Networks of News and Cross-Sectional Returns

Junjie Hu [*]
Wolfgang Karl Härdle [*] [*2] [*3] [*4] [*5]

[*] Humboldt-Universität zu Berlin, Germany
[*2] Xiamen University, China
[*3] Singapore Management University, Singapore
[*4] Charles University, Czech Republic
[*5] National Yang Ming Chiao Tung University, Taiwan

Networks of News and Cross-Sectional Returns[*]

Junjie Hu[†], Wolfgang Karl Härdle[§]

First draft: April 2021
This draft: October 2021

Abstract

We uncover networks from news articles to study cross-sectional stock returns. By analyzing a huge dataset of more than 1 million news articles collected from the internet, we construct time-varying directed networks of the S&P500 stocks. The well-defined directed news networks are formed based on a modest assumption about firm-specific news structure, and we propose an algorithm to tackle type-I errors in identifying the stock tickers. We find strong evidence for the comovement effect between the news-linked stocks returns and reversal effect from the lead stock return on the 1-day ahead follower stock return, after controlling for many known effects. Furthermore, a series of portfolio tests reveal that the news network attention proxy, *network degree*, provides a robust and significant cross-sectional predictability of the monthly stock returns. Among different types of news linkages, the linkages of within-sector stocks, large size lead firms, and lead firms with lower stock liquidity are crucial for cross-sectional predictability.

[†] School of Business and Economics, Humboldt-Universität zu Berlin. Spandauer Str. 1, 10178 Berlin, Germany. Email: junjie.hu@hu-berlin.de

[§] Blockchain Research Center, Humboldt-Universität zu Berlin, Germany; Wang Yanan Institute for Studies in Economics, Xiamen University, China; Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore; Faculty of Mathematics and Physics, Charles University, Czech Republic; National Yang Ming Chiao Tung University, Taiwan. Email: haerdle@hu-berlin.de

1. Introduction

It is undeniable that the interdependency between firms is a major driver for stock returns; the practical problem is though how to capture this time-varying dependency structure. Among all the information sources, the fast-growing resource of news articles provides us via textual analysis a perspective to observe and define the connections between firms. Taking advantage of the advancement of computation and information technology, e.g., "Text Mining", we are capable of capturing the linkages between firms from a huge dataset of more than 1 million news articles on NASDAQ.com. These linkages allow us to study cross-sectional stock returns based on news information.

Unlike the customer-supplier linkage in a supply chain, which has a natural representation of the linkage and direction, we first need to define the news linkage and its direction. With the directed linkages, we then study the impact of stock return via the news linkage and the structural information of the directed news networks.

We postulate a news linkage as a firm of main interest (*lead*) pointing to a related firm (*follower*) within a piece of news. By assuming that the firm-specific news has a structure of placing the firm of main interest on its headline and the related firms in its content, we uncover the directed news linkages based on the positions that those firms are identified, i.e., in the title or content. The assumption is modest because when writing firm-specific news reports, the authors will naturally emphasize the firm of interest on the headline and mention the other related firms in the content. Such definition on news linkage is different from the literature, where news linkage is formed if two firms are co-mentioned in one article (e.g., Scherbina and Schlusche (2015); Chen, Guo, and Tu (2021); Guo and Li (2019); Ge, Linton, and Li (2021)). Consequently, a directed news network is uncovered based on a collection of directed news linkages in a given period.

The task of firm tickers identification is the cornerstone of the later empirical analysis; however, the task is challenging as it could suffer from the type-I error if not handled carefully. We propose an algorithm that is dedicated to narrow down firm tickers identification of S&P500 stocks only in the three most commonly seen and reliable scenarios with some specified exceptional cases. This algorithm is a trade-off between fuzzy and inefficient searching, dealing with the problem caused by the various ways of firm names written by a human. On the one hand, using keywords and tickers to identify firms is intuitive. However, this fuzzy searching falsely matches a huge amount of firms that are not in the news, i.e., also known as the type-I error.[1] On the other hand, manually searching over the more than 1-million news articles in our sample is infeasible with limited time and resources. This algorithm will

---

[1]We also did such an experiment, some firms were falsely identified for more than couple thousand times.

surely miss many firm occurrences, but for scientific prudence, a subset of the true set of news linkages gives conservative results.

Applying the proposed algorithm on a dataset of more than 1-million economic and finance related news articles published by 50+ online news publishers on the NASDAQ.com, we uncover the news network of S&P500 stocks and conduct our empirical analysis focusing on two angles: *Lead return*, the return of a linkage-weighted portfolio consisting of all the lead firms of a firm in a given news network, a similar definition can be found in Cohen and Frazzini (2008), and Scherbina and Schlusche (2015). The network attention proxy, *network degree*, a well-known measure in complex network science and social network analysis (Albert, Jeong, and Barabási (1999); Barabási and Albert (1999); Jackson (2010)), intuitively counts the number of unique nodes that a firm directly connects with, therefore captures the "attention hype" of a firm.

Interestingly, we find that the detected news-linked firms, regardless of their business relationship, have a strong comovement effect on daily stock returns. We first calculate the daily stock residual return variables $r_{i,t}^{resid}$ from Fama-French 3-factor model (FF3, Fama and French (1993)) and Fama-French 5-factor model (FF5, Fama and French (2015)). As such variable $r_{i,t}^{resid}$ can be viewed as the part unexplainable by the known factors and firm alphas. By regressing $r_{i,t}^{resid}$ on the contemporaneous *lead return* variables with control variables of firm characteristics, including market capitalization, book-market value, and firm liquidity, the econometric model shows that the daily stock residual return positively relates to the contemporaneous *lead return* at 1% level. The standard error is corrected by clustering on both individual and time. The regression result implies that the follower stock return moves with its lead stock return to the same direction on average over time. The results stay significant after considering whether the lead and follower firms belong to the same sector, though the comovement effect is more intensive between firms within the same sector. Moreover, the effect is significant at 1% regardless of the sign of lead firm stock returns.

We further validate the comovement effect by forming portfolios based on *lead return* of each firm. Of course, the portfolio test is infeasible as it requires knowing the stock returns of all lead firms and the news linkages on day-t to form the portfolios on day-t, but the point here is to study the significance of stock returns comovement effects between the lead firms and follower firms. The trading rule is intuitive. First, all the S&P500 constituents are sorted into quintiles in ascending order based on the contemporaneous *lead return*. Then, the portfolios are equal-weighted and rebalanced each trading day. A significant difference in returns, averaging 165.15 bps per day, between the quintile of highest *lead return* and the lowest.

A reversal effect from *lead return* on the 1-day ahead follower stock return emerges after

controlling for the Fama-French factors and firm characteristics. We conduct the regressions of the 1-day ahead stock (residual) return variables $r_{i,t+1}$ $\left(r_{i,t+1}^{resid}\right)$ on the day-t *lead return* variables along with control variables including the firm characteristics and also contemporaneous stock return $r_{i,t}$. The regression results show that the *lead return* has a significant negative impact on the 1-day ahead $r_{i,t+1}^{resid}$, despite the mixed effect on $r_{i,t+1}$, implying the existence of reversal effects from *lead return* to stock return, nevertheless being dominated by many factors that offset the reversal effect. Furthermore, a daily rebalancing trading strategy based on sorting *lead return* reveals that it gives only marginal cross-sectional predictability of 1-day ahead stock return. In the meanwhile, a strong reversal effect can be found with the accumulated residual stock returns.

We employ the *network degree* as a proxy of measuring the "network attention hype" of a firm and find significant predictability of cross-sectional stock returns. *Network degree* is motivated by our observation that a firm with more attention from news media has a higher likelihood of being found with other firms, i.e., a higher network degree. By conjecturing that a "hyped" firm, i.e., a firm with greater network attention, is more likely to outperform, we provide evidence that such information can be profitable with a simple long-short strategy. The S&P500 stocks are sorted into quintiles each month based on the monthly aggregated network attention from the preceding month. The portfolio test shows that firms with higher network attention yield higher monthly returns in the next month. By investigating the attention variables of news networks formed by different types of news linkages, we find that the network attention, both being lead and follower, shows significant predictability; and the within-sector network attention variable carries much stronger predictability than the cross-sector network attention. Additionally, the lead firm size and stock liquidity are important considerations for predictability. Specifically, the network attention of the lead firm with a larger size and lower stock liquidity is more informative to predict the cross-sectional stock returns. To ensure that the predictability of *network degree* is not only caused by some of the known effects, we provide robust evidence of controlling for firm size, value, liquidity, momentum, and also the selection of network formation window size.

This paper contributes to a large literature that aims to distill interdependency information between firms from all the resources other than the stock market. For example, much of the literature focuses on the supply-chain networks and studies the risks implications at both macroeconomics level (e.g., Gabaix (2011); Acemoglu, Carvalho, Ozdaglar, and Tahbaz-salehi (2012)) and the firm level (e.g., Ahern and Harford (2014); Herskovic (2018); Herskovic, Kelly, Lustig, and Van Nieuwerburgh (2020)), and also stock return predictability (e.g., Cohen and Frazzini (2008)). Moreover, some papers investigate the stock performance by utilizing common shareholders structure (e.g., Anton and Polk (2014)), air

traveling connection (e.g., Da, Gurun, Li, and Warachka (2021)).

Textual news information has been drawing attention in academia as powerful statistical tools, and commercial databases became readily available (e.g., Antweiler and Frank (2004); Da, Engelberg, and Gao (2014); Zhang, Härdle, Chen, and Bommes (2016); Boudoukh, Feldman, Kogan, and Richardson (2018)). Recently, a new trend of studying the firm connection in text data has arisen. Scherbina and Schlusche (2015); Chen, Guo, and Tu (2021); Guo and Li (2019) employ the Thomson Reuters News Analytics (TRNA) dataset and construct the linkages of firms co-mentioned in the same article, Schwenkler and Zheng (2019) employ the natural language processing tools to identify the firm tickers. Scherbina and Schlusche (2015) defines the lead-follower relationship between the news-linked firms by liquidity characteristic and identify the competitive relationships. As such, they show the predictability of returns of the lead firms. Schwenkler and Zheng (2019) investigate the market contagion by various interconnectivity measures. Chen, Guo, and Tu (2021) focus on using the strength of connection of firms to predict the stock returns correlations and how centrality matter for the predictability. Ge, Linton, and Li (2021) calibrate a spatial factor model with the news-implied dependency and derive an asset pricing model.

Another strand of literature studying the interdependency of stocks focuses on modeling the stock price processes with statistical methods. For example, some papers study the connectedness of stocks by decomposing the covariance matrix (e.g., Diebold and Yılmaz (2014); Diebold and Yilmaz (2008)), some investigate the risk transmission between stocks (e.g., Hautsch, Schaumburg, and Schienle (2014); Härdle, Wang, and Yu (2016); Yang and Zhou (2017)), and some study the network effect with the vector autoregression framework (Barigozzi and Brownlees (2019); Zhu, Wang, Wang, and Härdle (2019)).

The rest of the paper is organized as follows. Section 2 first defines the directed news networks and then introduces the two variables along with by the econometric and portfolio methods employed in the empirical analysis. Section 3 details the proposed algorithm that identifies the tickers in S&P500 constituents and summarizes the constructed empirical news network. In Section 4 we give evidence on the comovement effect of the news-linked firms. Then we discuss the negative impact of *lead return* on the 1-day ahead stock returns. In Section 5 we focus on the monthly predictability of the network attention proxy, *network degree*, followed by the analysis of the importance of different types of linkages, and we conduct a series of robustness tests on the network attention predictability.

2. Methodologies and Models

We first define a directed news network and then introduce the two variables derived from the news network, which are the two angles that relate the news networks with stock returns. Different from econometric literature, e.g., Diebold and Yılmaz (2014), in which networks are uncovered under imposed models, the constructed adjacency matrix in this study are model-free and can be tested against econometric hypotheses. Finally, we introduce the econometric model and portfolio construction methods employed in the empirical analysis.

2.1. A Directed Network on Firm-Specific News

News naturally embeds lead-follower relations since authors decide which firms to mention in headlines and will naturally mention the consequences for related firms. To define the lead-follower relationship, we first make a modest assumption on the structure of those firm-specific news articles.

Assumption 2.1. *The firm-specific news articles have the firm of main interest on the news headline and the related firms in the news content.*

To understand the intuition behind this assumption, one can consider a situation where a news article is designated to cover firm j. The author most likely will mention firm j on the article headline and mention the related firms in the content consequently for various purposes, such as potential competitors, cooperators, suppliers, customers, etc. We give a typical example of the firm-specific news in Figure 12 of Appendix C, which demonstrate a piece of news about *Apple*, where some other related firms, e.g., *Intel*, *Nike*, and *Peleton*, are also mentioned in the content. Of course, not all the firm-specific news articles follow such structure; however, we argue that this assumption is naturally true for most firm-specific news and allows a clear definition of the lead-follower relationship.

Based on the assumption, we first define the firm that appears in the news title, firm j, as the *"lead"* firm. Correspondingly in the same news article, each firm i mentioned in the news content is defined as the *"follower"* of firm j. Therefore, given a set of textual news articles aggregated in period T, $D_T = \{m1, \ldots, m_d\}$, we identify the set of news linkage pairs $l_{ij,T}$ of (i, j) among n firms, i.e. $i, j \in \{1, \ldots, n\}$ as

$$l_{ij,T} \stackrel{\text{def}}{=} \bigcup_{m_d \in D_T} \left\{ (i,j)_{m_d} \,|\, j \text{ in } m_d \text{ title}, i \text{ in } m_d \text{ headline}, i \neq j \right\}. \tag{1}$$

With the well-defined news linkage pairs, we then define the *"News Network"*, $\mathcal{W}_T$ as follows. Given a weighted direct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in a time period T by the adjacency

matrix $\mathcal{W}_T \stackrel{\text{def}}{=} [\omega_{ij,T}]_{n \times n} = \left[n_{i,T}^{-1} a_{ij,T}\right]_{n \times n}$, where

$$a_{ij,T} \stackrel{\text{def}}{=} \# l_{ij,T}, \tag{2}$$

and $n_{i,T} = \sum_{j=1}^{n} a_{ij,T}$. The $a_{ij,T}$ counts the number of lead-follower pairs among all the articles in period T, and the adjacency matrix is row-normalized by n_i. Also, note that $\mathcal{W}_T$ is an asymmetric matrix with all zero diagonal values.

2.1.1. Network Decomposition by Nodes

The news networks can be decomposed by the nodes characteristics to further study the effects of various types of linkages. In this paper, we mainly study 3 types of decomposition. The first one is decomposition by sector. This decomposition tests whether a stock interacts with the stocks from different sectors differently compared to the ones from the same sector. Based on the sector defined by Global Industry Classification Standard (GICS), a news network $\mathcal{W}_T = [\omega_{ij,T}]$ can be fully decomposed into within-sector network $\mathcal{W}_T^w = \left[\omega_{ij,T}^w\right]$ and cross-sector network $\mathcal{W}_T^c = \left[\omega_{ij,T}^c\right]$, where

$$\begin{aligned}
\omega_{ij,T}^w &\stackrel{\text{def}}{=} \mathbf{I}\{\text{i,j from same sector}\} \cdot \omega_{ij,T} \\
\omega_{ij,T}^c &\stackrel{\text{def}}{=} \mathbf{I}\{\text{i,j from diff. sectors}\} \cdot \omega_{ij,T},
\end{aligned} \tag{3}$$

and $\mathbf{I}$ is the indicator function, T is the period of network formation.

