performance of the news-based network over compustat customer-supplier network and the industrial network reflects the novel information reflects the additional information contained in the novel dataset.

5 Dynamic Estimation

Equity returns comovement reflects both strong and weak cross-sectional dependence. It has been documented that asset returns depart from fundamentals during times of financial crisis and stocks dis-connect from the market factor (see Baily et al. (2019[5], 2020[6])). Our two stage factor plus spatial approach captures both sources of co-movement separately and thus provides an avenue to examine how weak cross-sectional dependence evolve over time. In this section, I consider a rolling window analysis with 251-day (the average number of trading days in a year) rolling sample from 03/01/2006 to 31/12/2013. In total, there are 1761 windows.

5.1 Time Evolution of Weak Cross-sectional Dependence

The composition of $S\&P500$ index changes periodically in response to acquisitions and the growth or shrinkage of company values. We update the list of sample companies on a yearly basis and include the securities that stay in the $S\&P500$ list and have no missing observations for that year. On average, there are 447 stocks on the list for each update. Then we use a rolling estimation with a 251-day window to gauge the time variations in local dependencies. For estimation window $[t, t+251]$, we conduct the two-stage procedure, and W_t used for the estimation of spatial-temporal model is constructed using all the news published one year during the year. In the end, 1761 sets of estimates are obtained.

Figure 2 plot $\hat{\psi}_{0,t}^{MG}$, the 251-day rolling mean group estimates of the the contemporaneous spatial parameter. For the window $[t-125, t+126]$, the mean group estimate of the contemporaneous spatial parameter is calculated as $\hat{\psi}_{0,t}^{MG} = \frac{1}{N}\sum_{i=1}^{N} \hat{\psi}_{0,i,t}$. 1761 rolling samples from 03/01/2006 to 31/12/2013 give rise to 1761 sets of estimates from 30/06/2006 to 01/07/2013. The figure reveals the increase in the intensities of local dependencies during times of financial turmoils. $\hat{\psi}_{0,t}^{MG}$ was low in the 2006 and early 2007, it increased gradually since 2007 following the liquidity crisis. By the end of the year, the public started to aware that big US banks might write off a huge amount of losses and a global financial crisis is unfolding. $\hat{\psi}_{0,t}^{MG}$ skyrocketed afterwards, peaking around the Lehman bankruptcy. After months, with the massive direct capital injection by the US government, the market calmed down and $\hat{\psi}_{0,t}^{MG}$ gradually recovered to pre-crisis level. Instead of staying low, the several waves of European Debt Crisis raised $\hat{\psi}_{0,t}^{MG}$ again, although by smaller magnitude.

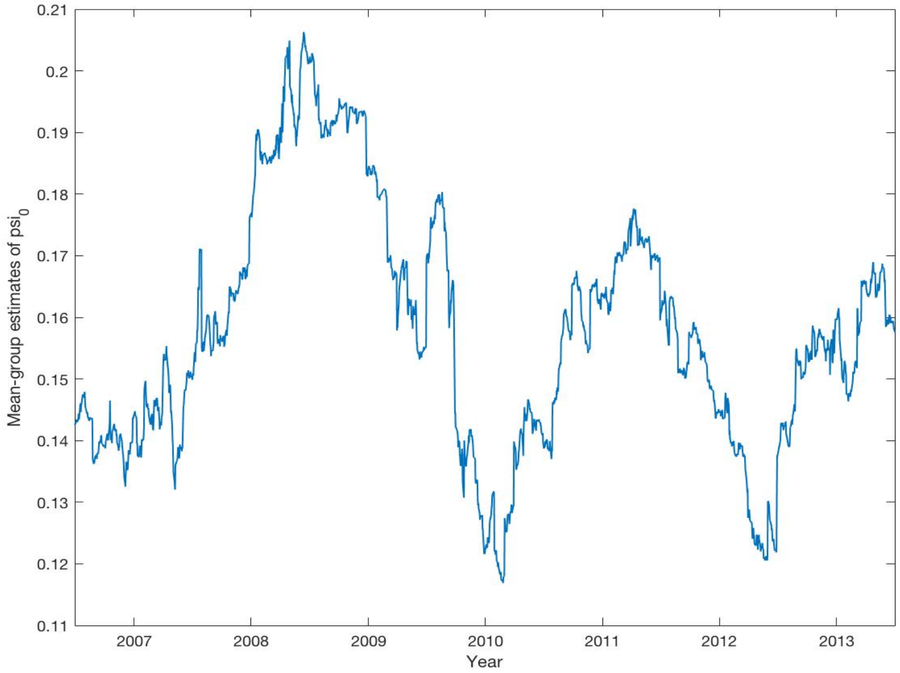
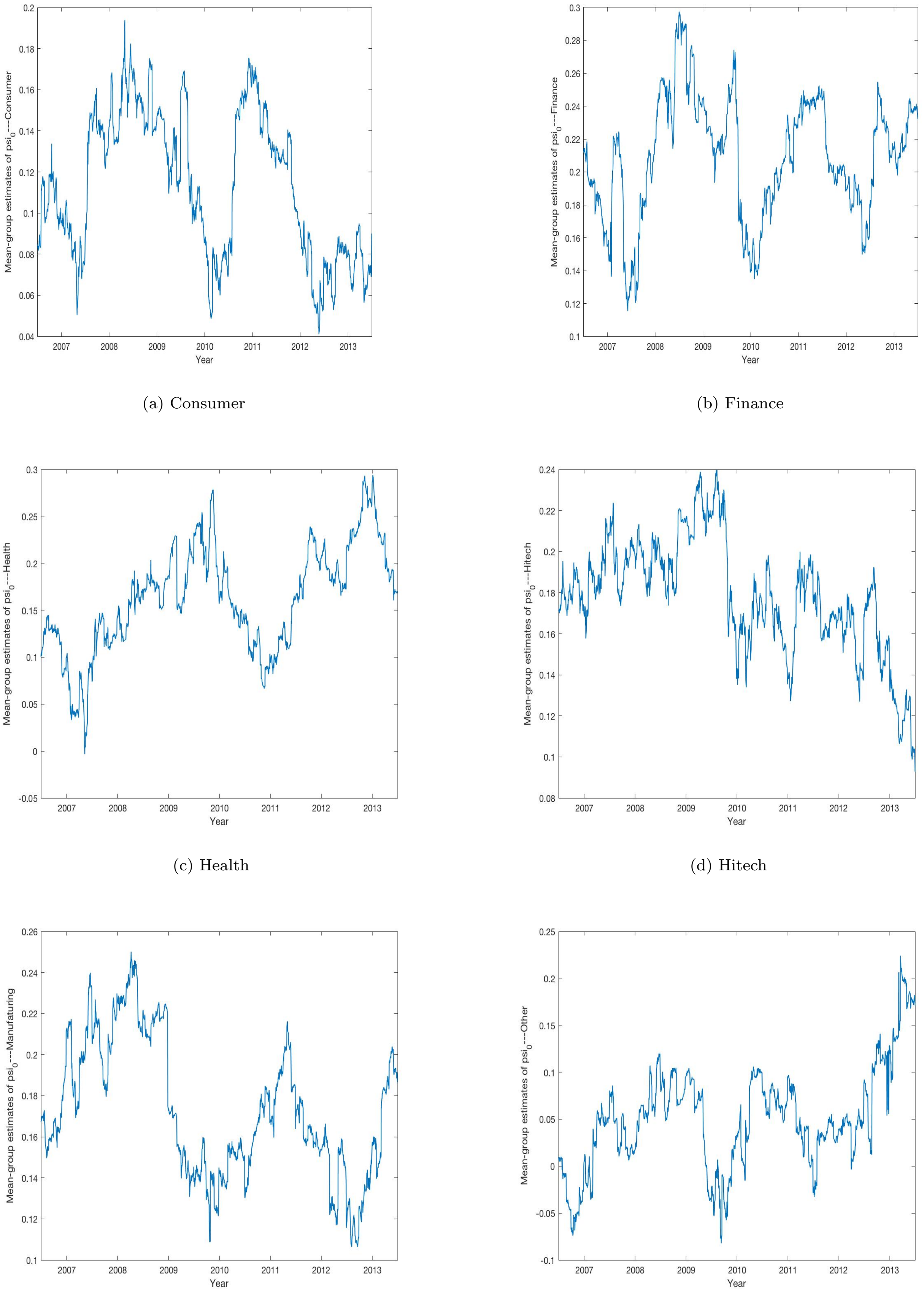


Figure 2: 251-day rolling $\psi_{0,t}^{MG}$ from 03/01/2006 to 31/12/2013. Note: for window $[t-125, t+126]$, we use the middle date of the window to denote the the mean group estimate $\hat{\psi}_{0,t}^{MG}$. That's why the x axis spans from 30/06/2006 to 01/07/2013.



(a) Consumer

(b) Finance

(c) Health

(d) Hitech

(e) Manufacturing

(f) Other

Figure 3: 251-day rolling $\psi_{0,g,t}^{MG}$ from 03/01/2006 to 31/12/2013 for different industry groups.

To examine the industrial heterogeneity of time variations in the strength of local spillovers, I plot the rolling mean group estimates of the contemporaneous spatial parameter for different industries in Figure 3. For each window, the mean group estimate of the contemporaneous spatial parameter for industry g is calculated as the sample average of individual specific contemporaneous spatial parameter from that industry at t, namely, $\hat{\psi}_{0,g,t}^{MG} = \frac{1}{N}\sum_{i\in g}\hat{\psi}_{0,i,t}$. The time series pattern of $\hat{\psi}_{0,g,t}^{MG}$ shows a considerable degree of heterogeneity. Table 7 presents the correlation coefficients matrix of $\hat{\psi}_{0,t}^{MG}$ and $\hat{\psi}_{0,g,t}^{MG}$ for $g =$ Consumer, Finance, Health, Manufacturing and other. $\hat{\psi}_{0,consumer,t}^{MG}$, $\hat{\psi}_{0,finance,t}^{MG}$ and $\hat{\psi}_{0,manufacturing,t}^{MG}$ exhibit similar pattern and they all have two obvious humps around the Great Financial Crisis and European Debt Crisis episodes. While hitech industry also experienced a rise in local risk spillovers during the Great Financial Crisis, it was not very affected by the European Debt Crisis. Health care stocks belong to the non-cyclical group and $\hat{\psi}_{0,health,t}^{MG}$ moves in opposite directions with others.