The second type decomposition is by the firm size of the lead firm. It is intuitive to suspect that the impact from a large market capitalization firm is much stronger than that of a small size firm. We define big firms as the firms with market capital higher than the 70% quantile of the market capitalization across all the S&P500 stocks, $q_{0.7}^{mv}$). And small firms are the firms with market capital lower than the 30% quantile of the market capitalization across all the S&P500 stocks, $q_{0.3}^{mv}$). We take the average value of market capitalization over the network formation time window T to avoid forward-looking bias. As such we have the news network consisting of only linkages with big size lead firm, $\mathcal{W}_T^{BigLead} = \left[\omega_{ij,T}^{BigLead}\right]$, and similarly, the news network of small lead firm, $\mathcal{W}_T^{SmallLead} = \left[\omega_{ij,T}^{SmallLead}\right]$, where

$$\begin{aligned}
\omega_{ij,T}^{BigLead} &\stackrel{\text{def}}{=} \mathbf{I}\{mv_j > q_{0.7}^{mv}\} \cdot \omega_{ij,T} \\
\omega_{ij,T}^{SmallLead} &\stackrel{\text{def}}{=} \mathbf{I}\{mv_j <= q_{0.3}^{mv}\} \cdot \omega_{ij,T},
\end{aligned} \tag{4}$$

The third decomposition is similar to the second one but with stock liquidity. A firm with

6

higher stock liquidity is deemed to be more efficient in incorporating market information. Hence, decomposition by the lead firm's stock liquidity can test whether there is a difference in information transmission between a more liquid stock and a less liquid stock. Similar to firm size decomposition, the breakpoints of low liquidity and high liquidity are 30% quantile $q_{0.3}^{turnover}$, and 70% quantile $q_{0.7}^{turnover}$, of the average turnover rate over the network formation period T. Then, we define the network of high liquidity lead, $\mathcal{W}_T^{HighLead} = \left[\omega_{ij,T}^{HighLead}\right]$, and network of low liquid lead, $\mathcal{W}_T^{LowLead} = \left[\omega_{ij,T}^{LowLead}\right]$, where

$$
\begin{aligned}
\omega_{ij,T}^{HighLead} &\overset{\text{def}}{=} \mathbf{I}\{turnover_j > q_{0.7}^{turnover}\} \cdot \omega_{ij,T} \\
\omega_{ij,T}^{LowLead} &\overset{\text{def}}{=} \mathbf{I}\{turnover_j <= q_{0.3}^{turnover}\} \cdot \omega_{ij,T},
\end{aligned}
\tag{5}
$$

2.2. News Network Derived Variables

Two variables, *lead return* and *network degree*, are derived from the news network along with stock return. The news network, $\mathcal{W}_T$, is uncovered with the definition mentioned above in period $T = [t - l : t]$, and stock return $r_{i,t}$ is the daily stock return of firm i in day-t.

2.2.1. News Network Lead Return

The first network derived variable is aimed to investigate the impacts from lead stocks returns to follower stocks returns. *Lead return*, $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t})$, of stock i at day-t is defined as the linear combination of all the returns of its lead companies $r_{j,t}$ given the news network $\mathcal{W}_{t-l:t} = [\omega_{ij,t-l:t}]$,

$$
\mathcal{LR}_{i,t}(\omega_{ij,t-l:t}) \overset{\text{def}}{=} \sum_{j=1}^{n} \omega_{ij,t-l:t} \cdot r_{j,t},
\tag{6}
$$

where $[t - l : t]$ is the time period of network formation, and l is the window size. Note that $\omega_{ij,t-l:t}$ is a coefficient determined by the amount of linkages between firm i and j, hence the network price proxy $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t})$ also captures the attention effect. For instance, firm j will be more influential to *lead return* if there is a larger volume of news connecting i and j. A similar definition can refer to the literature (e.g., Cohen and Frazzini (2008); Scherbina and Schlusche (2015)).

Lead return is employed here for two reasons. First of all, it is the most intuitive way to synthesize the stock returns from the lead firms to each company i. Secondly, *lead return* can be easily decomposed, e.g., by the characteristics of nodes and linkages in the news network, and its decomposed variants allow flexible econometric and portfolio tests. Finally,

the econometric model with *lead return* essentially tests the direct network effect between the lead stock return and follower stock return.

The timeline of the information used is carefully defined on a daily frequency to ensure no future information bias. As shown in Figure 1, the information set on day-t noted as $\mathcal{F}_t$ is defined as 09:00 am (day-t) until 09:00 am (day-$t+1$). Correspondingly, the stock logreturn on day-t is defined by the logarithmic difference between open price at day-$t+1$ and day-t, i.e $r_t \overset{\text{def}}{=} \log p_{t+1}^{open} - \log p_t^{open}$. The information used to construct factor on day-t is up to half an hour before the market opens at day-$t+1$. Note that we synchronize the article arriving time with the local exchanges time (New York Time, i.e., Eastern Time Zone (UTC-05:00) with daylight saving time (UTC-04:00), in our case).
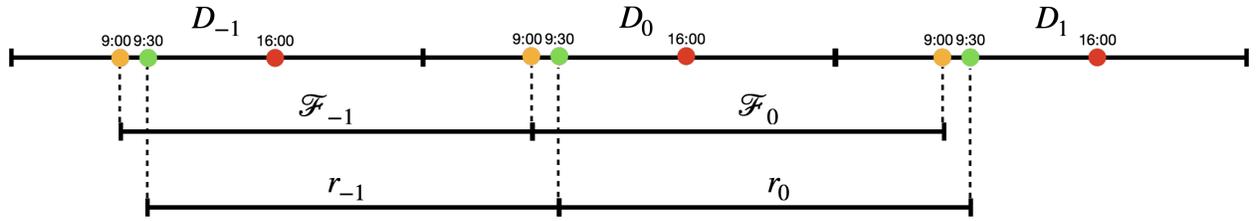

Figure 1: Definition of Timeline of Information
This figure demonstrates the definition of the information timeline for computing stock returns and news networks formation. This upper line is the calendar day synchronized to the New York time, the middle line defines the information set on each day-t, $\mathcal{F}_t$, as 09:00 am on day-t to 09:00 am next day, and the bottom line defines the stock return on day-t, r_t with open prices at 09:30 am on day-t and day-$t+1$.

Lead return, $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t})$, is essentially a function of the network $[\omega_{ij,t-l:t}]$. With the network decomposition defined in Section 2.1.1, *Lead return* also have the decomposed variants, accordingly.

We define the the *within-sector lead return* as $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t}^w)$ and the *cross-sector lead return* as $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t}^c)$, based on the decomposition in Equation (3). We further define the *big size lead return* as $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t}^{BigLead})$ and the *small size lead return* $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t}^{SmallLead})$ using the definition in Equation (4). Similarly, according to Equation (5), we define the *high liquid lead return* as $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t}^{HighLead})$ and *low liquid lead return* as $\mathcal{LR}_{i,t}(\omega_{ij,t-l:t}^{LowLead})$.

Furthermore, given the sign of lead firm return, *positive lead return* $\mathcal{LR}^+(\omega_{ij,t-l:t})$ and *negative lead return* $\mathcal{LR}^-(\omega_{ij,t-l:t})$ are defined as,

$$\mathcal{LR}_{i,t}^{+(-)}(\omega_{ij,t-l:t}) \overset{\text{def}}{=} \sum_{j=1}^n \omega_{ij,t-l:t} \cdot \left| r_{j,t}^{+(-)} \right|. \tag{7}$$

Notice that here we take the absolute value to compute $\mathcal{LR}^-(\omega_{ij,t-l:t})$. The decomposition by the sign of lead firm return helps us to understand if positive and negative returns of a

8

lead firm impact differently on the follower firm.

2.2.2. News Network Attention: Network Degrees

We employ the *network degree* as a "network attention" proxy. The attention proxy, *network degree*, $\mathcal{A}_{i,t}(\omega_{ij,t-l:t})$, for a firm i in a network $[\omega_{ij,t-l:t}]$ is defined as

$$\mathcal{A}_{i,t}(\omega_{ij,t-l:t}) = \#\{j : \omega_{ij,t-l:t} \neq 0\} + \#\{j : \omega_{ji,t-l:t} \neq 0\}, j = 1, 2, \cdots, n, \qquad (8)$$

where $\#\{j : \omega_{ij,t-l} \neq 0\}$ counts the number of non-zero entries in i_{th}-row in $[\omega_{ij,t-l:t}]$, likewise for $\#\{j : \omega_{ji,t-l} \neq 0\}$. This proxy is a well-known fundamental measure that counts the number of connections a node has, and the distribution of it describes how tight do nodes connect. In the following content, we use attention and *network degree* interchangeably.

Compared with counting the number of firm-specific articles of a firm, *network degree* counts the number of linkages of the firm, so that it reduces the noise caused by repetitive news on certain firms. The intuition of using *network degree* is that a firm with higher *network degree* implies more appearance with many other firms, either in headlines as lead firm or content as follower firm. For instance, in 2020, the automobile manufacturer Tesla had an unprecedented media exposure, when it is covered by media news in the headlines and mentioned in the news content of some other firms like General Motors and Ford. [2]

The attention proxy, *network degree*, $\mathcal{A}_{i,t}(\omega_{ij,t-l:t})$, is also a function of the given news network $\omega_{ij,t-l:t}$. Therefore, the attention variants can be defined based on the network decomposition in Section 2.1.1. We define the *network degree* of the network that consists linkages formed by firms from same (different) sector(s) as *within(cross)-sector attention* noted as $\mathcal{A}_{i,t}(\omega_{ij,t-l:t}^{w(c)})$; *network degree* of the network that consists linkages of big (small) size lead firm as *big (small) lead attention* noted as $\mathcal{A}_{i,t}(\omega_{ij,t-l:t}^{BigLead(SmallLead)})$; and *network degree* of the network that consists linkages of more (less) liquid lead firm as *high (low) lead attention* noted as $\mathcal{A}_{i,t}(\omega_{ij,t-l:t}^{HighLead(LowLead)})$.

One would notice that *network degree* consists of two parts that counting the row and column connections in the adjacency matrix $[\omega_{ij,t-l:t}]$. We define the *lead attention* as $\mathcal{A}^{lead}(\omega_{ij,t-l:t}) \overset{\text{def}}{=} \#\{j : \omega_{ji,t-l} \neq 0\}$, which essentially counts the number of different follower firms linked to lead firm i in a time period $[t - l : t]$. Similarly, *follower attention* is defined as $\mathcal{A}^{follower}(\omega_{ij,t-l:t}) \overset{\text{def}}{=} \#\{j : \omega_{ij,t-l} \neq 0\}$ that counts the number of different lead firms that link to firm i.

---

[2]Many automobile manufacturers like General Motors and Ford either released their electric vehicles or announced their plans for electric cars.

2.3. Econometric Model

To test the relationship between *lead returns* and stock returns, the following panel model is employed.

$$y_{i,t+h} = \alpha_i + \mathcal{LR}_{i,t}^{\top}(\omega_{ij,t-l:t})\beta + \sum_{k=1}^{m} \gamma_k \cdot control_{i,t}^{k} + \varepsilon_{i,t}, \qquad (9)$$

where the dependent variable, $y_{i,t+h}$, is the simple h-day ahead stock return $r_{i,t+h}$ or the stock residual return $r_{i,t+h}^{resid}$ from an asset pricing model. The main regressors, $\mathcal{LR}_{i,t}^{\top}(\omega_{ij,t-l:t})$, is a vector of *lead return* variables on the news network $[\omega_{ij,t-l:t}]$ constructed with l-day window size. Control variables can be employed to control for firm characteristics.

The parameters of fixed effect model in Equation (9) are estimated by pooled OLS estimator with standard error corrected by clustering over individuals and time. The β of our main interest estimates the averaged effect between *lead return* variables and the dependent variables across all the individual stocks. Specifically, β estimates the comovement effect between when $h = 0$, and lead-lag effect when $h >= 1$. α_i is employed to capture the possible fixed effects.

We employ the two stock residuals return variables from the Fama-Frech 3-factor model (FF3, Fama and Macbeth (1973)) and Fama-Frech 5-factor model (FF5, Fama and French (2015)). FF5 can be written as,

$$r_{i,t} - r_t^{f} = \theta_i + \theta_i^{m}(r_t^{m} - r_t^{f}) + \theta_i^{s} f_t^{smb} + \theta_i^{h} f_t^{hml} + \theta_i^{r} f_t^{rmw} + \theta_i^{c} f_t^{cma} + r_{i,t}^{resid}, \qquad (10)$$

where $r_{i,t}$ is the simple stock return, r_t^{f} is the risk-free return, r_t^{m} is the market return. The 3-factor model includes risk premium factor $r_t^{m} - r_t^{f}$, size factor f_t^{smb}, and value factor f_t^{hml}. The 5-factor model adds profitability factor f_t^{rmw} and investment factor f_t^{cma} to the 3-factor model. With all coefficients estimated by OLS in full sample, one can easily have the residual return from FF5, $r_{i,t}^{resid}(FF5)$, and FF3, $r_{i,t}^{resid}(FF3)$.

The residual return from Equation (10) is the part of stock returns unexplained by those factors and idiosyncratic stock alpha. Therefore, $r_{i,t}^{resid}(FF5)$ exclude the effects that are caused by the the Fama-Frech 5-factor. Similarly, $r_{i,t}^{resid}(FF3)$ excludes the effects from Fama-French 3-factor.

2.4. Portfolio Strategy by Sorting Characteristics

The portfolio strategy of sorting a characteristic aims to test the cross-sectional effect of the characteristic without imposing some assumptions, e.g., linear restriction, in econometric models. The trading rule is simple. Consider a characteristic to test, $x_{i,t}$, for each individual

stock i at time t, all the stocks are sorted by $x_{i,t}$ into K portfolios. Within each portfolio, stocks are equal-weighted or value-weighted. All the portfolios are held for one-period and then rebalanced for the next period.

The main point of portfolio strategy is to see whether a long-short trading rule based on sorting $x_{i,t}$ can deliver abnormal returns. In an efficient market, one shall not see an abnormal return with this simple trading strategy. While a positive abnormal return from the long-short portfolio suggests the existence of anomaly returns, which can be profited by investors.

2.4.1. Double Sorting Method for Robustness

Problem about sorting stocks based on their characteristic is that the characteristic only captures some of those known effects, e.g., size effect, therefore, it is imperative to test the robustness of the characteristic by controlling for those known effects.

The double-sorting method (Ang, Hodrick, Xing, and Zhang (2006)) is employed to relieve the results from the four possible effects including size, value, liquidity, and momentum. The double-sorting method is illustrated as follows. To control the effect from a characteristic, $c_{i,t}$, of firm i at time t, the assets are first sorted into K portfolios based on $c_{i,t}$. Then, within each portfolio, the assets are again sorted into K portfolios by the characteristic $x_{i,t}$ in ascending order. Each portfolio in the second sort is formed into a new portfolio across the K portfolios in the first sort. As such, the newly formed portfolios are relieved from the effect of $c_{i,t}$ since each new portfolio contains assets from all portfolios in the first sort.