	$S\&P500$	Consumer	Finance	Health	Hitech	Manufacturing	Other
$S\&P500$	1	0.77	0.79	0.04	0.39	0.65	0.3
Consumer	0.77	1	0.52	-0.18	0.45	0.5	-0.11
Finance	0.79	0.52	1	0.18	0.14	0.25	0.28
Health	0.04	-0.18	0.18	1	-0.16	-0.48	0.21
Hitech	0.39	0.45	0.14	-0.16	1	0.21	-0.46
Manufacturing	0.65	0.5	0.25	-0.48	0.21	1	0.14
Other	0.3	-0.11	0.28	0.21	-0.46	0.14	1

Table 7: Correlation coefficients of $\hat{\psi}_{0,t}^{MG}$ and $\hat{\psi}_{0,g,t}^{MG}$ for $g =$Consumer, Finance, Health, Manufacturing and other.

5.2 Time Evolution of Market Factor Strength

While weak CSD intensifies during periods of financial crisis and turmoils, strong CSD, as documented in Baily et al. (2019[5], 2020[6]), loses its power. According to asset pricing theories like capital asset pricing model (CAPM), all stocks should load significantly on market factor. In the papers, they propose a estimator of factor strength based on the number of statistically significant factor loadings, taking account of the multiple testing problem. For the factor model with $\mathbf{f}_t = (f_{1t}, \ldots, f_{kt})$ being the vector of factors.

$$r_{it} - r_{ft} = \alpha_i + \mathbf{b}_i'\mathbf{f}_t + \epsilon_{it} \text{ for } i = 1, \ldots, N \tag{15}$$

Their proposed an estimator of the factor strength for the jth factor $\hat{\alpha_j}$, which is calculated as

$$\hat{\alpha_j} = 1 + \frac{log(\hat{D_j}/N)}{log(N)} \text{ if } \hat{D_j} > 0 \tag{16}$$

where $\hat{D_j}$ is the total number of statistically significant loadings of factor j out of N cross-sectional regressions. The critical value of the test is adjusted for the multiple testing problem.

According to capital asset pricing model (CAPM), the market factor is a strong factor and all stocks load significantly on market factor as the number of stocks N grows large. This implies the market factor should have $\alpha_{market} = 1$. I re-do their exercise and conduct a rolling estimation of α_{market}. Figure 4 plots the rolling estimate of the strength of market factor. The time series is more volatile than that in Baily et al. (2020)[6] since I am using daily 251-day rolling window while they are using monthly 10-year rolling window. As is found in their work, market factor is pretty strong with its strength being very close to 1 all the time except for a short period during the financial crisis. This result, together with the time series patterns of the local risk spillovers show that the strength of strong and weak CSD tend to move in opposite directions. The correlation coefficient of $\hat{\psi}_{0,t}^{MG}$ and $\hat{\alpha}_{market}$ is -0.6. When market factor loses its importance during the financial crisis, weak cross-sectional dependence gains its power with the strength of local risk spillovers becoming stronger.

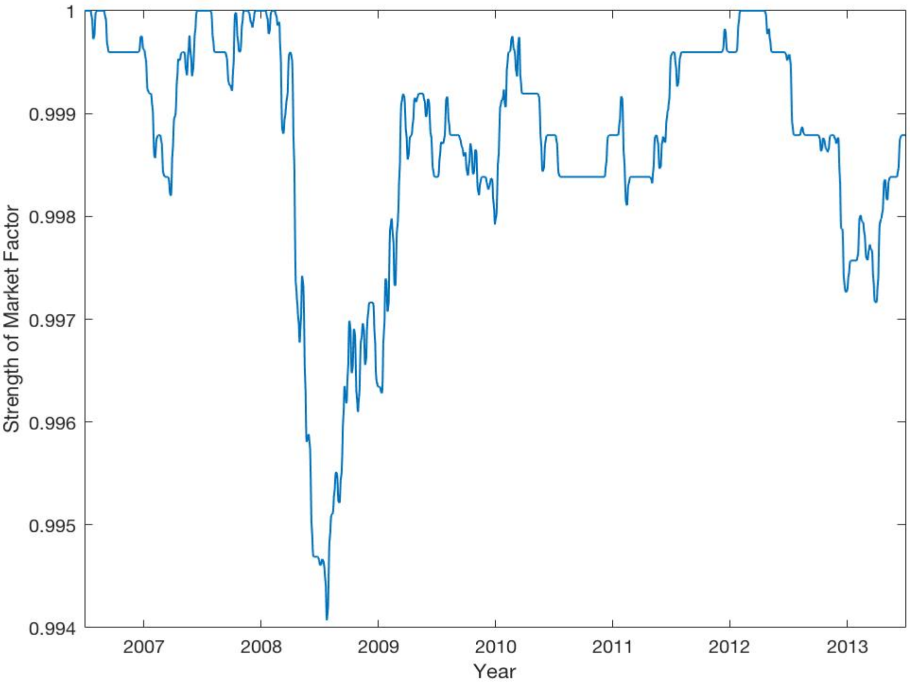


Figure 4: 251-day rolling of α_t from 03/01/2006 to 31/12/2013. Note: The factor strength parameter α is calculated as in Baily et al. (2020).

6 Robustness Check and Placebo Test

6.1 Robustness Check

In this section, I am going to test whether the results are sensitive to the the way we de-factor the panel of excess returns. As a robustness check, I de-factor by using unobserved factors instead of observed ones. A hierarchical principal components (PCA) procedure is applied, to remove both principal components at market level and industry level. Such a hierarchical can be written as:

$$r_{it} - r_{ft} = \alpha_i + \mathbf{b}_i'\mathbf{f}_t + \gamma_i'\mathbf{f}_{g,t} + \epsilon_{it} \tag{17}$$

where $\mathbf{f}_t$ is the vector of market factors that affect every stock and $\mathbf{f}_{g,t}$ is the vector of industrial factors that affect every stock in industry g. Applying the information criteria in Bai and Ng (2002)[3], we select the first 5 market principal components and 1 industry principal component. Then I run regression (2) and obtained the residuals $\hat{\epsilon}_{it}^{pca}$. Estimating the heterogeneous spatial-temporal model (2) using $\hat{\epsilon}_{it}^{pca}$, the results are close to the one presented in Table 3 and Table 4. The spatial-temporal estimates are slightly smaller, but the main conclusions remain valid. The results are given in the Table 14 and Table 15.

6.2 Placebo Test

In this section, I am going to conduct a placebo test by checking whether randomly generated would give rise to significant local dependencies. Our full sample news-based network has 6742 linked pairs out of $N * (N-1)/2 = 148785$ pairs of firms. So the linking probability is 4.5%. I generate 100 random graphs using $G(N,p)$ model, which is one version of the Erdős–Rényi (ER) random graph models. In the $G(N,p)$ model, a graph is constructed by connecting nodes randomly. There are N edges and each edge is included in the graph with probability p independent from every other edge. To have the same level of sparisty as our full sample news-based network, I let $p = 4.5\%$.

For each one of the randomly simulated E-R networks, I use it as the adjacency matrix W in equation (2) and estimate the spatial temporal-model. As expected, none of produce significant spatial parameters. The placebo test thus confirm that the text analysis approach does help us to identify the links among firms that are important for the transmissions of local shocks.

7 Conclusion

This paper investigates the local dependencies in idiosyncratic asset returns. Utilizing the novel text-based linkage data, I am able to construct the channels from which the local shocks transmits. I found that stocks linked via news paper co-mentioning exhibit excess comovement beyond that is predicted by standard asset pricing models. Local shocks travel over time and space, and the decay along spatial dimension is slower than that along time dimension. By analyzing the impulse responses, we are able to identify the major systemic risk contributors and receivers, which are of the interest to microprudential polices. From a macroprudential perspective, by separately addressing both strong and weak cross-sectional dependencies, I found that the strength of strong and weak CSD tend to move in opposite directions. When equities dis-connect from the market factor during the period of finanical turmoil, the strength of local risk spillovers becomes stronger.

The findings suggest text-based network as a promising alternative to existing network data. Our empirical studies show it is competitive in the modelling of local risk spillovers. The author believe that the text-based dataset can be applied to a wider context. For example, the modelling of the volatility spillovers and the use of text-based links as prior information in estimating links from large panel, etc.

References

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

[2] Michele Aquaro, Natalia Bailey, and M Hashem Pesaran. Estimation and inference for spatial models with heterogeneous coefficients: an application to us house prices. *USC-INET Research Paper*, (19-07), 2019.

[3] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[4] Natalia Bailey, Sean Holly, and M Hashem Pesaran. A two-stage approach to spatio-temporal analysis with strong and weak cross-sectional dependence. *Journal of Applied Econometrics*, 31(1):249–280, 2016.

[5] Natalia Bailey, George Kapetanios, and M Hashem Pesaran. Exponent of cross-sectional dependence for residuals. *Sankhya B*, 81(1):46–102, 2019.

[6] Natalia Bailey, George Kapetanios, and M Hashem Pesaran. Measurement of factor strenght: Theory and practice. 2020.

[7] Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

[8] Matteo Barigozzi and Christian Brownlees. Nets: Network estimation for time series. *Journal of Applied Econometrics*, 34(3):347–364, 2019.

[9] Matteo Barigozzi and Marc Hallin. A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):581–605, 2017.

[10] Trevor S Breusch and Adrian R Pagan. The lagrange multiplier test and its applications to model specification in econometrics. *The review of economic studies*, 47(1):239–253, 1980.

[11] Alexander Chudik and M Hashem Pesaran. Mean group estimation in presence of weakly cross-correlated estimators. *Economics Letters*, 175:101–105, 2019.

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

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

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

[15] Eugene F Fama and Kenneth R French. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

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

[17] Diego Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

[18] Galina Hale and Jose A Lopez. Monitoring banking system connectedness with big data. *Journal of Econometrics*, 212(1):203–220, 2019.

[19] Nikolaus Hautsch, Julia Schaumburg, and Melanie Schienle. Financial network systemic risk contributions. *Review of Finance*, 19(2):685–738, 2015.

[20] Gerard Hoberg and Gordon Phillips. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465, 2016.

[21] Gary Koop, M Hashem Pesaran, and Simon M Potter. Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, 74(1):119–147, 1996.

[22] Steven Kou, Xianhua Peng, and Haowen Zhong. Asset pricing with spatial interaction. *Management Science*, 64(5):2083–2101, 2018.

[23] Charles MC Lee, Paul Ma, and Charles CY Wang. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, 116(2):410–431, 2015.

[24] James P LeSage. An introduction to spatial econometrics. *Revue d'économie industrielle*, (123):19–44, 2008.

[25] James P LeSage and Yao-Yu Chih. Interpreting heterogeneous coefficient spatial autoregressive panel models. *Economics Letters*, 142:1–5, 2016.

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

[27] H Hashem Pesaran and Yongcheol Shin. Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29, 1998.

[28] M Hashem Pesaran and Ron Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of econometrics*, 68(1):79–113, 1995.

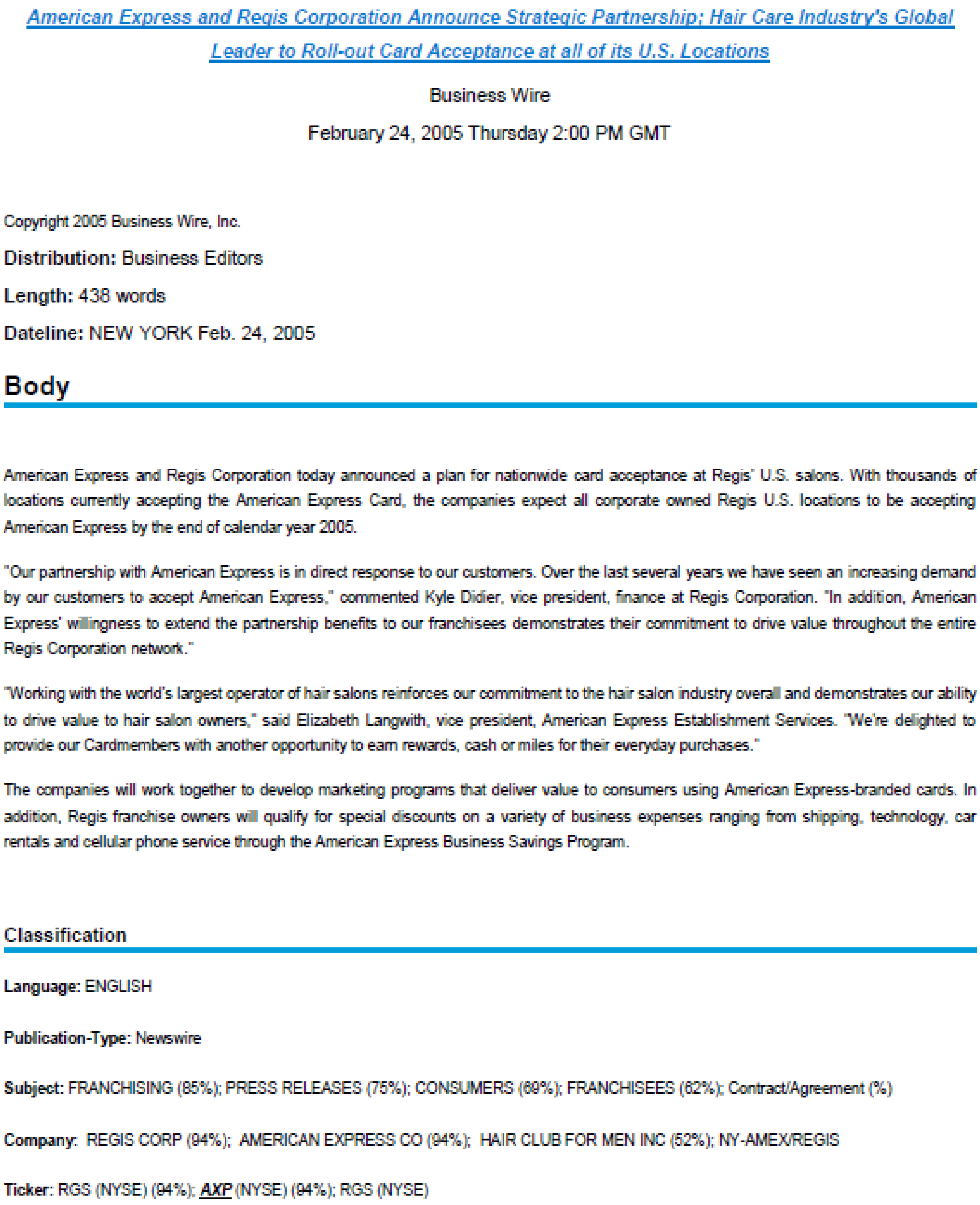
[29] S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011, 2012.

[30] Stephen Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

[31] Anna Scherbina and Bernd Schlusche. Economic linkages inferred from news stories and the predictability of stock returns. *Available at SSRN 2363436*, 2015.

[32] Gustavo Schwenkler and Hannan Zheng. The network of firms implied by the news. *Available at SSRN 3320859*, 2019.

[33] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.



American Express and Regis Corporation Announce Strategic Partnership; Hair Care Industry's Global Leader to Roll-out Card Acceptance at all of its U.S. Locations

Business Wire

February 24, 2005 Thursday 2:00 PM GMT

Copyright 2005 Business Wire, Inc.

Distribution: Business Editors

Length: 438 words

Dateline: NEW YORK Feb. 24, 2005

Body

American Express and Regis Corporation today announced a plan for nationwide card acceptance at Regis' U.S. salons. With thousands of locations currently accepting the American Express Card, the companies expect all corporate owned Regis U.S. locations to be accepting American Express by the end of calendar year 2005.

"Our partnership with American Express is in direct response to our customers. Over the last several years we have seen an increasing demand by our customers to accept American Express," commented Kyle Didier, vice president, finance at Regis Corporation. "In addition, American Express' willingness to extend the partnership benefits to our franchisees demonstrates their commitment to drive value throughout the entire Regis Corporation network."

"Working with the world's largest operator of hair salons reinforces our commitment to the hair salon industry overall and demonstrates our ability to drive value to hair salon owners," said Elizabeth Langwith, vice president, American Express Establishment Services. "We're delighted to provide our Cardmembers with another opportunity to earn rewards, cash or miles for their everyday purchases."

The companies will work together to develop marketing programs that deliver value to consumers using American Express-branded cards. In addition, Regis franchise owners will qualify for special discounts on a variety of business expenses ranging from shipping, technology, car rentals and cellular phone service through the American Express Business Savings Program.

Classification

Language: ENGLISH

Publication-Type: Newswire

Subject: FRANCHISING (85%); PRESS RELEASES (75%); CONSUMERS (69%); FRANCHISEES (62%); Contract/Agreement (%)

Company: REGIS CORP (94%); AMERICAN EXPRESS CO (94%); HAIR CLUB FOR MEN INC (52%); NY-AMEX/REGIS

Ticker: RGS (NYSE) (94%); ***AXP*** (NYSE) (94%); RGS (NYSE)

Figure 5: A typical business news in the dataset.