3. Empirical News Networks of S&P500 Stocks

3.1. Ticker Identification

Uncovering the networks from news articles is hindered by the lack of a reliable approach to recognize the firms. The challenges are twofold. First, most articles are written by financial analysts or reporters from different agencies, such as the media giant Reuters, the emerging research company Zacks. Each of them has its way of referring to a firm's name. Take the tech giant *"Apple.Inc"* for example, Some articles have a strict form of reference, e.g. *"(NASDAQ: AAPL)"*, others might allow any casual ways, e.g. *"Apple"*. Second, recognizing a firm's name is a typical task that heavily relies on context awareness. For example, "Booking" can either refer to the company or a human act. Previous literature deals with the problem in two ways. One way is to directly use the Thomson-Reuters News Analytics (TRNA) that labels the companies mentioned in the news (Scherbina and Schlusche (2015); Chen, Guo, and Tu

(2021); Guo and Li (2019)). Recently, Schwenkler and Zheng (2019) employs the Natural Language Processing (NLP) tools and clustering algorithms to match the company names. However, neither of the two ways above is perfect in the sense of transparency and precision.

3.1.1. An Algorithm for Identification

We propose to tackle this problem with a conservative algorithm. The algorithm is designed to reduce the type-I error, i.e., to reduce the "false link" error and buy the "false unlink" error. We only identify the names/tickers of the given set of firms with the following three strategies, and for each strategy, we set multiple exceptional cases:

S1: Tickers in the brackets.

> *S1* is the main way among the three to identify firms as it is the safest way. We first manually configure a set of exceptional cases that can *NOT* be identified. Note that this set is designed for the given set of firms to identify, e.g., in our scenario {"PEG," "COO," "C," "GPS"...}, then we match the characters that are in the brackets. After differentiating the exceptional characters, we can have the clean tickers in the brackets.

S2: Name segments of the firms.

> The reason of identifying firm name segments instead of their full names is that people rarely use the suffixes in the full name of firms, e.g., *"JPMorgan Chase & Co"* is usually referred to as *"JPMorgan Chase"* or *"JPMorgan"*. To identify the name segments, we first need to map the full names of all the firms to their n-gram name segments detailed in Algorithm2. Then, we simply match the name segments in the text to identify corresponding firms. We carefully configure the redundant strings for the given set of firms to identify.

S3: Tickers in plain text.

> *S3* is employed to identify firms with long tickers like "GOOGL" only in the news headlines and only when neither *S1* nor *S2* could find any matches. In our experiment, we find the false identification error increases quite much if *S3* is applied on the contents due to a lot of unexpected short initials, such as *"HD"*, which is commonly short for *"high-definition"* while it is also the ticker of the firm *"Home Depot Inc"*.

We treat headline text and content differently with the above three strategies. For headlines, we prioritize the strategies with the order of $S1 > S2 > S3$ by a weak assumption

that the way all the firm names mentioned in the same headline is consistent, i.e., the lower priority strategy effects only when, the higher one finds no result. And for content text, only the first two strategies, $S1 and S2$, are employed as $S3$ could cause some apparent misidentifications. A flowchart of the identification process is illustrated in Figure 2 with an example.
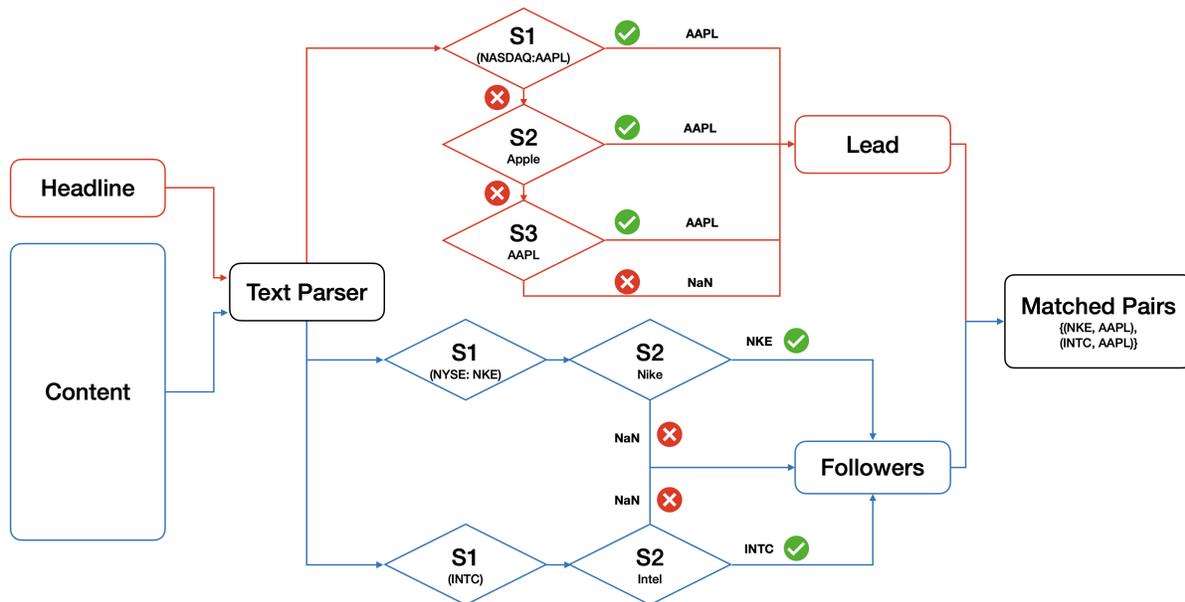
Figure 2: Flowchart of Identification Algorithm
A flowchart representation of how the algorithm treats the headline and content differently, and the priority of the three identification strategies.

3.1.2. Screening the Identified Pairs

We further employ two steps to screen the articles after the identification so that our study is focused on single firm-specific news. Firstly, we screen out the articles that have more than one lead firm identified. This step mainly helps to eliminate those market summary reports that have few major companies on their headlines. Such reports do not fit in the scope of our definition of lead firm. Secondly, we remove the articles that have more than 10 followers identified. This removal is aimed to overcome the errors caused by some of the market news mentioning an iconic firm in their headlines. A typical example is shown in the Appendix Figure 13, where 12 companies are identified in the content, apart from the single lead company. Apparently, articles like the sample are not firm-specific news. This truncation affects only a small number of linkages, as one can see that the frequency decreases dramatically as the number of followers increases shown in the follower distribution plot in Figure 3.
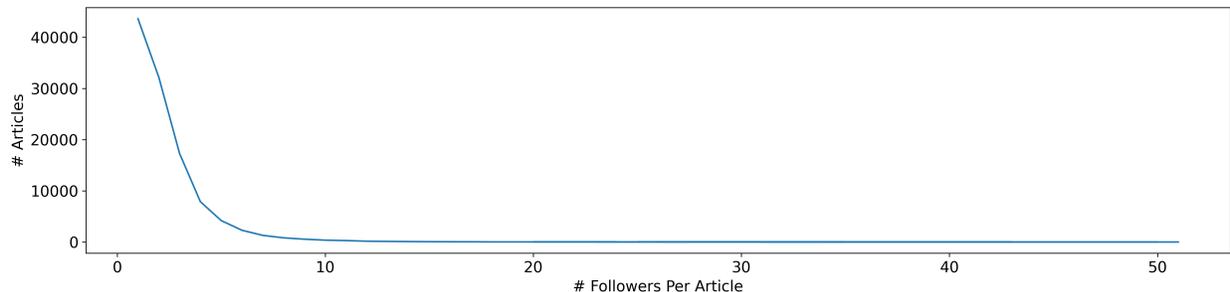
Figure 3: Distribution of Followers in Each Article
This figure shows the number of followers identified in each article before the screening. Most of the articles have the number of followers identified under 10.

We provide pseudocode in Algorithm 1 in Appendix. As the algorithm narrows down the scenarios that a firm could be mentioned, the design allows the errors that many linkages are not identified. However, for the purpose of testing the effects of the linkages, we will have more robust and conservative results. Surely, one can attempt to handle this task with more sophisticated tools, which, however, also introduce more statistical modeling errors.

3.2. News Linkages Identified Among S&P500 Constituents

In this paper, we focus only on the firms that are the constituents of S&P500. S&P500 consists of the companies with large capitalization listed in the U.S stock market and is one of the most commonly studied indices. It is also the underlying asset of many financial instruments, including ETFs, futures, and options. The choice is a trade-off between generality and efficiency of identification. One could adapt the algorithm to consider a broader range of assets, e.g., small firms or even listed funds, which, however, would cause more false links to be mismatched even after a huge effort of specifying special cases.

To identify the linkages from the media news, we collect a huge dataset of public articles from more than 50 online publishers on nasdaq.com news platform. The early period is removed due to the sparsity of the news arrivals, and we utilize the 5-year sample period from Jan. 2016 until the end of 2020 and obtain 1,027,272 articles in total, averaging 764 articles each trading day. Apart from the financial market and stock news as the main type of articles, the dataset also consists of analytical reports, comments, and even investor meeting notes and covers topics including global economics, commodities, etc. Note that all the articles are economic and financial news, and we keep all the articles of different topics as the input of the algorithm.

As shown in Figure 4, along with the booming of digital news, the volume of daily articles gets higher. Apart from that, the volume of news articles evolves quite periodically. More

specifically, we can observe the 4-peak each year, caused by the massive quarterly and annual financial reports releases. We mark the holidays (in green star) and the weekends (in the red cross) in Figure 4 to contrast the news volumes on the trading days. Thus, one can see that most articles are released on trading days.
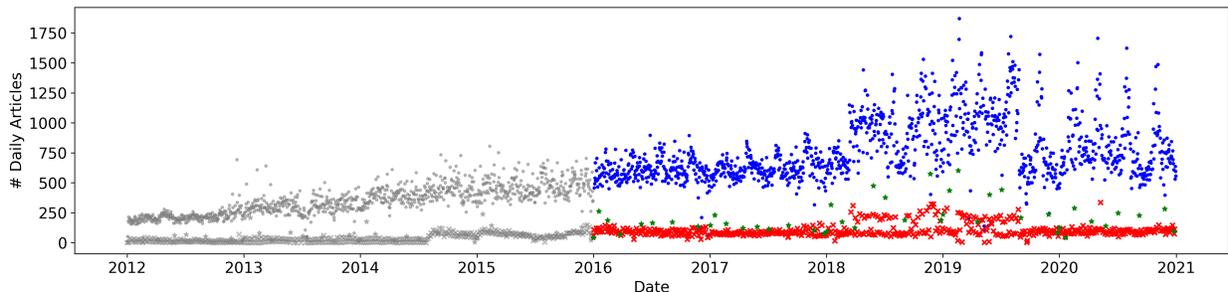
Figure 4: Number of Articles in The Database
This plot shows the number of articles daily aggregated from Jan. 2012 until Dec. 2020. We only utilize sample starting from Jan. 2016, so the grey dots are the samples not used. The blue dots indicate the trading days, the red crosses represent the weekends, and the green stars are the market holidays.

Applying the algorithm on the collected dataset, 100,960 articles out of the total 1,027,272 articles are identified with S&P500 constituents, and we obtain 204,074 pairs of linkages in those articles during the whole 5-year sample period. Among that, 195,017 pairs of linkages are identified in trading days. As shown in Figure 5, we can observe a strong periodicity of the daily linkages number. The number of pairs in trading days shows four local maximum each year, which is consistent with the daily volume of news arrivals.

A typical example of the articles that are identified with linkages of the news network is documented in Appendix Figure 12. The algorithm uncovers two pairs of linkages: {(INTC, AAPL), (NKE, AAPL)}. From the title, it is easy to recognize that this article is mainly covering the listed company, *"Apple"*. And the article content also mentions *"Intel"* and *"Nike"*, both of which are the business partners of *"Apple"*. One might have noticed that the listed company, *"Peloton Interactive"*, is not identified as the company was not included in S&P500 during that sample period.

One interesting fact is the fairly large amount of cross-sector linkage. The identified result shows that almost half of the linkages are cross-sector pairs, i.e., the lead and follower are from different sectors. Among the 204,074 pairs of linkages, 99,047 pairs are cross-sector pairs, and 105,027 pairs are intra-sector. The top-3 cross-sector linkage types are:

- Communication Services / Information Technology: 15,071
- Financial / Information: 8,778
- Consumer Discretionary/ Information Technology: 7,993

15
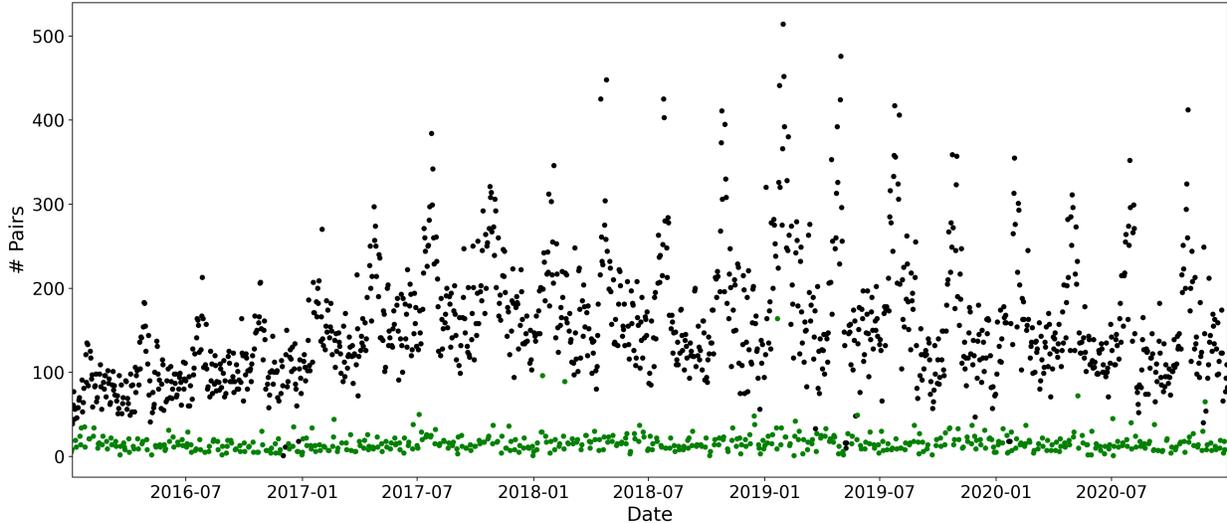
Figure 5: Number of Linkages Identified Daily
This figure shows the number of linkages identified each day over the 5-year sample period, from Jan. 2016 to Dec. 2020. The dark color bar represents the trading days and the green bar indicates the weekends and market holidays.

3.3. Summary of News Network Variables

A news network, $[\omega_{ij,t-l:t}]_{n \times n}$, is formed with linkages aggregated in period from day-$t-l$ to day-t, and the n nodes of the network are the monthly updated S&P500 constituents obtained from Bloomberg. As such, the news networks is time-varying. The S&P500 stocks market data is provided by Compustat on Wharton Research Data Services (WRDS).

We employ a 1-year window network, $[\omega_{ij,t-365:t}]$, to compute the *lead return* variables, and a 1-month window network, $[\omega_{ij,t-30:t}]$, for the *network degree* variables. Panel A in Table 1 documents the *lead return* variables with 1-year window news networks along the daily return variables, and Panel B reports the *network degree* variables with 1-month window news network, as well as the monthly stock return variables.

In the literature of discussing the formation of networks, two of the models are the most prominent and well investigated. The first model named Erdős-Rényi network depicts the linkages being randomly formed with probability p, which is governed by the binomial distribution or Poisson distribution. Secondly, the scale-free network characterized by power-law distribution was discussed in the late 1990s and early 2000s. It is well documented that many networks in social science and economics carry with the scale-free property(Barabási and Albert (1999); Albert, Jeong, and Barabási (1999); Jackson (2010)). The power-law distribution can be written as, $p(d) = cd^{-\gamma}$, where c is a constant scaler and d is the degree of nodes. The degree exponent γ is the coefficient in the linear function of $\log(d)$ when the distribution is represented in the log-log form, i.e $\log(p(d)) = \log(c) - \gamma \log(d)$.

16

Table 1
Summary Statistics

	Mean	Std.	Median	10%	25%	75%	95%
Panel A: Daily Return & *Lead Return* of 1-Year Window Network $T = [t - 365 : t]$							
r	0.026	2.187	0.079	-1.989	-0.821	0.940	2.905
$r^{resid}(FF3)$	-0.016	2.077	0.019	-1.872	-0.818	0.831	2.677
$r^{resid}(FF5)$	-0.015	2.065	0.020	-1.860	-0.814	0.828	2.659
$\mathcal{LR}(\omega_{ij,T})$	0.033	1.497	0.045	-1.346	-0.513	0.656	2.011
$\mathcal{LR}(\omega_{ij,T}^{w})$	0.028	1.686	0.000	-1.509	-0.576	0.704	2.252
$\mathcal{LR}(\omega_{ij,T}^{c})$	0.036	1.475	0.000	-1.303	-0.455	0.619	2.003
$\mathcal{LR}^{+}(\omega_{ij,T})$	0.657	0.894	0.402	0.000	0.130	0.854	2.143
$\mathcal{LR}^{-}(\omega_{ij,T})$	0.625	1.002	0.319	0.000	0.093	0.761	2.232
$\mathcal{LR}(\omega_{ij,T}^{BigLead})$	0.042	1.418	0.006	-1.292	-0.467	0.637	1.955
$\mathcal{LR}(\omega_{ij,T}^{SmallLead})$	-0.001	2.101	0.000	-1.755	-0.561	0.647	2.695
$\mathcal{LR}(\omega_{ij,T}^{HighLead})$	0.018	2.080	0.000	-1.817	-0.643	0.768	2.756
$\mathcal{LR}(\omega_{ij,T}^{LowLead})$	0.035	1.284	0.000	-1.176	-0.421	0.579	1.777
Panel B: Monthly Return & *Network Degree* of 1-Month Window Network $T = [t - 30 : t]$							
r	0.623	9.848	1.275	-9.432	-3.527	5.640	13.976
$r^{resid}(FF3)$	-0.276	8.180	0.199	-8.785	-4.011	4.037	10.929
$r^{resid}(FF5)$	-0.253	8.126	0.222	-8.753	-3.934	4.062	10.886
$\mathcal{A}(\omega_{ij,T}))$	8.778	14.215	5.000	0.000	2.000	11.000	27.000
$\mathcal{A}^{lead}(\omega_{ij,T})$	4.390	5.743	3.000	0.000	1.000	6.000	14.000
$\mathcal{A}^{follower}(\omega_{ij,T})$	4.388	10.283	2.000	0.000	0.000	5.000	15.000
$\mathcal{A}(\omega_{ij,T}^{w})$	4.105	4.846	3.000	0.000	1.000	6.000	13.000
$\mathcal{A}(\omega_{ij,T}^{c})$	4.673	10.834	2.000	0.000	0.000	5.000	16.000
$\mathcal{A}(\omega_{ij,T}^{BigLead})$	4.838	10.016	1.000	0.000	0.000	5.000	22.000
$\mathcal{A}(\omega_{ij,T}^{SmallLead})$	1.156	2.617	0.000	0.000	0.000	1.000	6.000
$\mathcal{A}(\omega_{ij,T}^{HighLead})$	2.119	4.660	0.000	0.000	0.000	2.000	10.000
$\mathcal{A}(\omega_{ij,T}^{LowLead})$	3.287	6.678	1.000	0.000	0.000	4.000	14.000

Panel A reports the summary statistics of daily stock (residual) return variables and *lead return* variables on a 1-year window size news network. Panel B reports the summary statistics of monthly (residual) return variables and *attention* variables on a 1-month window news network. All statistics in Panel A are given in percentage, as well as the (residual) return variables in Panel B. Columns from left to right are mean, standard deviation, median, 10% percentile, 25% percentile, 75% percentile, 95% percentile of respective variables.

The degree distribution of the monthly aggregated news network falls quite well into the power-law distribution with $\gamma = 2.12$ and $R^2 = 0.914$ as shown in Figure 6 where the relationship between $\log(Frequency)$ and $\log(Degree)$ is approximated with a linear function. The power-law distributed news network implies the existence of highly connected

nodes or hubs (Barabsi (2016)), i.e., a portion of the companies are mentioned much more frequently than the other companies.
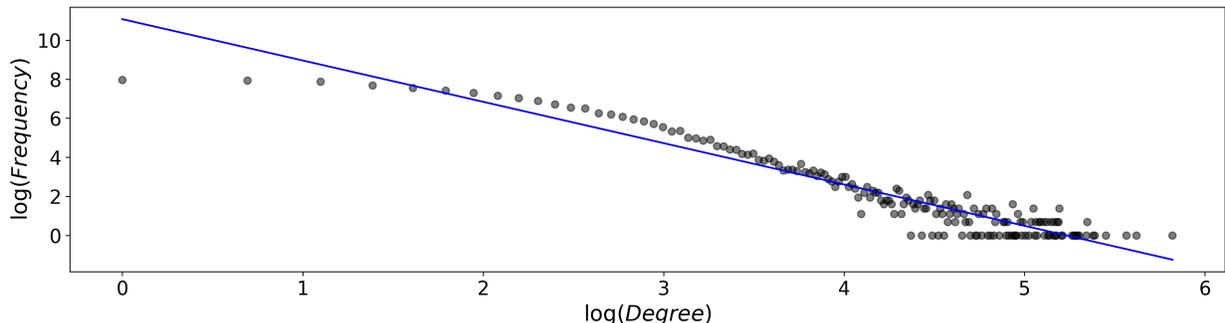
Figure 6: Network Degree Distribution of Network Aggregated Monthly
The log-log plot is the logarithmic frequency, $\log(frequency)$, against its logarithmic network degree, $\log(\mathcal{A})$. The solid line is the linear fitting estimated by OLS, and the slop coefficient $\gamma = 2.12$ and $R^2 = 0.914$. Such a pattern is often deemed as power-law distribution.
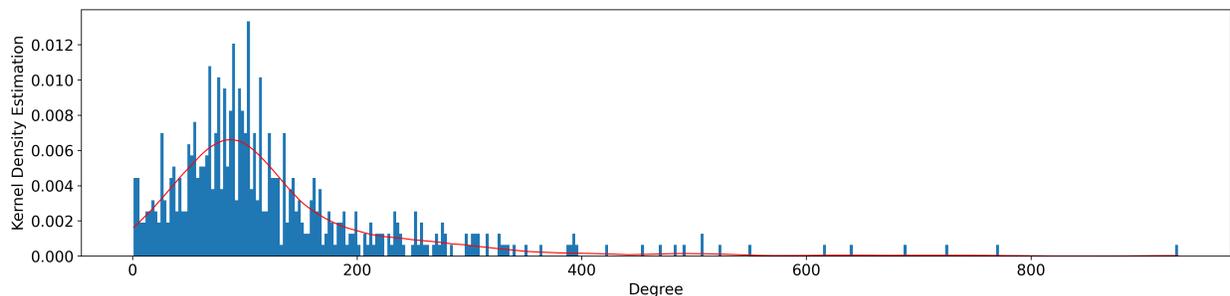
Figure 7: Network Degree Distribution of Full-Sample Network
This figure shows the network degree distribution of the network formed with full-sample data. The bars represent the histogram and the solid line is the kernel density estimated with Epanechnikov kernel.

Interestingly, when the news linkages are aggregated for the whole sample period, the degree distribution of the full sample news network can be well described as a randomized network as shown in Figure 7. Furthermore, the network degree of that network is closer to a Poisson distribution by Epanechnikov kernel. Note that the full-sample news network is constructed with all the historical S&P500 constituents, i.e., network size of 596.

4. *Lead Returns* and Cross-Sectional Stock Returns

The section shows how the returns of the lead firms, *lead return*, interact with returns of the follower firms in two cases. We first investigate the contemporaneous relationship. We find a strong comovement effect between the news-linked stock daily returns, even after

controlling for those known factors and firm characteristics. Then we focus on the case of 1-day ahead stock return, in which we give evidence of the reversal effect from *lead return* on the 1-day ahead stock return; however, the negative impact is offset.

4.1. Simultaneous Returns Test

The fundamental question on the constructed news network is how the linkages between firms identified relate to the performance of the stock returns. In this subsection, we provide evidence to justify the validity of news linkages by both econometric and portfolio tests.

Surprisingly, we found a strong stock price comovement effect after controlling for the known factors and firm characteristics, without recognizing an economic relationship between the news-linked firms, for example supplier-customer relationship. This fact implies two statements. First, the linkages in news articles are informative in stock returns. Second, the lead and follower firms are more likely to have a positive economic correlation than a rivalry relationship, as their daily stock returns move in the same direction.
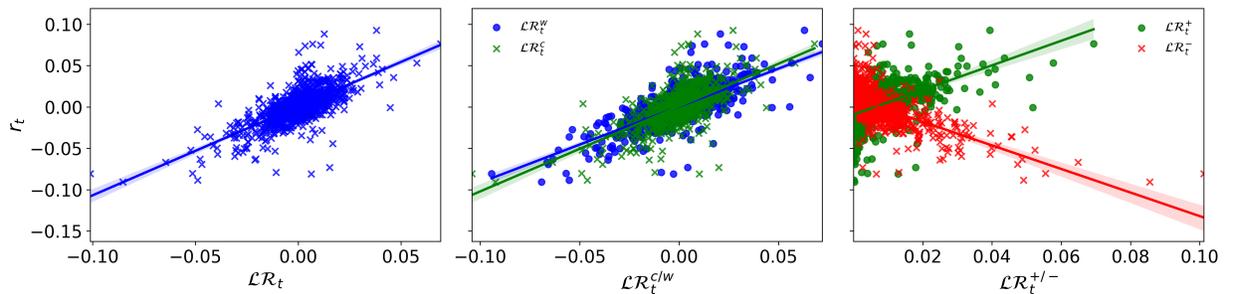
Figure 8: Stock Return Against Simultaneous Lead Return, AAPL
Plots from left to right are AAPL daily return r_t against its lead return $\mathcal{LR}(\omega_{ij,t-365:t})$, sector decomposed variants $\mathcal{LR}\left(\omega^{w/c}_{ij,t-365:t}\right)$, and the return sign decomposed variants $\mathcal{LR}^{+/-}(\omega_{ij,t-365:t})$, respectively. The solid lines are estimated by OLS with bootstrap of 95% confidence interval.

The left panel of Figure 8 shows that the daily return of Apple, r_t, positively correlates with its contemporaneous *lead return* formed with a 1-year rolling window network, $\mathcal{LR}(\omega_{ij,t-365:t})$. The strong positive correlation stays after decomposing *lead return* by sector, $\mathcal{LR}(\omega^{w(c)}_{ij,t-365:t})$, as shown in the middle of Figure 8. It is shown in the right plot that the effect remains true when *lead return* is separated by the sign of lead firm returns, $\mathcal{LR}^{+(-)}(\omega_{ij,t-365:t})$.

Motivated by this scatter plot, we conjecture that the news linkages capture the comovement effect of stock returns. Therefore, we test the relationship between *lead returns* and the contemporaneous stock returns by the panel model in Equation 9 with horizon $h = 0$.

One would naturally believe that the comovement of stock returns can be explained by the market trend, the within-sector comovement, and etc. Here we employ the two stock

Table 2
Estimation of Stock Returns Comovement Effects

	$r_{i,t}^{resid}(FF3)$					$r_{i,t}^{resid}(FF5)$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\mathcal{LR}_{i,t}(\omega_{ij,T})$	0.752 (56.984)					0.744 (58.020)				
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{w})$		0.622 (36.463)					0.610 (36.702)			
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{c})$		0.121 (11.496)					0.127 (12.503)			
$\mathcal{LR}_{i,t}^{+}(\omega_{ij,T})$			0.741 (54.584)					0.732 (55.576)		
$\mathcal{LR}_{i,t}^{-}(\omega_{ij,T})$			−0.761 (−51.768)					−0.755 (−52.526)		
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{BigLead})$				0.529 (41.797)					0.526 (42.255)	
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{SmallLead})$				0.214 (17.092)					0.211 (17.480)	
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{HighLead})$					0.246 (16.860)					0.244 (17.262)
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{LowLead})$					0.536 (40.903)					0.532 (41.498)

A 1-year window $T = [t - 365 : t]$ for news network formation. The left (right) panel reports the dependent variable of residual returns from Fama-French 3(5)-factor models. Estimated coefficients of *lead returns* for each model are reported in each column, respectively. *t*-statistics are reported in the parentheses, with standard error corrected by clustering on individual and time. Control variables, including market value $\log(MV)$, book-market ratio B/M, and turnover rate *volume/share*, are employed in each model. All the financial statement related variables are 2 quarters lagged.

residual return variable, $r_{i,t}^{resid}(FF3)$ and $r_{i,t}^{resid}(FF5)$, so that we estimate the comovement effect between *lead return* and the part of stock returns unexplained by those factors. Four fundamental variables of stocks are used as control variables, including the daily market value in the logarithmic form $\log(MV)$, daily book-market value B/M, daily turnover rate *volume/share*. Following Fama and French (2015), the financial statement data including, market share, and book equity B, are two quarters lagged values when they are used to compute the control variables.

We first test *lead returns* formed with 1-year rolling window, $\mathcal{LR}(\omega_{ij,t-365:t})$. Due to the sparsity of the news arrivals, the 1-year window is a reasonable choice between testing the persistent linkages and capturing the innovative linkages. The results are consistent with a 6-month rolling window. Column (1) in Table 2 reports that *lead return* carries positive (0.752) comovement effect, with the significant (*t*-statistics= 56.984) relationship between

$\mathcal{LR}(\omega_{ij,t-365:t})$ and the residual return from FF3 $r_{i,t}^{resid}(FF3)$ at 1% level. This implies that a one-standard-deviation increase in *lead return* (1.497%) accompanies with an average 112 bps (1.497% * 0.752) of excessive return increase across all the stocks on the same trading day. Furthermore, a robust check with FF5 reported in Column (6) shows the consistent result. It is evidenced that the follower daily stock returns comove with their news-linked lead firms daily stock returns.