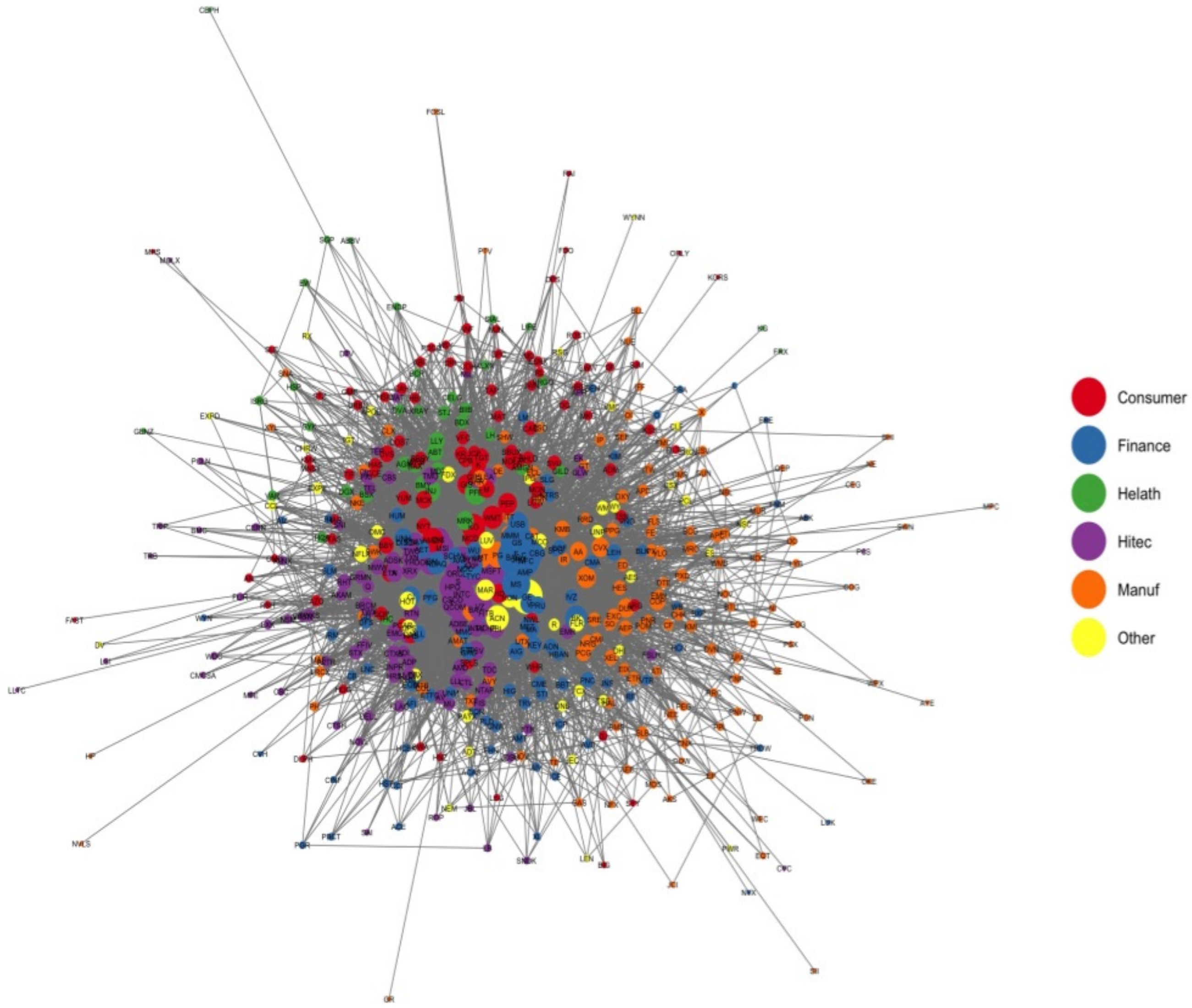


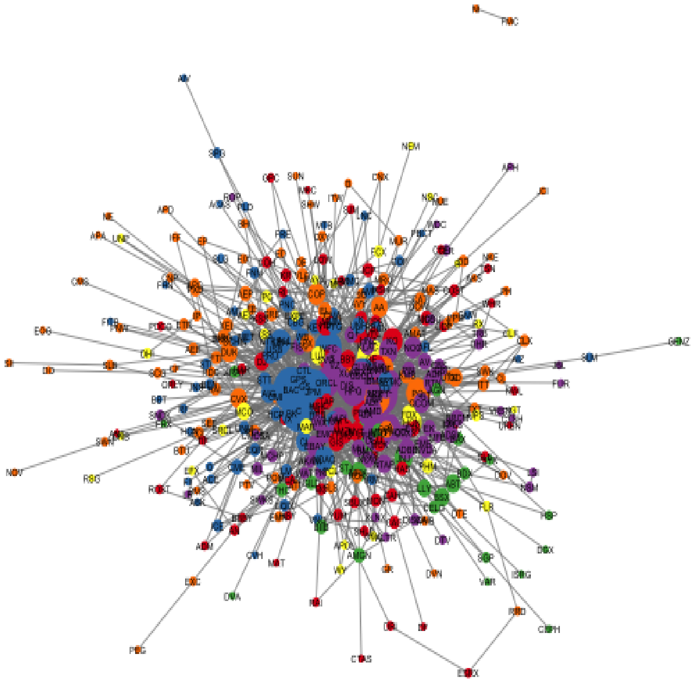
Figure 6: **News-based networks of $S\&P500$ companies identified using all the business news from Business Wire from 2006 to 2013**. The figure plots all the links identified in the sample period. For visualization, only companies with links are plotted. The color of a node indicate which industry the company is in. The industry classification is given in section 4.1. The data source of SIC code for each sample company is from CRSP/Compustat Merged. The size of a node is proportional to the network degree of the company (how many other companies is the node linked with). If there is an edge between two nodes, this indicates there is a link between the nodes.

Linked pair	Link type
(MSFT,INTC)	Strategic partnerships, product developments
(ACE,CB)	Strategic partnerships,merger and acquisition
(C,LM)	Outsourcing, Strategic partnerships
(MSFT,APPL)	Products developments, competition
(BAC,WFC)	Joint venture, strategic partnerships
(CCE,KO)	Partners, common major owners
(MSFT,HPQ)	Strategic partnerships, products developments, competition
(MSFT,ORCL)	Strategic partnerships, products developments
(AXP,V)	Competition, legal
(PEP,KO)	Competition
(BAC,V)	Strategic partnerships, joint venture, products developments
(JPM,GS)	Joint investment banking, competition
(MS,AXP)	Strategic partnerships, joint venture, products developments
(WFC,JPM)	Joint venture, strategic partnerships
(C,JPM)	Joint venture, strategic partnerships
(T,VZ)	Competition
(C,MS)	Joint venture, strategic partnerships
(MSFT,CSCO)	Strategic partnerships, products developments
(NVDA,INTC)	Strategic partnerships, products developments, competition
(Q,CTL)	Competition
(MSFT,ADBE)	Strategic partnerships, products developments
(JPM,GS)	Joint venture, strategic partnerships, joint investment banking, competition
(BA,MSFT)	Strategic partnerships, product developments
(PFE,BMY)	Joint reserach and development
(MSFT,ACN)	Strategic partnerships, products developments
(AMD,INTC)	Supplier-customer
(C,BAC)	Competition
(GE,BA)	Supplier-customer
(BK,JPM)	Business-swap, acquire business lines
(MSFT,INTU)	Strategic partnerships, products developments
(MSFT,CTXS)	Strategic partnerships, products developments
(DISCA,HAS)	Joint venture, supplier-customer
(BSX,STJ)	Legal settlement, competition, joint development effort
(LLY,PFE)	Joint reserach and development
(GS,C)	Strategic partnerships, joint financing
(NOC,BA)	Strategic partnerships, supplier-customer
(K,PG)	Acquire business lines
(AMZN,APPL)	Strategic partnerships, products developments
(GE,JPM)	Strategic partnerships, joint financing
(VZ,MSI)	Alliance, products developments

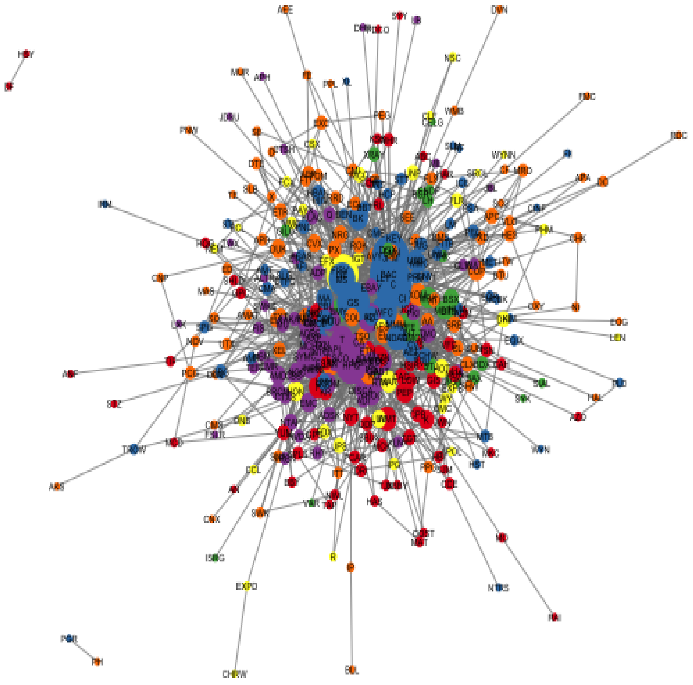
Table 8: Link validation. Note: The table shows the type of economic linkages that the article co-mentioning imply. Since those pairs were co-mentioned quite frequently, for each pair, we randomly read 5 news that have co-mentioned the two firms and infer link type from the news. Thus the listed link types are representative but not exhaustive. Due to space limitations, we only show the validation results for the most frequently co-mentioned pairs.

Table 9: **Summary statistics of news-based links from 2006 to 2013**. I collect all distinct business news within the sample period that tagged the $S\&P500$ companies. Firm links are constructed using the methodology mentioned in section 2. Most pairs of firms are co-mentioned multiple times within the sample period, and we consider both weighted links and unweighted links. For the weighted version, a typical entry w_{ij} of W gives the number of times i and j that are co-mentioned in the sample (co-mentioning multiple times in the same month count only once). On the other hand, for the unweighted version, the entries of W are 0/1 dummies. $w_{ij} = 1$ if i and j are co-mentioned at least once $w_{ij} = 0$ if i and j are never mentioned together in any articles. The unweighted version of the statistics are given in the parentheses below the corresponding weighted statistics. The number of total links identified for industry g is $\sum_{i \in g}^{N} d_i$, where d_i is the links of firm i from industry g. The average and the 90th percentile of d_i are given in the two following rows. I further break down links to intra-industry and inter-industry links. The industry definitions are given in section 4.1

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	11114.00	4843.00	5287.00	13447.00	2821.00	40185.00
	(3403.00)	(2115.00)	(2350.00)	(3616.00)	(870.00)	(13485.00)
Average degree	110.04	47.02	36.46	123.37	70.53	73.60
	(33.69)	(20.53)	(16.21)	(33.17)	(21.75)	(24.70)
90th percentile of degree	169.00	104.80	74.20	247.20	166.90	154.00
	(66.00)	(43.80)	(34.00)	(71.40)	(50.10)	(52.50)
#of intra industry links	6582.00	2020.00	2232.00	8802.00	1660.00	21660.00
	(1372.00)	(724.00)	(864.00)	(1624.00)	(316.00)	(5028.00)
#of inter industry links	4532.00	2823.00	3055.00	4645.00	1161.00	18525.00
	(2031.00)	(1391.00)	(1486.00)	(1992.00)	(554.00)	(8457.00)
%of firms with inter industry links	0.89	0.94	0.87	0.89	0.92	0.90
	(0.89)	(0.94)	(0.87)	(0.89)	(0.92)	(0.90)