As shown in the Column (2) of Table 2, the regression replaces *lead return* from Column (1) with its decomposed variants by sector. Both *within-sector lead returns* and *cross-sector lead returns* positively correlated with the residual returns from FF3 of individual stocks at 1% level. The $\mathcal{LR}_{i,t}(\omega_{ij,t-365:t}^{w})$ shows a stronger positive relationship (0.622, *t*-statistics= 36.463) compared to the $\mathcal{LR}_{i,t}(\omega_{ij,t-365:t}^{c})$ (0.121, *t*-statistics= 11.496), implying that the comovement effect between news-linked firms is more prominent within the same sector. However, the cross-sector comovement effect is not ignorable.

The comovement effect still exists after accounting for the sign of lead firm returns. As reported in Column (3) of Table 2, the regression shows that *positive lead return* $\mathcal{LR}_{i,t}^{+}(\omega_{ij,t-365:t})$ positively relates (0.741, *t*-statistics= 54.584) with $r_{i,t}^{resid}(FF3)$ and *negative lead return* $\mathcal{LR}_{i,t}^{-}(\omega_{ij,t-365:t})$ has negative relationship (-0.761, *t*-statistics= -51.768) with $r_{i,t}^{resid}(FF3)$.

Then, we see interesting results after testing the firm size and firm liquidity decomposed variants reported in Column (4)-(5) of Table 2, which shows the asymmetric comovement effects depending on the size and stock liquidity of lead firms. Although all the *lead return* variables have significant relationship between the stock residual returns, one can see the *big size lead return* $\mathcal{LR}_{i,t}(\omega_{ij,t-365:t}^{BigLead})$ has a much stronger comovement effect (0.529, *t*-statistics= 41.797) with stock residual return compared with *small size lead return* $\mathcal{LR}_{i,t}(\omega_{ij,t-365:t}^{SmallLead})$ (0.214, *t*-statistics= 17.092), and surprisingly, the *low liquidity lead return* $\mathcal{LR}_{i,t}(\omega_{ij,t-365:t}^{LowLead})$ shows stronger comovement effect (0.536, *t*-statistics= 40.903) as opposed to a weaker comovement effect from *high liquidity lead return* $\mathcal{LR}_{i,t}(\omega_{ij,t-365:t}^{HighLead})$ (0.246, *t*-statistics= 16.860).

Consistently, with replacing the residual returns from FF3 with the ones from FF5, all the above conclusions stay true and significant at 1% level as shown in Columns from (6)-(10) in Table 2.

4.1.1. *Infeasible Portfolio Test*

The econometric insight of the result above make it evident that if a firm has higher *lead return*, it has a higher stock return at the same day. We apply an infeasible portfolio test in the real stock market to relieve the model restrictions and draw a consistent conclusion.

We conduct the portfolio test by sorting the S&P500 constituents based on the con-

temporaneous *lead returns* as the method described in Section 2.4. The news networks are constructed with a 1-year rolling window. Also, the 6-month rolling window networks (un-reported) deliver similar results. The stocks are sorted into quintiles, and the portfolios are rebalanced each day. Firms in quintile 1 have the lowest *lead return* values, while firms in quintile 5 have lead firms performing the best. One can notice that none of the portfolios is feasible. However, the point of those infeasible portfolios is meant to prove the comovement effect between the news-linked firms with a simple trading rule.

Table 3
Infeasible Portfolios of Sorting Contemporaneous *lead return* $\mathcal{LR}(\omega_{ij,t-365:t})$

						FF3		FF5	
Rank	Mean	SR	%MV	B/M	Liquidty	α	R^2	α	R^2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	−82.41	−9.14	16.69	44.13	1.02	−85.03 (−25.42)	15.32	−85.02 (−25.60)	15.73
2	−25.74	−3.44	21.14	40.35	0.87	−28.31 (−10.78)	18.22	−28.40 (−10.86)	18.61
3	3.97	0.48	22.66	39.62	0.85	1.47 (0.59)	18.28	1.38 (0.56)	18.76
4	31.30	4.18	21.88	39.54	0.86	28.86 (11.21)	18.55	28.78 (11.28)	19.10
5	82.73	9.59	17.64	43.05	0.99	80.46 (24.73)	16.62	80.37 (24.99)	17.29

The table reports the performance of the equal-weighted portfolios sorted by *lead return*, $\mathcal{LR}(\omega_{ij,t-365:t})$, computed with news linkages aggregated in a 1-year rolling window. All the port-folios are infeasible. Column (1)-(5) are average simple daily return in bps; average annualized Sharpe-ratio, average market value proportion in percentage, average book-to-market ratio in per-centage, average volume-to-share ratio in percentage. The last 4 columns report the alpha, R^2, and robust Newey-West t-statistics in parentheses against Fama-French 3-factor and Fama-French 5-factor, respectively.

As reported in the first two columns of Table 3, the average simple returns of the equal-weighted portfolios monotonic increase from −82.41 bps to 82.73 bps per day, and the an-nualized daily Sharpe-ratio increases from −9.14 to 9.95. The last four columns of Table 3 under the heading FF3 and FF5 report the estimated alpha and R^2 for each portfolio in Fama-French 3 and 5-factor models, and the robust Newey–West t-statistics of the alpha is in the parentheses. Portfolio 5 has a significant positive alpha (80.46 bps) against the Fama-French 3-factor model compared with portfolio 1, which has a significant negative al-pha (−85.03 bps). Columns (3)-(5) under the heading "%MV", "B/M", "Liquidity" report

the market value proportion, book-to-market ratio, volume-to-share ratio for each quintile portfolio. One can see that there is no monotonic pattern for all those characteristics. For example, both the quintile 1 and quintile 5 portfolios have relatively lower market value, higher book-to-market ratio, and higher stock liquidity. Figure 9 showcases the cumulative returns of three of the quintile portfolios (lowest, medium, and highest) benchmarked by an equal-weighted S&P500 market portfolio. The cumulative returns of quintile 1 and quintile 5 diverge quickly over time, implying that a portfolio with a higher simultaneous *lead return* value consistently outperforms the portfolio with lower simultaneous *lead return*.
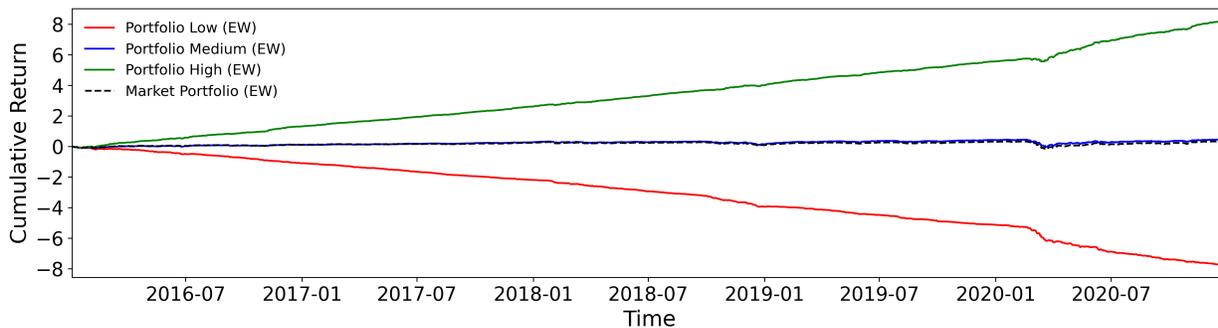
Figure 9: Portfolios Cumulative Returns of Sorting *Lead Return* $\mathcal{LR}(\omega_{ij,t-365:t})$
S&P500 constituents are sorted into quintile equal-weighted portfolios. The three solid lines in red, blue, green refer to cumulative returns of portofolios with lowest, medium, and highest *lead return* value. The black dash line is the benchmark cumulative return of an equal-weighted S&P500 market portfolio.

We also report robust check on other *lead return* variants including *within(cross)-sector lead returns* $\mathcal{LR}(\omega^{w(c)}_{ij,t-365:t})$, *positive(negative) lead returns* $\mathcal{LR}^{+(-)}(\omega_{ij,t-365:t})$ in Appendix, see Table 10 and Table 11. All the robust check results draw a consistent conclusion. However, it is worth noting that the difference between quintile 5 and quintile 1 sorted by *cross-sector lead return* (reported in the bottom panel of Table 10) is much smaller than that of those sorted by *within-sector lead return*. Those evidences strongly suggest that the comovement effect does exist between firms from different sectors, though it is less significant.

4.2. Lead Return Portfolio and Predictability

4.2.1. Momentum v.s Reversal

With the evidence on stock return comovement between news-linked firms, it is appealing to investigate how the returns of a lead firm impact the returns of its follower stocks in the future. It is documented in the early studies, e.g., Cohen and Frazzini (2008), that the customers' stock returns of a firm have a momentum effect on the firm's monthly stock returns. But the momentum effect is not true in our case for two plausible reasons. Firstly,

information transmitting much faster causes higher efficiency in the current stock market than 10-15 years ago. Secondly, caused by the fast information transmission, the *"overreaction"* phenomenon become significant, which explains the strong comovement effect between news-linked firms.

As shown in Figure 10, three scatter plots of the 1-day ahead stock returns r_{t+1} of Apple against multiple of its *lead return* variables, implying that there is potential reversal effect of *lead returns*. However, such a potential reversal effect could vanish for other stocks. To test the average effect across all S&P500 constituents, we conduct the panel regression in Equation (9) with $h = 1$. In addition to the firm characteristics employed testing comovement effect in 4.1, we also use the 1-day lagged stock return r_t as a control variable. 1-year rolling window size is used to construct the news networks.
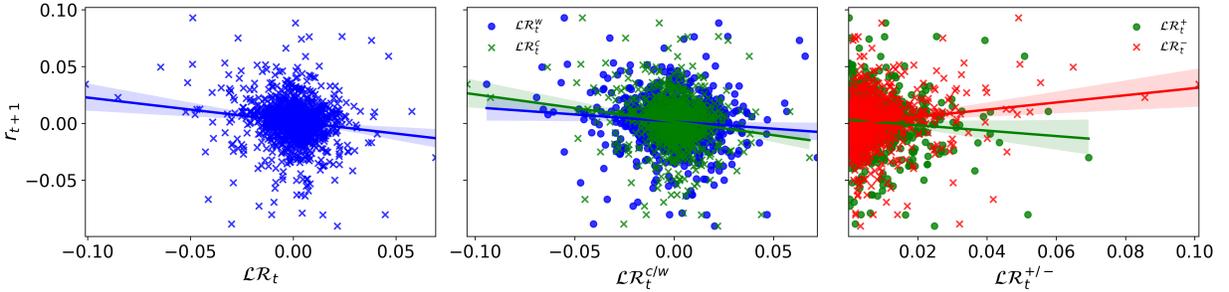
Figure 10: Stock Return Against 1-Day Ahead *lead return*, AAPL
Plots from left to right are AAPL 1-day ahead daily return r_{t+1} against *lead return* $\mathcal{LR}(\omega_{ij,t-365:t})$, sector decomposed variants $\mathcal{LR}\left(\omega_{ij,t-365:t}^{w/c}\right)$, and the return sign decomposed variants $\mathcal{LR}^{+/-}(\omega_{ij,t-365:t})$, respectively. The solid lines are estimated by OLS with bootstrap of 95% confidence interval.

Panels from left to right in Table 4 are estimations of the impacts of *lead returns* on 1-day ahead stock return r_{t+1}, stock residual return from FF3, $r_{t+1}^{resid}(FF3)$, and stock residual return from FF5, $r_{t+1}^{resid}(FF5)$, respectively. Column (1) of Table 4 reports that *lead return* shows insignificant effect (0.003) on the 1-day ahead stock return with (*t*-statistics=0.601), suggesting that one-standard deviation increase in $\mathcal{LR}(\omega_{ij,t-365:t})$ will cause only a 0.4 bps (1.497% ∗ 0.003) stock return increase in then next day r_{t+1}.

We can see the mixed results when investigating the *lead return* decomposed by sector and sign of lead returns. Column (2) of Table 4 reports the impact from the sector decomposed *lead return* variables. The *cross-sector lead return* shows significant reversal impact on 1-day ahead return with an estimated coefficient of -0.011 (*t*-statistics=-2.811) at 1% level, whereas the *within-sector lead return* indicates significant momentum effect on 1-day ahead stock return with an estimated coefficient of 0.010 (*t*-statistics=2.323) at 1% level. Column (3) of Table 4 reports the estimation when *lead return* is decomposed by the sign of lead

Table 4
Estimation of *lead returns* Impact on 1-Day Ahead Stock Returns

	$r_{i,t+1}$			$r_{i,t+1}^{resid}(FF3)$			$r_{i,t+1}^{resid}(FF5)$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\mathcal{LR}_{i,t}(\omega_{ij,T})$	0.003 (0.601)			−0.127 (−24.626)			−0.125 (−24.227)		
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{w})$		0.010 (2.323)			−0.061 (−12.452)			−0.062 (−13.036)	
$\mathcal{LR}_{i,t}(\omega_{ij,T}^{c})$		−0.011 (−2.811)			−0.087 (−20.445)			−0.082 (−19.787)	
$\mathcal{LR}_{i,t}^{+}(\omega_{ij,T})$			0.032 (5.124)			−0.066 (−9.697)			−0.064 (−18.650)
$\mathcal{LR}_{i,t}^{-}(\omega_{ij,T})$			0.021 (3.503)			0.178 (27.138)			0.176 (56.187)

A 1-year window size news network formation $T = [t − 365 : t]$. The first panel under the heading of r_{t+1} reports the estimation of the impact from *lead return* on stock returns. The middle (right) panel contains the estimation of the impact from *lead returns* on stock residual returns from FF3 (FF5). Estimated coefficients of *lead return* variables for each model are reported in corresponding column, and *t*-statistics is reported in the parentheses, with standard error corrected by clustering on individuals and time. Control variables, including market value $\log(MV)$, book-market ratio B/M, and turnover rate *volume/share* and 1-day lagged return r_t, are employed in each model. All the financial statement related variables are 2 quarters lagged.

firm return. The *positive lead return* component shows significant momentum effect with an estimated coefficient of 0.032 (*t*-statistics=5.124), while the *negative lead return* component indicates significant reversal effect with an estimated coefficient of 0.021 (*t*-statistics=3.503), both at 1% level.

A possible interpretation of the ambiguous reversal effect is the influences from other more dominant factors, such as the Fama-French factors. Columns (4)-(9) of Table 4 report the estimation in which the dependent variable, 1-day ahead stock return, r_{t+1}, is replaced by the stock residual returns, r_{t+1}^{resid}, unexplained by the Fama-French factor models. One can see the consistent and significant reversal effect of *lead return* variables on residual stock return. For instance, Column (4) of Table 4 shows that $\mathcal{LR}(\omega_{ij,t−365:t})$ impacts negatively on $r_{t+1}^{resid}(FF3)$ with coefficient estimated −0.127 (*t*-statistics=−24.626), implying a one-standard-deviation increase in $\mathcal{LR}(\omega_{ij,t−365:t})$ will decrease the residual stock return from FF3 by 19 bps (1.497% ∗ 0.127). The estimated results strongly suggest that *lead returns* contributes negatively to the stock returns in the next trading day. However, the reversal effect is offset by the momentum effect caused by other factors and firm characteristics.