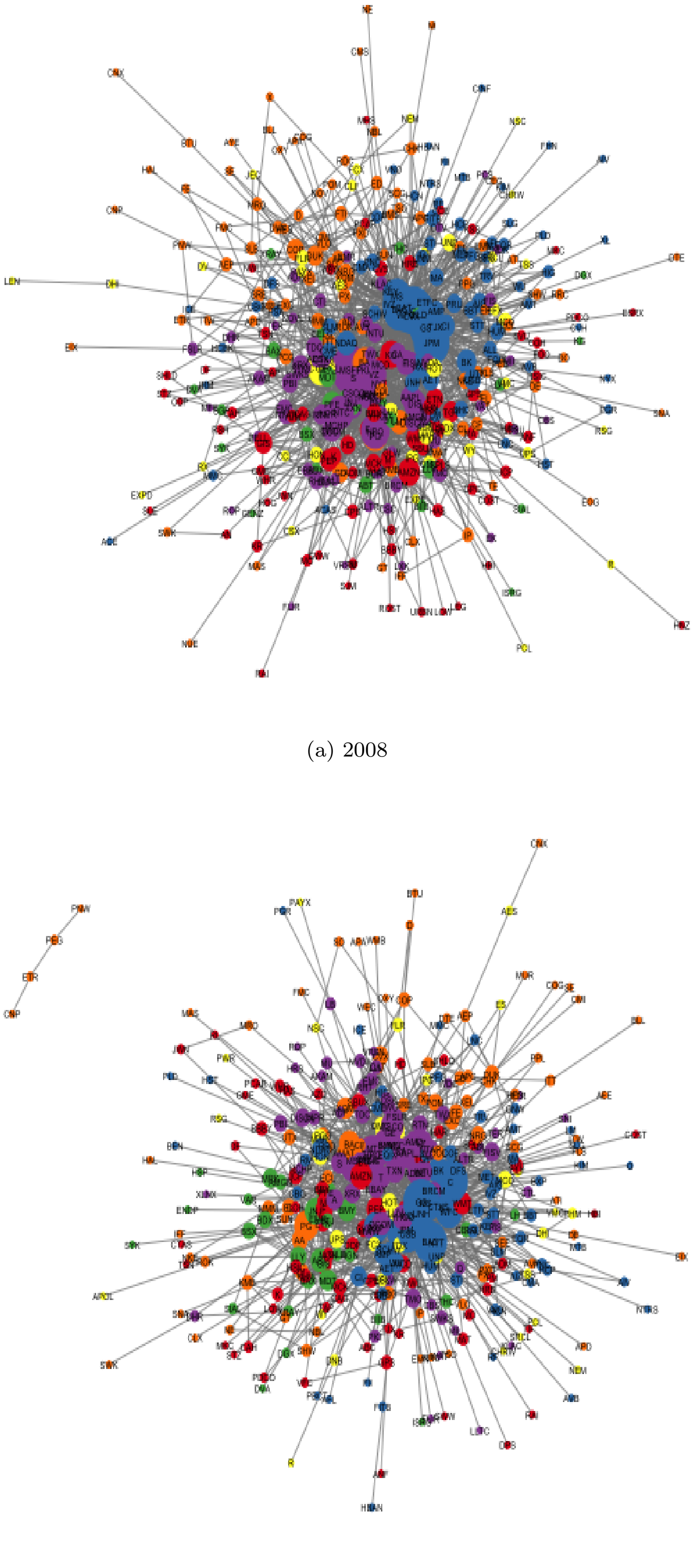


(a) 2006



(b) 2007

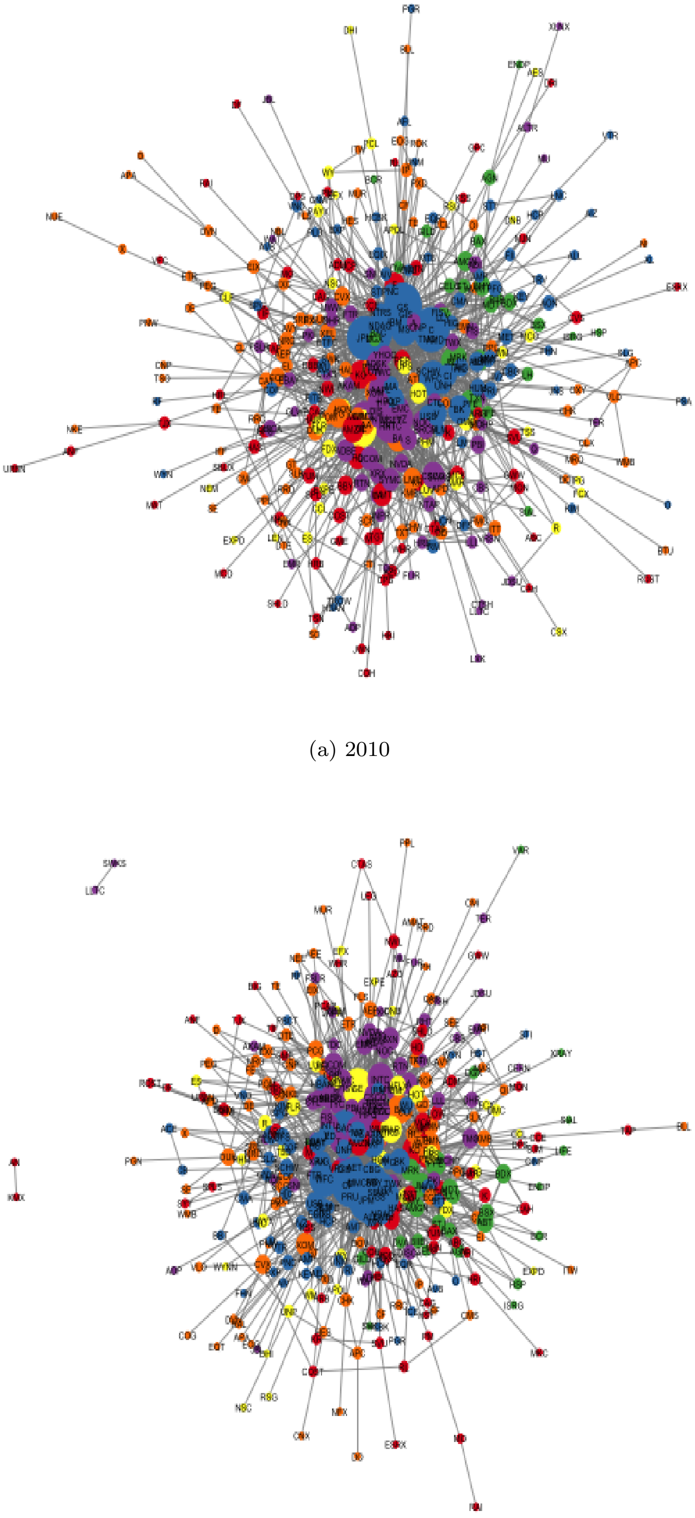
Figure 7: Yearly news-based networks of $S\&P500$ companies. For each year, all the business news from Business Wire that have mentioned sample companies are used to identify links across companies. Only companies with links are plotted. The color code is the same as the aggregate graph and node size is proportional to the network degree of the company (how many other companies is the node linked with).



(a) 2008

(b) 2009

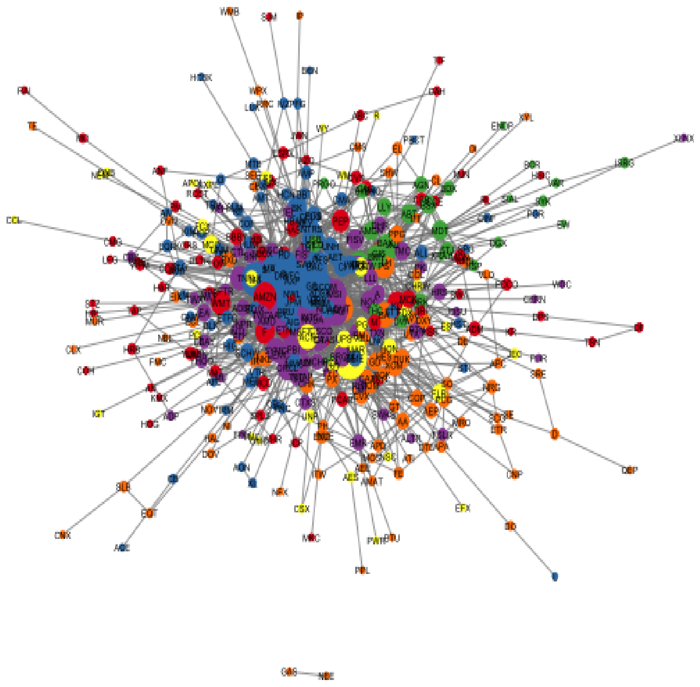
Figure 7: Yearly news-based networks of $S\&P500$ companies. For each year, all the business news from Business Wire that have mentioned sample companies are used to identify links across companies. Only companies with links are plotted. The color code is the same as the aggregate graph and node size is proportional to the network degree of the company (how many other companies is the node linked with).