4.2.2. Portfolio Test on lead returns

A portfolio test is performed to relieve the linearity restriction in the econometric model by sorting S&P500 stocks into quintile based on the preceding day *lead returns*. Here we mainly analyze the results for equal-weighted portfolios, and results from value-weighted portfolios are consistent.

$$\mathcal{LR}_{agg}(\omega_{ij,t-365:t}) = \mathcal{LR}^+(\omega^w_{ij,t-365:t,}) + \mathcal{LR}^-(\omega^w_{ij,t-365:t,}) + \mathcal{LR}^-(\omega^c_{ij,t-365:t,}) - \mathcal{LR}^+(\omega^c_{ij,t-365:t,})$$

$$(11)$$

The simple aggregated *lead return* proxy is an effort to capture the different impacts from all the decomposed *lead return* variables. However, the optimization of utilizing those variables is beyond this study.

We show that *lead return* variables carry only marginal predictability of 1-day ahead stock return. the portfolios are rebalanced by the end of each trading day and holds for the next trading day. As reported in Panel A of Table 5 sorting $\mathcal{LR}(\omega_{ij,t-365:t})$, although quintile 5 outperforms the equal-weighted market portfolio by a margin of 3.4 bps (Sharpe-ratio=0.38) against 2.8 bps (Sharpe-ratio=0.31), and the difference between quintile 5 and quintile 1 is positive and generates positive alphas with respect to FF3 and FF5 by 2.37 bps and 2.23 bps, respectively, nevertheless insignificant statistically. It is difficult to conclude the cross-sectional predictability in $\mathcal{LR}(\omega_{ij,t-365:t})$.

Panel B reports the performance of assets sorted by $\mathcal{LR}_{agg}(\omega_{ij,t-365:t})$, which shows a more clear pattern compared with the previous proxy $\mathcal{LR}(\omega_{ij,t-365:t})$, implying that it is possible to improve the predictability by calibrating the combination of those *lead return* variables. The quintile 5 portfolio generates 3.95 bps with a Sharpe-ratio of 0.43 compared with an average return of 1.73 bps and Sharpe-ratio of 0.35 from the equal-weighted market portfolio. Similar to Panel A, the difference between quintile 5 and quintile 1 offers positive alphas, and the Sharpe-ratio is only slightly better than the equal-weighted market portfolio. Note that we do not consider the transaction cost in the above portfolio test, which is essential for the daily rebalancing portfolio test.

Furthermore, we show that a reversal effect can be found in the stock residual returns r^{resid}_{t+1} of the portfolios sorted by the 1-day lagged *lead return*, $\mathcal{LR}(\omega_{ij,t-365:t})$. Motivated by the econometric results in Table 4, we compute the residual returns of the stocks in each of the portfolios, named "residual portfolios" of sorting $\mathcal{LR}(\omega_{ij,t-365:t})$ reported in Table 6. The portfolios are the same as the ones reported in Panel A of Table 5, only the performance is evaluated based on stock residual returns instead of simple stock returns.

Panel A in Table 6 reports the FF3 residual return $r^{resid}_{t+1}(FF3)$ performance of equal-

<div align="center">

Table 5
Portfolios by Sorting Preceding *lead return* variables (EW)

</div>

Rank	Mean	SR	%MV	B/M	Liquidty	FF3 α	FF3 $\alpha(t)$	FF5 α	FF5 $\alpha(t)$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				PANEL A: SORTING $\mathcal{LR}(\omega_{ij,t-365:t})$					
1	1.46	0.12	16.69	44.13	1.02	−1.17	−0.39	−1.16	−0.38
2	1.70	0.16	21.14	40.35	0.87	−0.95	−0.37	−0.99	−0.38
3	3.93	0.47	22.66	39.62	0.85	1.45	0.57	1.38	0.55
4	3.69	0.43	21.88	39.54	0.86	1.24	0.47	1.12	0.43
5	3.40	0.38	17.64	43.05	0.99	1.20	0.42	1.07	0.39
5-1	1.94	0.30				2.37	1.00	2.23	0.95
				PANEL B: SORTING $\mathcal{LR}_{agg}(\omega_{ij,t-365:t})$					
1	2.22	0.23	15.06	42.24	0.90	−0.42	−0.16	−0.48	−0.19
2	2.05	0.21	20.37	40.65	0.89	−0.69	−0.26	−0.74	−0.28
3	2.44	0.27	23.45	39.86	0.88	−0.04	−0.02	−0.09	−0.04
4	3.30	0.37	22.24	40.42	0.90	0.98	0.37	0.90	0.34
5	3.95	0.43	18.88	43.56	1.02	1.72	0.60	1.62	0.57
5-1	1.73	0.35				2.14	1.27	2.10	1.25
Mkt	2.80	0.31				0.32	0.12	0.25	0.10

Both Panel A and Panel B report the daily return, r_{t+1}, performance metrics of the equal-weighted portfolios, and all portfolios are daily rebalanced. portfolios in Panel A are sorted by *lead returns*, $\mathcal{LR}(\omega_{ij,t-365:t})$, whereas portfolios in Panel B are sorted by $\mathcal{LR}_{agg}(\omega_{ij,t-365:t})$ described in Equation 11. News networks are constructed with 1-year rolling window news linkages. Column (1) reports the simple average return denoted in bps. Columns (2)-(5) are average annualized Sharpe-ratio, average market value proportion in percentage, the average book-to-market ratio in percentage, the average volume-to-share ratio in percentage. The last 4 columns report the alpha and robust Newey-West t-statistics in parentheses against Fama-French 3-factor and Fama-French 5-factor models, respectively.

weighted portfolio, where one can see clear monotonic decreasing pattern in terms of average residual return, Sharpe-ratio, which is a clear evidence for the negative impact from *lead return* on the 1-day ahead stock return. Also, the difference between returns in quintile 5 and quintile 1 has significant negative alpha −6.35 with t-statistics=−2.98 relative to FF3 and −6.43 with t-statistics=−3.01 relative to FF5. Consistent and more significant results is shown in Panel B of Table 6 reporting the FF5 residual return $r_{t+1}^{resid}(FF5)$.

Table 6
Table 6
Residual Portfolios by Sorting Preceding *lead return* (EW)

Rank	Mean	SR	%MV	B/M	Liquidty	FF3		FF5	
						α	$\alpha(t)$	α	$\alpha(t)$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
PANEL A: RESIDUAL RETURNS FROM FF3, $r_{t+1}^{resid}(FF3)$									
1	1.65	0.15	16.69	44.13	1.02	2.12	0.71	2.06	0.69
2	−0.53	−0.14	21.14	40.35	0.87	−0.01	0.00	−0.09	−0.03
3	−0.58	−0.15	22.66	39.62	0.85	0.04	0.02	−0.03	−0.01
4	−2.80	−0.47	21.88	39.54	0.86	−2.23	−0.85	−2.34	−0.90
5	−4.91	−0.74	17.64	43.05	0.99	−4.24	−1.52	−4.37	−1.58
5-1	−6.56	−1.49				−6.35	−2.98	−6.43	−3.01
PANEL B: RESIDUAL RETURNS FROM FF3, $r_{t+1}^{resid}(FF5)$									
1	2.37	0.25	16.69	44.13	1.02	2.84	0.96	2.79	0.94
2	−0.32	−0.11	21.14	40.35	0.87	0.18	0.07	0.11	0.04
3	−0.53	−0.14	22.66	39.62	0.85	0.05	0.02	−0.02	−0.01
4	−2.95	−0.49	21.88	39.54	0.86	−2.40	−0.92	−2.5	−0.96
5	−5.24	−0.79	17.64	43.05	0.99	−4.60	−1.67	−4.71	−1.72
5-1	−7.61	−1.75				−7.44	−3.54	−7.51	−3.57

Both Panel A and Panel B report the performance metrics of equal-weighted portfolios sorted by *lead returns*, $\mathcal{LR}(\omega_{ij,t-365:t})$, and all portfolios are daily rebalanced. Panel A reports the residual of daily return with respect to the Fama-French 3-factor model, i.e $r_{t+1}^{resid}(FF3)$, whereas Panel B reports the residual of daily return with respect to the Fama-French 5-factor model, i.e $r_{t+1}^{resid}(FF5)$. News networks are constructed with 1-year rolling window news linkages. Column (1) reports the simple average return denoted in bps. Columns (2)-(5) are average annualized Sharpe-ratio, average market value proportion in percentage, the average book-to-market ratio in percentage, the average volume-to-share ratio in percentage. The last 4 columns report the alpha and robust Newey-West t-statistics in parentheses against Fama-French 3-factor and Fama-French 5-factor models, respectively.

5. News *Network Degree* and Cross-Sectional Stock Returns

In this section, we first show that the network attention proxy, *network degree* $\mathcal{A}(\omega_{ij,T})$, defined in Section 2.2.2 can be a good predictor for the cross-sectional stock returns with a monthly rebalancing trading rule. Then we demonstrate that the predictability of *network degree* varies depending on the characteristics of lead firms, including firm size and stock liquidity. Finally, we conduct a series of robust checks to dissolve the suspicion on the predictability of *network degree*.

5.1. Hypothesis of News Network Attention

It is an interesting problem that how would a company perform if it had more attention from the news media. On the one hand, greater attention could be a positive signal for the popularity of a company. Consider one company going "hype" as a result of news reports covering it either in the headline (lead) or content (follower). This media attention could be caused by a new product release that many of its customers find desirable, or delivering certain innovation that could expand its the market share. Such that an investor would expect positive returns on that company's stock. On the other hand, bad news of a company also draws much attention from the media. For instance, a controversial market strategy made by a company causing accusations or even boycotts would definitely get this company covered by many news reports. In this case, more attention could be a signal of worsening performance in its stock.

The null hypothesis is that the attention proxy does not carry predictability on its stock performance. Under the null hypothesis, portfolios sorted by the attention proxy would not generate significantly different returns.

We conjecture that the media news are inclined to spend more attention on positive news instead of negative ones. In other words, a firm with higher attention from media would outperform those firms with lower attention from media. By employing *network degree* to capture the attention effect as described in Section 2.2.2, we perform portfolio test by sorting *network degree* variables.

We examine the cross-sectional predictability of monthly stock return with the method in Section 2.4. The network attention proxy, *network degree* $\mathcal{A}(\omega_{ij,t-30:t})$ is computed on 1-month aggregated news network allowing the strategy to be more feasible. All S&P500 constituents are sorted into quintiles based on *network degree* at the end of each month, and the quintile portfolios are rebalanced at the end of each month and held for one month. Within each of the quintile portfolios, stocks are equal-weighted, and the weights are rebalanced monthly. In Table 7, we report the equal-weighted portfolios.

The main point of the portfolio test is to see whether a long-short trading strategy based on sorting *network degree* can deliver abnormal returns. Under the efficient market hypothesis, one shall not see an abnormal return with this simple trading strategy. On the other hand, a positive abnormal return would suggest the existence anomaly returns with news media attention proxy, *network degree*, derived from networks of news.

29

5.2. Portfolios of Sorting Network Degree

Following the trading strategy described above, Table 7 reports the equal-weighted port-folios of sorting the preceding monthly *network degree*, $\mathcal{A}(\omega_{ij,t-30:t})$. The quintile portfolios are ranked in ascending order, where quintile 1 consists of assets with the lowest *network degree* value from the preceding month, and quintile 5 consists of assets with the highest network degree. As one can see the 1-month aggregated news network governed by the power-law distribution in Figure 6 of Section 2.2.2, many firm have network degrees equal to 0, which are not informative in our prediction. Therefore, we focus on the portfolio per-formance of stocks with network degrees larger than 0, i.e dropping out the firms with 0 network degrees when constructing portfolios. The consistent result of keeping all the firms is reported in Table 13 in Appendix.
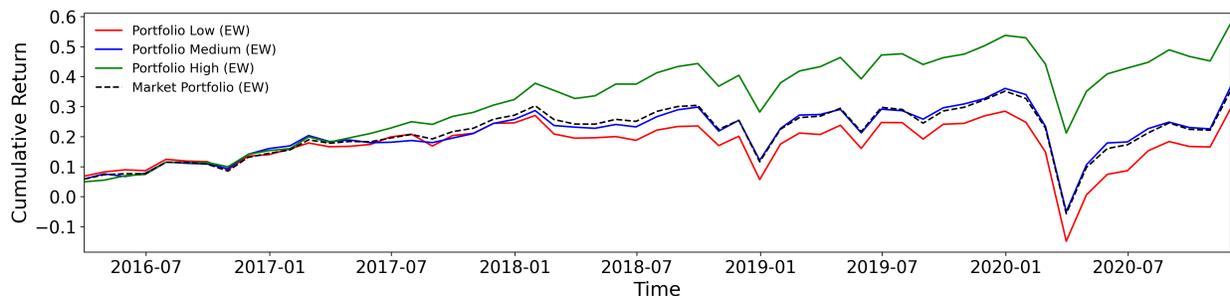
Figure 11: Portfolios Cumulative Returns of Sorting Preceding Monthly At-tention Proxy, *Network Degree* $\mathcal{A}(\omega_{ij,t-30:t})$
S&P500 constituents are sorted into quintile equal-weighted portfolios based on $\mathcal{A}(\omega_{ij,t-30:t})$ from the preceding month. All portfolios are rebalanced at the end of each month and held for one month, and firms with zero network degrees are removed. The three solid lines in red, blue, green refer to cumulative returns of portfolios with lowest, medium, and highest *network degree* value. The black dash line is the benchmark cumulative return of an equal-weighted S&P500 market portfolio.

We first demonstrate the cumulative returns of portfolios constructed by sorting the preceding monthly $\mathcal{A}(\omega_{ij,t-30:t})$ in Figure 11. The quintile 5 portfolio consists of stocks with highest network degree denoted in green line outperforms all the other portfolios significantly. The equal-weighted market portfolio in black dash line performs close to the quintile 3 portfolio in blue line. Finally, the quintile 1 portfolio in red line has the lowest performance over the sample period.

Column (1) in Table 7 reports the average simple return in percentage for each quintile portfolio across the 5-year sample period. Most importantly, quintile 5 generates 101 bps per month with a Sharpe-ratio 0.58, outperforming the other quintile portfolios and the equal-weighted market portfolio (64 bps and with Sharpe-ratio 0.31). One can also see that

30

Table 7
Portfolios by Sorting Preceding *Network Degree* $\mathcal{A}$
(Equal-Weighted, Drop $\mathcal{A}_{i,t} = 0$)

Rank	Mean	SR	%MV	B/M	Liquidty	FF3		FF5	
						α	R^2	α	R^2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	0.51	0.23	8.38	42.86	0.89	−0.75	91.17	−0.72	91.50
						(−2.55)		(−2.50)	
2	0.41	0.17	10.21	42.29	0.89	−0.87	92.89	−0.83	93.79
						(−3.75)		(−3.86)	
3	0.64	0.31	13.15	42.35	0.91	−0.64	90.96	−0.61	91.52
						(−2.75)		(−2.70)	
4	0.51	0.22	19.28	38.90	0.93	−0.81	87.79	−0.76	89.09
						(−2.81)		(−2.69)	
5	1.01	0.58	48.98	36.82	1.00	−0.27	93.03	−0.25	93.80
						(−1.48)		(−1.39)	
5-1	0.50	0.79				0.48	37.51	0.47	38.89
						(2.50)		(2.51)	
Mkt	0.64	0.31				−0.69	92.56	−0.65	93.34
						(−3.03)		(−2.97)	

The table reports the performance of the equal-weighted portfolios sorted by the network attention proxy, *network degree* $\mathcal{A}(\omega_{ij,t-30:t})$, computed with news linkages from the preceding month. Firms with zero network degrees are removed before forming the portfolios. Columns (1)-(5) are average simple daily return in percentage; the average annualized sharpe-ratio, average market value proportion in percentage, the average book-to-market ratio in percentage, and the average volume-to-share ratio in percentage. The last 4 columns report the alpha, R^2, and robust Newey-West t-statistics in parentheses with respect to the Fama-French 3-factor and the Fama-French 5-factor models, respectively.

the simple return is almost monotonic increasing, except for quintile 4, which generates a 51 bps per month, just slightly higher than quintile 1 with 48 bps.

As reported in the row with index "*5-1*" of Table 7, the zero-cost long-short portfolio generates a positive return of 50 bps per month and a high Sharpe-ratio of 0.79. Moreover, the long-short portfolio shows a significant positive alpha of 48 bps (t-statistics=2.07) relative to FF3 and 47 bps (t-statistics=1.97) with respect to FF5. A similar result of value-weighting portfolios is reported in Table 14 in Appendix.

The positive return and alphas from the simple zero-cost long-short portfolio provide evidence for the predictability of the attention proxy, *network degree*, on stock return, and it is feasible for investors to form portfolios. Additionally, quintile 5 can be viewed as an

enhanced market portfolio that consists of around 100 stocks from S&P 500 constituents, for its much higher average return and Sharpe-ratio than the market index.

5.3.   What Are The Important News Linkages?

Are all the detected linkages equally important, or are certain types of linkages more crucial for predicting the cross-sectional stock returns? In this subsection, we reveal that *network degree* variables of different types of linkages carry different levels of predictability by studying network attention variants detailed in Section 2.2.2.

5.3.1.   Lead and Follower Attention

First of all, one might wonder whether the news attention of being lead firm is more important than being a follower firm for stock returns prediction. We analyze the predictability of the *lead attention* and *follower attention* defined in Section 2.2.2.

Panel A of Table 8 reports the portfolio test results on the two network attention variants, where Panel A(1) shows the portfolios of sorting $\mathcal{A}^{lead}(\omega_{ij,t-30:t})$, and Panel A(2) contains the $\mathcal{A}^{follower}(\omega_{ij,t-30:t})$ sorted results. The zero-cost long-shot trading strategy, indexed with "*5-1*" in Panel A, of both sorting cases show similar significant abnormal returns with respect to FF3 and FF5. However, the performance of the quintile portfolio is not monotonically increasing in both of the sorting cases. In Panel A(1), we can see that quintile 5 outperforms any other quintile portfolios, resulting in the long-short strategy producing a significant alpha with respect to FF3 (51 bps, *t*-statistics=2.92) and FF5 (51 bps, *t*-statistics=3.01), as shown in Column (6)-(9). A similar result for the *follower attention* sorting case in Panel A(2) reveals that quintile 5 stands out despite quintile 1-4 having similar performance. As such, the long-short strategy generates significant alpha with respect to FF3 (39 bps, *t*-statistics=2.91) and FF5 (37 bps, *t*-statistics=2.82). *Lead attention* is relatively more important for the cross-sectional stock return prediction, but *follower attention* also contributes to the predictability of using *network degree*.

5.3.2.   Within-Sector and Cross-Sector Attention

It is a stylized fact that firms from the same sector tend to have a closer relationship than those from different sectors. As shown in Section 4.1, comovement effect is more significant between the stocks from the same sector. We are motivated to understand whether the firms from the same sector are more informative in terms of stock return prediction. Base on the sector decomposed attention variants, we further test the predictability of *within-sector attention* $\mathcal{A}(\omega_{ij,t-30:t}^{w})$, and *cross-sector attention* $\mathcal{A}(\omega_{ij,t-30:t}^{c})$.

As reported in Panel B of Table 8, we can see that *within-sector attention* does provide stronger predictability compared with *cross-sector attention* in terms of significant positive abnormal return from the long-short strategy. Specifically, the long-short portfolio by sorting $\mathcal{A}(\omega_{ij,t-30:t}^w)$ produces an alpha of 59 bps (*t*-statistics=3.14) with respect to FF3 and 57 bps (*t*-statistics=3.26) relative to FF5. And the long-short portfolio sorted by $\mathcal{A}(\omega_{ij,t-30:t}^c)$ has positive alphas with respect to FF3 and FF5 (27 bps, 26 bps) but insignificant (*t*-statistics=1.35, *t*-statistics=1.31). Moreover, a clear monotonic increase pattern on the quintile portfolios performance is found with sorting $\mathcal{A}(\omega_{ij,t-30:t}^w)$, see Columns (1)-(2) in Panel B(1). While assets sorted by $\mathcal{A}(\omega_{ij,t-30:t}^c)$ do not have a monotonic increase performance pattern.

From the statistical point of view, one can conclude that *within-sector attention* provides better and more stable cross-sectional predictability. However, it does not mean that one can always have higher performance with *within-sector attention*. As one can see that quintile 5 sorted by $\mathcal{A}(\omega_{ij,t-30:t}^w)$ in Panel B(1) with 89 bps (Sharpe-ratio=0.49) underperforms quintile 5 sorted by $\mathcal{A}(\omega_{ij,t-30:t}^c)$ in Panel B(2), which has 97 bps (Sharpe-ratio=0.52).

5.3.3. *Big Size and Small Size Lead Firm Attention*

It is commonly believed that media are more likely to cover the news of larger firms. Here we would like to see how important it is to appear in the news content focusing on large firms compared to the news revolving around the small firms. The *big lead attention* $\left[\omega_{ij,t-30:t}^{BigLead}\right]$ and *small lead attention* $\left[\omega_{ij,t-30:t}^{SmallLead}\right]$ defined in Section 2.2.2 are employed for the comparison.

Panel C in Table 8 reports the performance of the portfolios sorted by the two attention proxy variants, where we can see *big lead attention* shows significant predictability by providing positive abnormal returns from the long-short strategy. In contrast, *small lead attention* gives mixed results and negative insignificant abnormal returns from the long-short strategy. As reported in Column (1)-(2) of Panel C(1), quintile portfolio performance, mean return and Sharpe-ratio, increase monotonic along with the increase of $\mathcal{A}(\omega_{ij,t-30:t}^{BigLead})$, and Column (6)-(9) report that the long-short strategy indexed with "*5-1*" provides an alpha of 68 bps (*t*-statistics=2.65) with respect to FF3 and also an alpha of 64 bps relative to FF5 (*t*-statistics=2.62). From the long-short strategy result of sorting $\mathcal{A}(\omega_{ij,t-30:t}^{SmallLead})$ reported in Column (6)-(9) of Panel C(2), it reveals that a firm with higher $\mathcal{A}(\omega_{ij,t-30:t}^{SmallLead})$ is likely to underperform, which is opposite to $\mathcal{A}(\omega_{ij,t-30:t}^{BigLead})$, but nevertheless insignificant. Note that the smaller firms have fewer news coverages, and we drop the stocks that have attention variables equal to zero. Hence, quintile portfolios sorted by $\mathcal{A}(\omega_{ij,t-30:t}^{SmallLead})$ have fewer stocks, averaging 38 stocks for each quintile portfolio in each month.

The above analysis shows that the news linkages formed with the big market capital lead firms can be more informative in terms of providing the cross-sectional stock return predictability. Specifically, the more linkages a firm connecting to big firms as a follower, the more likely it will have better performance in the next month. On the other side, the news linkages formed with smaller market capital lead firms give opposite implications, however insignificant.

<div align="center">

Table 8
Portfolio Tests on Linkage Types
(Equal-Weighted, Drop $\mathcal{A}_{i,t} = 0$)

</div>

						FF3		FF5	
Rank	Mean	SR	%MV	B/M	Liquidty	α	$\alpha(t)$	α	$\alpha(t)$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)

PANEL A: LEAD ATTENTION V.S FOLLOWER ATTENTION

PANEL A(1): SORTED BY $\mathcal{A}^{lead}$

1	0.32	0.12	10.52	41.60	0.90	-0.92	-3.63	-0.90	-3.57
2	0.78	0.37	12.57	41.11	0.88	-0.51	-1.62	-0.47	-1.60
3	0.59	0.27	13.82	42.98	0.89	-0.63	-2.38	-0.57	-2.29
4	0.51	0.24	19.57	39.59	0.94	-0.81	-3.27	-0.78	-3.16
5	0.91	0.50	43.52	36.91	1.08	-0.41	-1.96	-0.39	-1.79
5-1	0.60	1.07				0.51	2.92	0.51	3.01

PANEL A(2): SORTED BY $\mathcal{A}^{follower}$

1	0.63	0.29	8.26	42.84	0.93	-0.74	-3.06	-0.70	-3.03
2	0.51	0.21	9.96	41.24	0.91	-0.8	-2.92	-0.76	-2.87
3	0.6	0.26	13.33	41.79	0.95	-0.66	-2.38	-0.62	-2.28
4	0.53	0.26	20.90	38.88	0.96	-0.73	-3.69	-0.70	-3.47
5	0.97	0.54	47.56	36.84	0.90	-0.35	-1.73	-0.33	-1.70
5-1	0.34	0.61				0.39	2.91	0.37	2.82

PANEL B: WITHIN-SECTOR ATTENTION V.S CROSS-SECTOR ATTENTION

PANEL B(1): SORTED BY $\mathcal{A}(\omega_{ij,t-30:t}^{w})$

1	0.32	0.13	9.24	39.85	0.91	-0.98	-3.75	-0.95	-3.74
2	0.54	0.24	11.14	41.59	0.90	-0.71	-2.34	-0.67	-2.27
3	0.55	0.25	14.63	41.68	0.90	-0.75	-3.00	-0.72	-2.97

continued

Table 8
continued

Rank	Mean	SR	%MV	B/M	Liquidty	FF3		FF5	
						α	$\alpha(t)$	α	$\alpha(t)$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
4	0.61	0.29	22.00	40.63	0.91	−0.68	−3.12	−0.64	−2.93
5	0.89	0.49	42.99	39.96	0.99	−0.39	−1.93	−0.38	−1.92
5-1	0.56	1.02				0.59	3.14	0.57	3.26

PANEL B(2): SORTED BY $\mathcal{A}(\omega_{ij,t-30:t}^{c})$

Rank	Mean	SR	%MV	B/M	Liquidty	α	$\alpha(t)$	α	$\alpha(t)$
1	0.50	0.22	9.73	44.56	0.89	−0.66	−2.48	−0.63	−2.48
2	0.66	0.29	11.54	41.72	0.91	−0.65	−2.53	−0.63	−2.52
3	0.63	0.31	13.91	39.39	0.93	−0.68	−3.06	−0.65	−2.96
4	0.35	0.14	19.65	34.79	0.96	−0.93	−2.82	−0.87	−2.77
5	0.97	0.52	45.17	36.45	1.04	−0.38	−1.89	−0.36	−1.70
5-1	0.47	0.73				0.27	1.35	0.26	1.31

PANEL C: LEAD FIRM SIZE ATTENTION

PANEL C(1): SORTED BY $\mathcal{A}(\omega_{ij,t-30:t}^{BigLead})$

Rank	Mean	SR	%MV	B/M	Liquidty	α	$\alpha(t)$	α	$\alpha(t)$
1	0.41	0.15	6.45	47.06	1.04	−0.89	−2.98	−0.85	−2.91
2	0.54	0.21	8.06	43.72	1.03	−0.85	−2.48	−0.80	−2.43
3	0.74	0.41	12.83	38.69	0.96	−0.52	−2.62	−0.51	−2.57
4	0.83	0.51	22.95	33.09	0.72	−0.44	−2.50	−0.43	−2.44
5	1.05	0.67	49.71	34.58	0.75	−0.21	−1.95	−0.22	−1.79
5-1	0.64	0.59				0.68	2.65	0.64	2.62

PANEL C(2): SORTED BY $\mathcal{A}(\omega_{ij,t-30:t}^{SmallLead})$

Rank	Mean	SR	%MV	B/M	Liquidty	α	$\alpha(t)$	α	$\alpha(t)$
1	0.54	0.25	19.91	37.66	0.87	−0.77	−2.84	−0.71	−2.76
2	0.78	0.42	20.04	39.26	0.89	−0.48	−2.29	−0.47	−2.31
3	0.29	0.11	20.46	43.65	0.95	−0.97	−2.81	−0.94	−2.72
4	0.54	0.20	21.21	47.73	1.08	−0.80	−1.82	−0.71	−1.70
5	−0.03	−0.05	18.38	55.66	1.59	−1.40	−2.40	−1.30	−2.31
5-1	−0.57	−0.52				−0.62	−1.42	−0.59	−1.35

PANEL D: LEAD FIRM LIQUIDITY ATTENTION

PANEL D(1): SORTED BY $\mathcal{A}(\omega_{ij,t-30:t}^{HighLead})$

Rank	Mean	SR	%MV	B/M	Liquidty	α	$\alpha(t)$	α	$\alpha(t)$
1	0.54	0.28	15.87	38.22	0.80	−0.66	−2.73	−0.63	−2.59

Table 8
continued

Rank	Mean	SR	%MV	B/M	Liquidty	FF3 α	FF3 $\alpha(t)$	FF5 α	FF5 $\alpha(t)$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
2	0.69	0.36	17.17	37.97	0.80	-0.56	-2.81	-0.56	-2.79
3	0.57	0.24	18.62	42.15	0.95	-0.92	-3.57	-0.89	-3.54
4	0.09	0.00	19.32	46.18	1.20	-1.23	-3.25	-1.16	-3.20
5	0.33	0.09	29.01	47.47	1.71	-1.17	-2.38	-1.07	-2.30
5-1	-0.21	-0.23				-0.51	-1.28	-0.44	-1.16
PANEL D(2): SORTED BY $\mathcal{A}(\omega_{ij,t-30:t}^{LowLead})$									
1	0.32	0.12	8.07	43.42	1.04	-0.96	-4.43	-0.93	-4.51
2	0.95	0.48	9.65	41.60	1.01	-0.30	-1.22	-0.28	-1.18
3	0.51	0.24	13.41	38.52	0.84	-0.80	-2.87	-0.76	-2.78
4	1.09	0.75	21.44	34.70	0.67	-0.05	-0.20	-0.03	-0.14
5	1.05	0.71	47.44	34.37	0.59	-0.11	-0.73	-0.11	-0.75
5-1	0.73	0.95				0.85	4.90	0.81	4.99

This table reports the portfolio performance of sorting network attention proxies, *network degree* $\mathcal{A}$, computed with different types of linkages. All portfolios are equal-weighted, and rebalanced monthly. We remove the firms that have the sorting variable equal to 0. In each panel, Columns (1)-(5) are the average simple daily return in percentage; average annualized sharpe-ratio, average market value proportion in percentage, average book-to-market ratio in percentage, average volume-to-share ratio in percentage. The last 4 columns report the alpha and robust Newey-West t-statistics with respect to the Fama-French 3-factor and the Fama-French 5-factor models, respectively. Panel A reports the differences in portfolio performance between the *lead attention* $\mathcal{A}^{lead}(\omega_{ij,t-30:t})$ (Panel A1) and the *follower attention* $\mathcal{A}^{follower}(\omega_{ij,t-30:t})$ (Panel A2). Panel B reports the differences in portfolio performance between the *within-sector attention* $\mathcal{A}(\omega_{ij,t-30:t}^{w})$ (Panel B1) and the *cross-sector attention* $\mathcal{A}(\omega_{ij,t-30:t}^{c})$ (Panel B2). Panel C reports the differences in portfolio performance between the *big lead attention* $\mathcal{A}(\omega_{ij,t-30:t}^{BigLead})$ (Panel C1) and the *small lead attention* $\mathcal{A}(\omega_{ij,t-30:t}^{SmallLead})$ (Panel C2). Panel D reports the differences in portfolio performance between the *high lead attention* $\mathcal{A}(\omega_{ij,t-30:t}^{HighLead})$ (Panel D1) and the *low lead attention* $\mathcal{A}(\omega_{ij,t-30:t}^{LowLead})$ (Panel D2). All the attention variants are defined in Section 2.2.2.