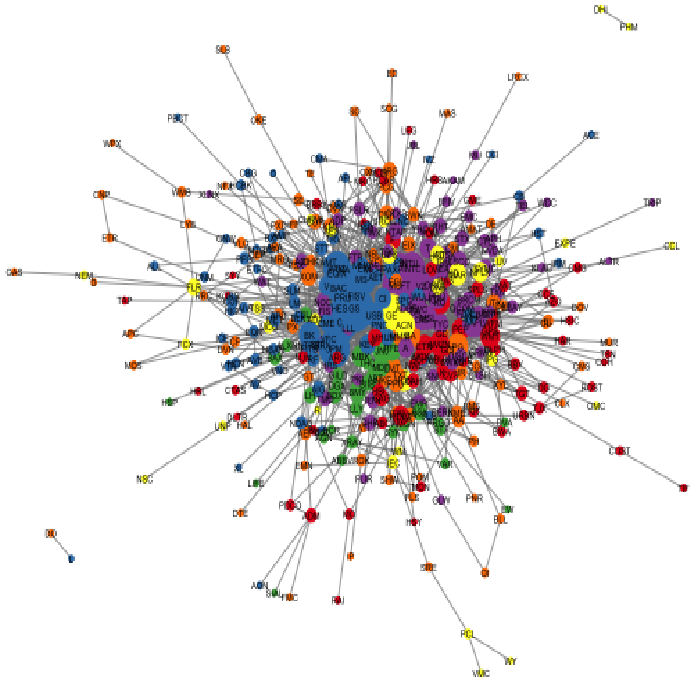
(a) 2010

(b) 2011

Figure 7: Yearly news-based networks of $S\&P500$ companies. For each year, all the business news from Business Wire that have mentioned sample companies are used to identify links across companies. Only companies with links are plotted. The color code is the same as the aggregate graph and node size is proportional to the network degree of the company (how many other companies is the node linked with).



(a) 2012



(b) 2013

Figure 7: Yearly news-based networks of $S\&P500$ companies. For each year, all the business news from Business Wire that have mentioned sample companies are used to identify links across companies. Only companies with links are plotted. The color code is the same as the aggregate graph and node size is proportional to the network degree of the company (how many other companies is the node linked with).

Table 10: **Summary statistics of news-based links from 2006 to 2013**. For each year from 2006 to 2013, I collect all distinct business news within that year and construct firm links using the methodology mentioned in section 2. The summary statistics of that year's news-based links are given in sub-tables. For each year, most pairs of firms are co-mentioned multiple times, and we consider both weighted links and unweighted links. For the weighted version, a typical entry w_{ij} of W gives the number of times i and j that are co-mentioned in the sample (co-mentioning multiple times in the same month count only once). On the other hand, for the unweighted version, the entries of W are 0/1 dummies. $w_{ij} = 1$ if i and j are co-mentioned at least once $w_{ij} = 0$ if i and j are never mentioned together in any articles. The unweighted version of the statistics are given in the parentheses below the corresponding weighted statistics. The number of total links identified for industry g is $\sum_{i \in g}^{N} d_i$, where d_i is the links of firm i from industry g. The average and the 90th percentile of d_i are given in the two following rows. I further break down links to intra-industry and inter-industry links. The industry definitions are given in section 4.1

(a) 2006

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1208.00	544.00	644.00	1762.00	276.00	4671.00
	(739.00)	(409.00)	(486.00)	(954.00)	(190.00)	(2959.00)
Average degree	15.10	6.80	5.24	20.97	8.1	10.57
	(9.24)	(5.11)	(3.95)	(11.35)	(5.58)	(6.69)
90th percentile of degree	29.20	17.10	10.80	43.00	20.70	23.90
	(22.30)	(12.00)	(9.00)	(25.50)	(12.00)	(16.00)
#of intra industry links	678.00	204.00	262.00	1204.00	150.00	2524.00
	(334.00)	(136.00)	(184.00)	(536.00)	(88.00)	(1296.00)
#of inter industry links	530.00	340.00	382.00	558.00	126.00	2147.00
	(405.00)	(273.00)	(302.00)	(418.00)	(102.00)	(1663.00)
%of firms with inter industry links	0.59	0.81	0.63	0.81	0.65	0.69
	(0.59)	(0.81)	(0.63)	(0.81)	(0.65)	(0.69)

(b) 2007

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1295	553	735	1691	372	4946
	(790)	(385)	(504)	(827)	(228)	(2958)
Average degree	14.89	6.74	5.93	19.89	10.94	10.89
	(9.08)	(4.69)	(4.06)	(9.73)	(6.71)	(6.52)
90th percentile of degree	28.80	14.00	13.00	47.60	23.40	23.70
	(18.20)	(10.90)	(9.00)	(22.60)	(15.00)	(15.00)
#of intra industry links	682.00	240.00	312.00	1202.00	210.00	2688.00
	(312.00)	(146.00)	(202.00)	(470.00)	(104.00)	(1266.00)
#of inter industry links	613.00	313.00	423.00	489.00	162.00	2258.00
	(478.00)	(239.00)	(302.00)	(357.00)	(124.00)	(1692.00)
%of firms with inter industry links	0.61	0.72	0.64	0.72	0.74	0.68
	(0.61)	(0.72)	(0.64)	0.72)	(0.74)	(0.68)

Table 10: Continued

(a) 2008

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1324.00	611.00	714.00	1633.00	337.00	4876.00
	(857.00)	(432.00)	(504.00)	(890.00)	(205.00)	(3088.00)
Average degree	15.22	7.36	5.71	18.56	9.91	10.58
	(9.85)	(5.20)	(4.03)	(10.11)	(6.03)	(6.70)
90th percentile of degree	30.40	16.80	11.60	38.60	20.00	23.00
	(20.80)	(12.00)	(8.00)	(23.00)	(13.00)	(14.00)
#of intra industry links	774.00	236.00	312.00	1100.00	178.00	2630.00
	(408.00)	(154.00)	(194.00)	(476.00)	(88.00)	(1340.00)
#of inter industry links	550.00	375.00	402.00	533.00	159.00	2246.00
	(449.00)	(278.00)	(310.00)	(414.00)	(117.00)	(1748.00)
%of firms with inter industry links	0.61	0.82	0.64	0.75	0.76	0.71
	(0.61)	(0.82)	(0.64)	(0.75)	(0.76)	(0.71)

(b) 2009

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1091.00	435.00	505.00	1317.00	295.00	3852.00
	(696.00)	(329.00)	(362.00)	(733.00)	(187.00)	(2472.00)
Average degree	12.54	5.12	4.01	14.16	8.68	8.21
	(8.00)	(3.87)	(2.87)	(7.88)	(5.50)	(5.27)
90th percentile of degree	33.80	9.60	10.00	27.80	21.70	18.00
	(17.80)	(8.00)	(7.50)	(18.60)	(14.70)	(12.00)
#of intra industry links	616.00	158.00	212.00	892.00	168.00	2078.00
	(302.00)	(110.00)	(140.00)	(398.00)	(86.00)	(1056.00)
#of inter industry links	475.00	277.00	293.00	425.00	127.00	1774.00
	(394.00)	(219.00)	(222.00)	(335.00)	(101.00)	(1416.00)
%of firms with inter industry links	0.60	0.66	0.53	0.65	0.68	0.62
	(0.60)	(0.66)	(0.53)	(0.65)	(0.68)	(0.62)

(c) 2010

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1072.00	471.00	538.00	1234.00	286.00	3832.00
	(677.00)	(328.00)	(377.00)	(699.00)	(180.00)	(2442.00)
Average degree	12.04	5.48	4.45	13.41	8.67	8.24
	(7.61)	(3.81)	(3.12)	(7.60)	(5.45)	(5.25)
90th percentile of degree	27.80	12.00	11.00	28.00	21.00	17.60
	(15.80)	(9.00)	(8.00)	(16.90)	(11.80)	(11.60)
#of intra industry links	628.00	194.00	232.00	774.00	178.00	2050.00
	(320.00)	(126.00)	(148.00)	(372.00)	(94.00)	(1094.00)
#of inter industry links	444.00	277.00	306.00	460.00	108.00	1782.00
	(357.00)	(202.00)	(229.00)	(327.00)	(86.00)	(1348.00)
%of firms with inter industry links	0.63	0.63	0.62	0.64	0.70	0.64
	(0.63)	(0.63)	(0.62)	(0.64)	(0.70)	(0.64)

Table 10: Continued

(a) 2011

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1330.00	453.00	549.00	1334.00	317.00	4282.00
	(815.00)	(324.00)	(416.00)	(767.00)	(204.00)	(2742.00)
Average degree	15.29	5.27	4.54	15.16	10.23	9.35
	(9.37)	(3.77)	(3.44)	(8.72)	(6.58)	(5.99)
90th percentile of degree	33.00	12.50	10.00	30.60	24.00	20.00
	(16.80)	(7.50)	(8.00)	(19.30)	(14.00)	(13.30)
#of intra industry links	846.00	196.00	214.00	852.00	192.00	2348.00
	(424.00)	(118.00)	(146.00)	(412.00)	(102.00)	(1236.00)
#of inter industry links	484.00	257.00	335.00	482.00	125.00	1934.00
	(391.00)	(206.00)	(270.00)	(355.00)	(102.00)	(1506.00)
%of firms with inter industry links	0.62	0.72	0.60	0.70	0.81	0.67
	(0.62)	(0.72)	(0.60)	(0.70)	(0.81)	(0.67)

(b) 2012

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1262.00	483.00	574.00	1354.00	367.00	4338.00
	(768.00)	(366.00)	(407.00)	(794.00)	(213.00)	(2770.00)
Average degree	14.34	5.49	4.82	14.72	10.79	9.31
	(8.73)	(4.16)	(3.42)	(8.63)	(6.26)	(5.94)
90th percentile of degree	38.90	12.30	11.00	32.70	27.60	19.50
	(20.00)	(8.30)	(8.00)	(19.00)	(12.00)	(13.00)
#of intra industry links	814.00	200.00	272.00	870.00	244.00	2444.00
	(412.00)	(144.00)	(180.00)	(442.00)	(120.00)	(1330.00)
#of inter industry links	448.00	283.00	302.00	484.00	123.00	1894.00
	(356.00)	(222.00)	(227.00)	(352.00)	(93.00)	(1440.00)
%of firms with inter industry links	0.64	0.74	0.61	0.66	0.76	0.68
	(0.64)	(0.74)	(0.61)	(0.66)	(0.76)	(0.68)

(c) 2013

	Finance	Consumer	Manuf	Hitech	Health	All
#of total links identified	1158.00	477.00	518.00	1133.00	365.00	3990.00
	(752.00)	(359.00)	(388.00)	(695.00)	(230.00)	(2660.00)
Average degree	13.47	5.42	4.11	13.49	10.14	8.60
	(8.74)	(4.08)	(3.08)	(8.27)	(6.39)	(5.73)
90th percentile of degree	29.00	11.60	10.00	23.70	24.50	19.00
	(18.00)	(9.00)	(7.00)	(16.70)	(15.50)	(13.00)
#of intra industry links	736.00	180.00	206.00	678.00	226.00	2076.00
	(398.00)	(128.00)	(144.00)	(368.00)	(118.00)	(1190.00)
#of inter industry links	422.00	297.00	312.00	455.00	139.00	1914.00
	(354.00)	(231.00)	(244.00)	(327.00)	(112.00)	(1470.00)
%of firms with inter industry links	0.64	0.68	0.60	0.75	0.75	0.66
	(0.64)	(0.68)	(0.60)	(0.75)	(0.75)	(0.66)

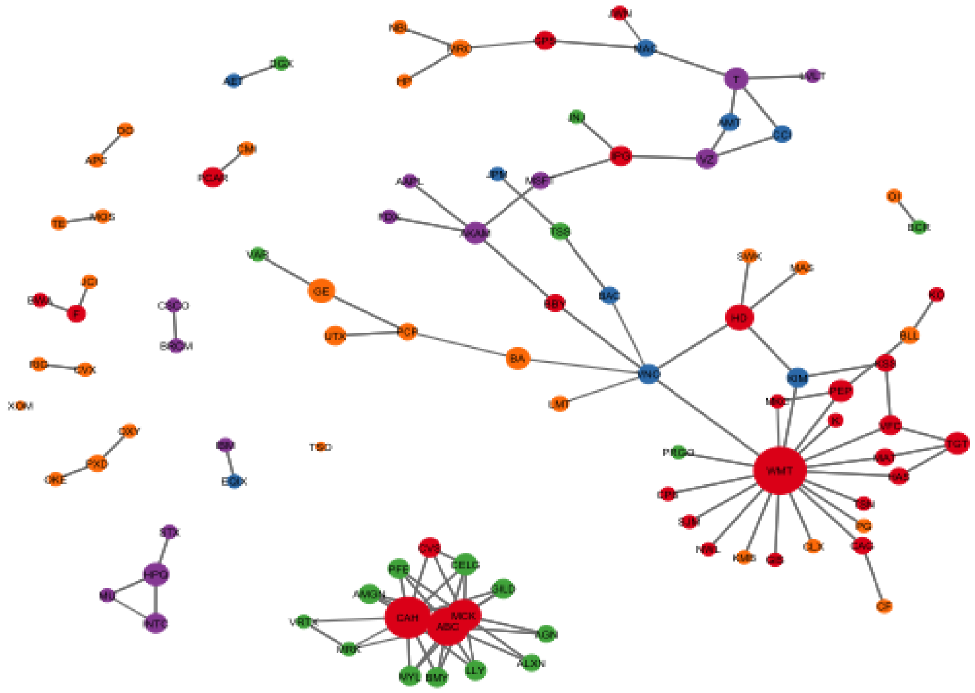
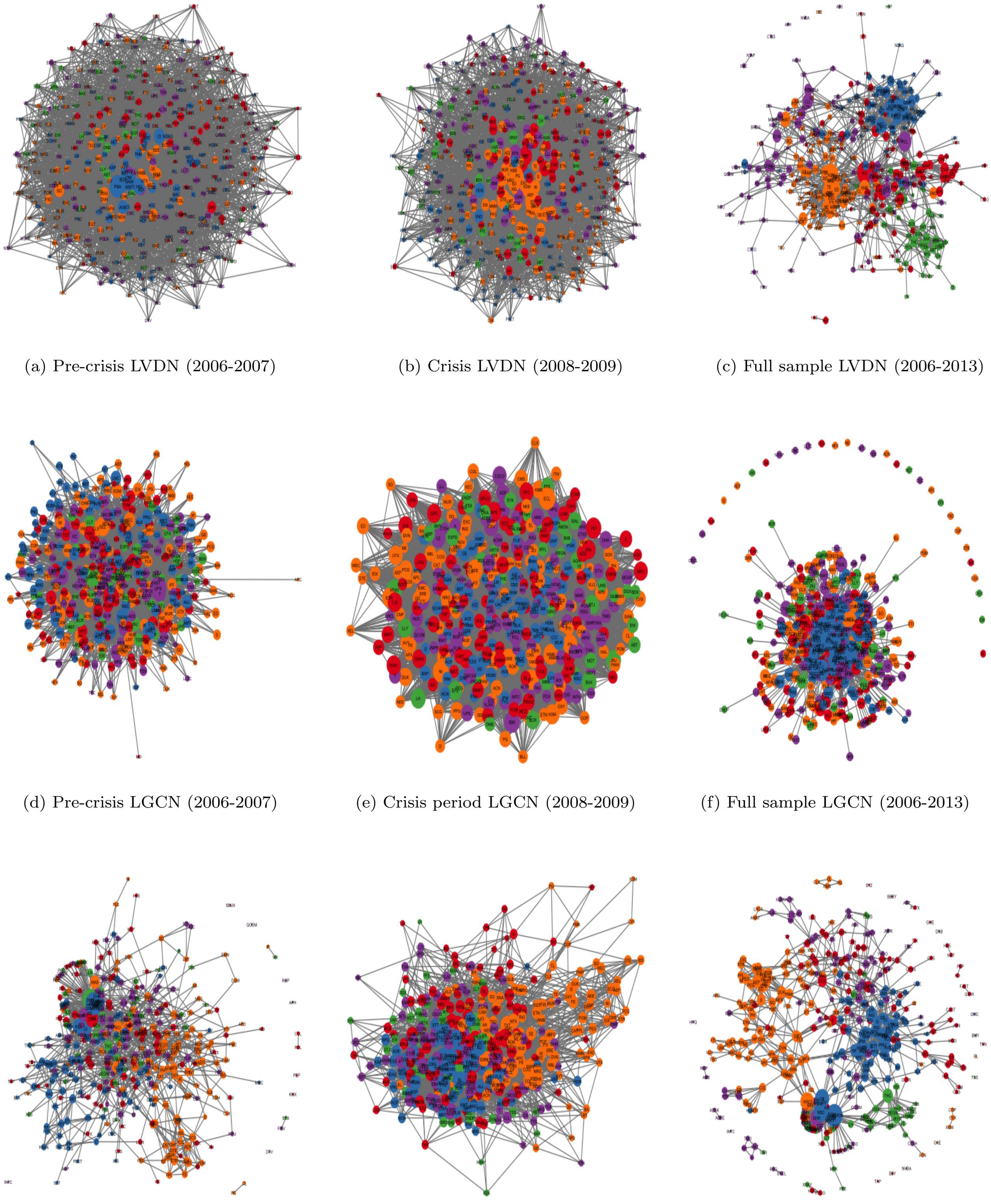


Figure 8: Customer-supplier links among $S\&P500$ companies $(2006-2013)$. Data source: Compustat.



(a) Pre-crisis LVDN (2006-2007)

(b) Crisis LVDN (2008-2009)

(c) Full sample LVDN (2006-2013)

(d) Pre-crisis LGCN (2006-2007)

(e) Crisis period LGCN (2008-2009)

(f) Full sample LGCN (2006-2013)

(g) Pre-crisis PCN (2006-2007)

(h) Crisis PCN (2008-2009)

(i) Full sample PCN (2006-2013)

Figure 9: Long-run variance Decomposition network (LVDN), Long-run Granger causality network (LGCN) and Partial correlation network (PCN) applying the high-dimensional method from Barigozzi and Hallin (2017) on the de-factored returns from equation (1). Note: The LGCN and PCN are sparse given that the high-dimensional VAR and correlation matrix are regularized. LVDN, on the other hand, is dense (for the 3 samples, the link densities are all over 75%. Hard thresholding is applied and only links that contribute to more than 1% of the future variances are kept, thus the plotted LVDN is sparse. Link densities of LVDN applying different thresholds are presented in Table 8. Red, blue, green, purple and orange correspond to consumer, finance, health, hitech and manufacturing industry, respectively. Node size is proportional to out degree.

Threshold	Pre-crisis (2006-2007)	Crisis (2008-2009)	Full sample (2006-2013)
0	131935	162091	111165
1	3682	3835	1639
2	711	770	574
3	275	307	310
4	138	201	205
5	102	158	149

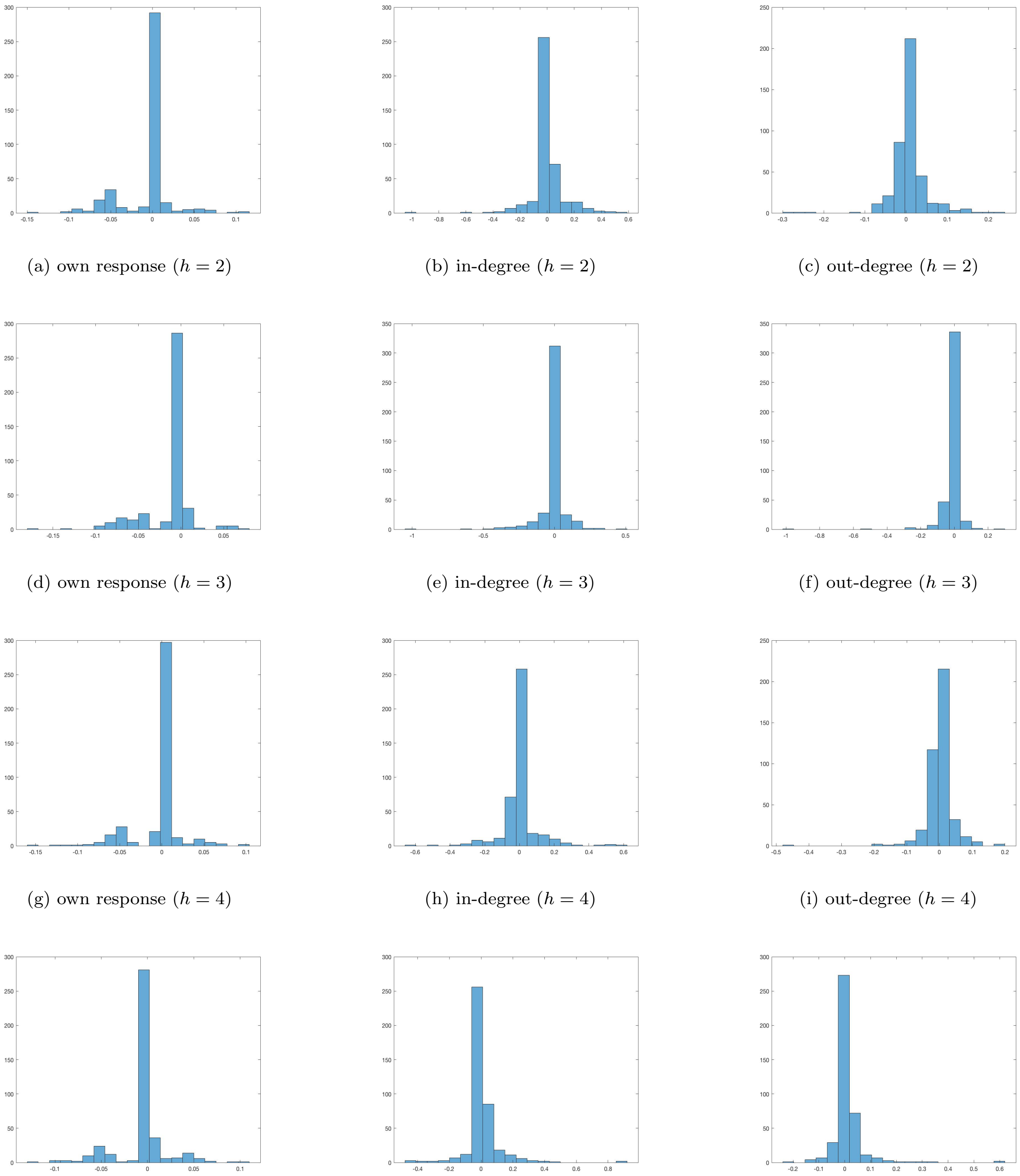
Table 11: Number of Long-run variance Decomposition network (LVDN) links after applying different hard thresholds (in percentage of future variance explained) for pre-crisis, crisis and full sample periods. Note: if threshold k is applied, a link from i to j is kept only if the shocks to i (the cumulative effect up to 10 lags) contribute to at least $k\%$ of j's variance.

	Pre-crisis (2006-2007)	Crisis (2008-2009)	Full sample (2006-2013)
LGCN	2005	9319	721
PCN	1614	3666	1486

Table 12: Number of Long-run Granger causality network (LGCN) and the Partial correlation network (PCN) links identified from pre-crisis, crisis and full sample periods.

Threshold	Pre-crisis LVDN	Text-based networks
0	0.783	0.03
1	0.04	0.06
2	0.06	0.16
3	0.07	0.26
4	0.05	0.31
5	0.04	0.34

Table 13: Percentages of crisis period Long-run variance Decomposition network (LVDN) links that get identified using alternative pre-crisis network information. Note: Different hard thresholds are applied to the LVDN given the network implied by LVDN is very dense (the link densities for pre-crisis and crisis sample are 77.5% and 95.3%, respectively). We do not need to apply thresholding to text-based network since it is already very sparse (the link density of the full sample network is 4.5%, and for the short pre-crisis sample the density is even smaller.



(a) own response ($h = 2$)

(b) in-degree ($h = 2$)

(c) out-degree ($h = 2$)

(d) own response ($h = 3$)

(e) in-degree ($h = 3$)

(f) out-degree ($h = 3$)

(g) own response ($h = 4$)

(h) in-degree ($h = 4$)

(i) out-degree ($h = 4$)

(j) own response ($h = 5$)

(k) in-degree ($h = 5$)

(l) out-degree ($h = 5$)

Figure 10: Histogram for own response, in-degree and out-degree at horizon $h = 2, 3, 4, 5$.

	(1)AR terms					(2) spatial-temporal terms						(3) σ
	λ_1	λ_2	λ_3	λ_4	λ_5	ψ_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	σ
Median	-0.026	-0.014	-0.014	-0.009	-0.002	0.221	0.021	0.003	0.005	0.005	0.003	1.395
MG Estimates	-0.026	-0.014	-0.016	-0.009	-0.005	0.270	0.032	0.004	0.002	0.008	0.008	1.491
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.019)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.025)
% Sig (at 5%)	40.9%	23.2%	22.0%	18.6%	19.6%	77.2%	24.5%	22.1%	15.7%	16.4%	14.5%	-
Non-zero coef.	413	413	413	413	413	408	408	408	408	408	408	413

Table 14: **QML estimation results of heterogeneous spatial temporal model (2) using $\hat{\epsilon}_{it}^{pca}$**

	(1)AR terms					(2) spatial-temporal terms						(3) σ
	λ_1	λ_2	λ_3	λ_4	λ_5	ψ_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	σ
Panel A: Consumer												
Median	-0.024	-0.017	-0.010	-0.014	-0.004	0.164	0.015	0.003	0.002	0.002	0.011	1.359
MG Estimates	-0.026	-0.019	-0.010	-0.011	-0.006	0.201	0.010	0.015	0.003	-0.001	0.008	1.431
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.039)	(0.011)	(0.009)	(0.008)	(0.010)	(0.010)	(0.051)
% Sig(at 5%)	32.5%	22.1%	19.5%	16.9%	19.5%	74.0%	15.6%	9.1%	11.7%	11.7%	6.5%	-
Non-zero coef.	77	77	77	77	77	77	77	77	77	77	77	77
Panel B: Finance												
Median	-0.032	-0.014	-0.024	-0.013	0.000	0.179	0.035	-0.015	0.015	0.000	0.020	1.574
MG Estimates	-0.035	-0.019	-0.028	-0.022	-0.002	0.257	0.074	-0.016	0.009	0.026	0.028	1.751
	(0.008)	(0.006)	(0.005)	(0.005)	(0.005)	(0.055)	(0.024)	(0.018)	(0.017)	(0.015)	(0.017)	(0.070)
% Sig(at 5%)	50.7%	41.3%	42.7%	33.3%	33.3%	78.7%	42.7%	36.0%	32.0%	28.0%	23.0%	-
Non-zero coef.	75	75	75	75	75	74	74	74	74	74	74	75
Panel C: Health												
Median	-0.009	-0.010	-0.002	-0.002	0.005	0.149	0.031	0.015	0.014	0.009	0.005	1.379
MG Estimates	-0.012	-0.005	-0.008	-0.006	0.007	0.203	0.032	0.016	0.013	0.013	0.024	1.459
	(0.006)	(0.005)	(0.005)	(0.005)	(0.004)	(0.068)	(0.015)	(0.014)	(0.010)	(0.017)	(0.015)	(0.087)
% Sig(at 5%)	22.9%	11.4%	14.3%	14.3%	17.1%	74.3%	20.0%	17.1%	8.6%	11.4%	17.6%	-
Non-zero coef.	35	35	35	35	35	34	34	34	34	34	34	35
Panel D: Hitech												
Median	-0.041	-0.020	-0.014	-0.009	-0.004	0.136	-0.004	0.002	0.020	0.005	-0.006	1.440
MG Estimates	-0.033	-0.019	-0.013	-0.008	-0.008	0.183	0.011	0.005	0.004	0.006	-0.008	1.553
	(0.005)	(0.003)	(0.003)	(0.003)	(0.003)	(0.040)	(0.011)	(0.011)	(0.009)	(0.010)	(0.011)	(0.058)
% Sig(at 5%)	53.4%	17.8%	9.6%	13.7%	11.0%	65.8%	15.1%	17.8%	5.5%	12.3%	11.0%	-
Non-zero coef.	73	73	73	73	73	73	73	73	73	73	73	73
Panel E: Manufacturing												
Median	-0.013	-0.003	-0.016	-0.003	-0.005	0.407	0.010	0.012	-0.001	0.003	0.003	1.255
MG Estimates	-0.018	-0.006	-0.017	-0.003	-0.010	0.411	0.025	0.007	-0.003	0.005	0.002	1.288
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.033)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.039)
% Sig(at 5%)	33.6%	19.1%	19.1%	19.1%	20.9%	79.1%	18.2%	20.9%	11.8%	14.5%	16.7%	-
Non-zero coef.	110	110	110	110	110	108	108	108	108	108	108	110
Panel F: Other												
Median	-0.045	-0.018	-0.014	-0.008	0.001	0.208	0.044	-0.002	-0.003	0.011	0.008	1.471
MG Estimates	-0.032	-0.018	-0.019	-0.006	0.001	0.225	0.049	-0.005	-0.012	-0.004	0.003	1.590
	(0.006)	(0.004)	(0.005)	(0.003)	(0.004)	(0.065)	(0.018)	(0.017)	(0.014)	(0.016)	(0.015)	(0.078)
% Sig(at 5%)	51.2%	23.3%	25.6%	7.0%	9.3%	76.7%	30.2%	20.9%	14.0%	7.0%	11.9%	-
Non-zero coef.	43	43	43	43	43	42	42	42	42	42	42	43

Table 15: **QML estimation results of heterogeneous spatial temporal model (2) using $\hat{\epsilon}_{it}^{pca}$, parameters summarized by industry.**