5.3.4. *More Liquid and Less Liquid Lead Firm Attention*

Stock liquidity is often investigated because firms with higher stock liquidity are deemed to incorporate news information faster into their stock prices. In comparison, less liquid stocks could carry material information ignored by many investors, then are informative for prediction. Similar to the analysis on the firm size of the lead firm, we further test how the

linkages formed with high and low liquidity lead firms carry the cross-sectional predictability differently. The *high lead attention* $\left[\omega_{ij,t-30:t}^{HighLead}\right]$ and *low lead attention* $\left[\omega_{ij,t-30:t}^{LowLead}\right]$ defined in Section 2.2.2 are investigated.

Panel D of Table 8 reports the stocks sorting results, where one can see that the portfolio test result is mixed. On the one hand, as shown in Panel D(2), one can see the long-short strategy ("*5-1*") by sorting $\mathcal{A}(\omega_{ij,t-30:t}^{LowLead})$ generates significant positive alphas with respect to FF3 (85 bps (*t*-statistics=4.90)) and FF5 (81 bps (*t*-statistics=4.99)). And the long-short strategy ("*5-1*") by sorting $\mathcal{A}(\omega_{ij,t-30:t}^{HighLead})$ reported in Panel D(1) shows insignificant negative alphas with respect to FF3 and FF5, implying that news linkages of low stock liquidity lead firms carry higher predictability of monthly stock returns. On the other hand, neither of those two attention variants can sort portfolios with performance monotonic increase.

5.4.   *Robust Check on Network Degree Predictability*

The portfolio test in Section 5.2 showcases how the network attention proxy, *network degree* can bring predictability on the cross-sectional stock returns. However, it is still a question if the predictability comes from some of the known characteristics that are also captured by the news *network degree*. As reported in Columns (3)-(5) of Table 7, where all the three characteristics, %MV, B/M, and Liquidity, change monotonically, implying a strong correlation between those three characteristics and the portfolio ranking. For example, quintile 5 contains more large firms than quintile 1 and has higher liquidity. Hence, it is imperative to test the robustness of *network degree* predictability to size, value, liquidity, momentum, and window size choice. We first employ the double-sorting method described in Section 2.4.1 to relieve the results from the first four possible effects mentioned above. After that, we check the robustness to the network formation window size. Note that the financial statement data, including the market share and book equity B, are two quarters lagged values when they are used to compute the sorting characteristics.

5.4.1.   *Controlling for Firm Size*

News medias pay more attention on the firms with larger market capitalization, suggesting that *network degree* is correlated with the firm market value at a certain level even though *network degree* exclude most of the repetitive reports by counting the number of linkages. In the sample period, within the S&P500 constituents, the larger firms outperform the smaller ones, so it is reasonable to question that the predictability of news *network degree* is related to the firm size. We relieve the size effect by employing the double-sorting method on firm market capital.

Table 9
Table 9
Robustness Check on Size, Value, Liquidity, and Momentum

Rank	*1*	*2*	*3*	*4*	*5*		*5-1*	
							α-FF3	α-FF5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			PANEL A: CONTROLLING FOR SIZE					
Mean	0.42	0.52	0.65	0.67	0.76	0.34	0.28	0.30
SR	0.19	0.23	0.32	0.32	0.36	0.62	(1.84)	(2.07)
			PANEL B: CONTROLLING FOR VALUE					
Mean	0.49	0.54	0.39	0.62	0.93	0.44	0.55	0.54
SR	0.21	0.25	0.16	0.28	0.52	0.73	(3.68)	(3.89)
			PANEL C: CONTROLLING FOR LIQUIDITY					
Mean	0.51	0.44	0.67	0.52	0.84	0.33	0.36	0.35
SR	0.22	0.18	0.32	0.23	0.46	0.51	(1.88)	(1.86)
			PANEL D: CONTROLLING FOR MOMENTUM					
Mean	0.45	0.54	0.54	0.56	0.94	0.49	0.48	0.47
SR	0.20	0.25	0.24	0.25	0.52	0.88	(2.96)	(3.00)

The table reports the portfolio performance of sorting the network attention proxy, *network degree* $\mathcal{A}(\omega_{ij,t-30:t})$, after controlling for the firm size (Panel A), firm value (Panel B), firm liquidity (Panel C), and momentum effect (Panel D). The double-sorting method is employed to control for those characteristics. Columns (1)-(5) in each panel are the average simple return and Sharpe-ratio of each quintile portfolio. Column (6) in each panel is the average simple return and Sharp-ratio of the long-short portfolio, i.e., the differences in returns between quintile 5 and quintile 1. Columns (7) and (8) in each panel are the alphas and robust Newey–West t-statistics with respect to the Fama-French 3-factor and 5-factor models, respectively.

As reported in Panel A of Table 9, after controlling for firm size, average return of the long-short strategy is reduced from 50 bps in Table 7 to 34 bps, nevertheless *network degree* still provides relatively strong predictability from two aspects. First, the long-short strategy *5-1* generates positive alphas with respect to FF3 (28 bps) and FF5 (30 bps), although alpha with respect to FF3 is only significant at 10% level (t-statistics=1.84), but significant in FF5 at 5% level (t-statistics=2.07) as reported in the last two columns. The second aspect is that there is a monotonic increase pattern of portfolio performance in terms of average return and Sharpe-ratio as shown in the first five columns of Table 9.

5.4.2. Controlling for Firm Value

Much of the literature documents that small growth firms outperform those large firms. However, this phenomenon is not observed by sorting *network degree* in our sample period as reported in Column (4) of Table 7, where quintile 5 with the best performance has the lowest average book-to-market value. Hence, a double-sorting is employed on book-to-value to test the robustness to firm value.

The firm value controlled portfolio result is reported in Panel B of Table 9, where average return of the long-short strategy is reduced to 44 bps compared with the non-controlling result of 50 bps. But the *network degree* sorted long-short portfolio remains positive alphas of 55 bps and 54 bps with respect to FF3 and FF5 at 1% level (with *t*-statistics=3.68 and 3.89) as reported in the last two columns in Panel B of Table 7.

5.4.3. Controlling for Liquidity

As shown in Column (5) of Table 7, the liquidity proxy, volume/share, increases along with the portfolio rank. Such a pattern is consistent with the finding in Gervais, Kaniel, and Mingelgrin (2001) that stocks tend to outperform when they experienced higher trading volume in the preceding month.

We control the liquidity effect with the double-sorting method applied to the turnover rate. As reported in Panel C of Table 9, which suggests that *network degree* does capture quite amount of liquidity effect. An average return of the long-short strategy decreases to 33 bps from 50 bps in a non-controlled case. The long-short strategy still generates positive alphas with respect to FF3 (36 bps) and FF5 (35 bps), but only significant at 10% level with *t*-statistics=1.88 and 1.86. Although abnormal returns from the long-short strategy are less significant compared with cases controlling other effects, one can still conclude that *network degree* provides extra information in the cross-sectional return predictability.

5.4.4. Controlling for Momentum

It was well documented that the stocks that experienced higher returns in the past will tend to have higher returns, which is called momentum effect (Jegadeesh and Titman (1993)). One would suspect that *network degree* correlates with the last monthly stock performance, in other words, the predictability of *network degree* can be explained by momentum effect. Therefore, we conduct the robustness check by applying the double-sorting method on monthly return in the preceding month.

The long-short strategy produces an average return of 49 bps, which is close to the average return of 50 bps in the non-controlled case. The FF-3 and FF-5 alphas remain positive

and significant at 1% level with *t*-statistics=2.96 and 3.00. Also, portfolio performance increases monotonically from quintile 1 to 5 by sorting *network degree*. It's evidenced that the momentum effect can not account for the cross-sectional stock return predictability from *network degree*.

5.4.5. Robustness for Network Formation Window Size

Note that the choice of network formation window size is not an outcome of pretests. We choose a 1-month window size to form the news network simply because the portfolio is rebalanced monthly, which is a convention in much of the literature that tests the cross-sectional predictability. The result is quite robust when we extend the 1-month window size 1-week and 2-week more. The long-short strategy with 1-more week window size still generates significant alphas of 47 bps and 45 bps relative to FF3 and FF5 with *t*-statistics=2.39 and 2.37, compared with alphas of 48 (*t*-statistics=2.50) bps and 47 *t*-statistics=2.51 bps in 1-month window size case reported in Table 7. As expected, with longer window size, the cross-sectional stock return predictability of *network degree* decay. When the network formation window size is extended with 2 more weeks, the long-short strategy produces 36 bps (*t*-statistics=1.93) with respect to FF3 and 35 bps (*t*-statistics=1.86) with respect to FF5. Overall, the result is robust to the selection of network formation window size.

6. Conclusion

This paper studies the news-based interdependency network between firms joint with cross-sectional returns and their consequences. By postulating the structure of firm-specific news, we define the lead-follower linkage embedded in the news. After that, we propose an algorithm to tackle the type-I error in identifying firm tickers from the news text so that we can construct the directed networks of S&P500 firms from news articles.

By investigating the stock returns of lead firms, *lead return*, and follower firms, we find that firms linked by news articles, regardless of the business relationship between lead and follower firms, tend to have strong comovement effects on daily returns, after controlling for factors including idiosyncratic alphas, market premium, size factor, profitability factor, etc. Moreover, the econometric model reveals that *lead return* has a reversal effect on 1-day ahead stock return. However, in the cross-sectional portfolio test, the reversal effect is not strong enough to dominate the effects caused by other factors. Therefore, *lead return* does not carry significant cross-sectional predictability.

We find that the monthly aggregated *network degree*, capturing the attention of the news network, is a good predictor for the cross-sectional stock returns. A monthly rebalancing

portfolio test reveals that *network degree* provides significant positive alphas relative to the Fama-French 3-factor and 5-factor models. Among the different types of linkages, we show that *network degree* of (1) both lead and follower networks, (2) within-sector network, (4) large size lead firm network, (5) and lower liquidity lead firm network, generate more significant positive alphas. Finally, we show the robustness of the cross-sectional predictability of *network degree* against some of the known factors and window size of network formation.

Our results provide two insights for portfolio management. First, investors can optimize their portfolios by accounting for the comovement effects between their assets. Second, the portfolio test on *network degree* can be easily transformed into trading strategies, either a smart-beta index based on the high *network degree* portfolios or an alpha investment strategy based on the long-short portfolios.

References

ACEMOGLU, D., V. M. CARVALHO, A. OZDAGLAR, AND A. TAHBAZ-SALEHI (2012): "The Network Origins Of Aggregate Fluctuations," *Econometrica*, 80(5), 1977–2016.

AHERN, K. R., AND J. HARFORD (2014): "The Importance Of Industry Links In Merger Waves," *The Journal of Finance*, 69(2), 527–576.

ALBERT, R., H. JEONG, AND A.-L. BARABÁSI (1999): "Diameter of the World-Wide Web," *Nature*, 401(6749), 130–131.

ANG, A., R. J. HODRICK, Y. XING, AND X. ZHANG (2006): "The Cross-section Of Volatility And Expected Returns," *The Journal of Finance*, 61(1), 259–299.

ANTON, M., AND C. POLK (2014): "Connected Stocks," *The Journal of Finance*, 69(3), 1099–1127.

ANTWEILER, W., AND M. Z. FRANK (2004): "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *The Journal of Finance*, 59(3), 1259–1294.

BARABÁSI, A.-L., AND R. ALBERT (1999): "Emergence of scaling in random networks," *science*, 286(5439), 509–512.

BARABSI, A.-L. (2016): *Network Science*. Cambridge University Press, USA, 1st edn.

BARIGOZZI, M., AND C. BROWNLEES (2019): "Nets: Network estimation for time series," *Journal of Applied Econometrics*, 34(3), 347–364.

BOUDOUKH, J., R. FELDMAN, S. KOGAN, AND M. RICHARDSON (2018): "Information, Trading, and Volatility: Evidence from Firm-Specific News," *The Review of Financial Studies*, 32(3), 992–1033.

CHEN, Z., L. GUO, AND J. TU (2021): "Media Connection and Return Comovement," *Journal of Economic Dynamics and Control*, p. 104191.

COHEN, L., AND A. FRAZZINI (2008): "Economic Links And Predictable Returns," *The Journal of Finance*, 63(4), 1977–2011.

DA, Z., J. ENGELBERG, AND P. GAO (2014): "The Sum of All FEARS Investor Sentiment and Asset Prices," *The Review of Financial Studies*, 28(1), 1–32.

DA, Z., U. G. GURUN, B. LI, AND M. WARACHKA (2021): "Investment in a Smaller World: The Implications of Air Travel for Investors and Firms," *Management Science*, 67(1), 417–435.

DIEBOLD, F. X., AND K. YILMAZ (2008): "Measuring Financial Asset Return and Volatility Spillovers, with Application to Global Equity Markets," *The Economic Journal*, 119(534), 158–171.

DIEBOLD, F. X., AND K. YILMAZ (2014): "On The Network Topology Of Variance Decompositions: Measuring The Connectedness Of Financial Firms," *Journal of Econometrics*, 182(1), 119–134, Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions.

FAMA, E. F., AND K. R. FRENCH (1993): "Common Risk Factors In The Returns On Stocks And Bonds," *Journal of Financial Economics*, 33(1), 3–56.

——— (2015): "A Five-factor Asset Pricing Model," *Journal of Financial Economics*, 116(1), 1–22.

FAMA, E. F., AND J. D. MACBETH (1973): "Risk, Return, And Equilibrium: Empirical Tests," *Journal of Political Economy*, 81(3), 607–636.

GABAIX, X. (2011): "The Granular Origins Of Aggregate Fluctuations," *Econometrica*, 79(3), 733–772.

GE, S., O. LINTON, AND S. LI (2021): "News-Implied Linkages and Local Dependency in the Equity Market," .

GERVAIS, S., R. KANIEL, AND D. H. MINGELGRIN (2001): "The High-volume Return Premium," *The Journal of Finance*, 56(3), 877–919.

GUO, L., AND F. W. LI (2019): "Media-based Inter-industry Network And Information Transmission," *ArXiv*.

HÄRDLE, W. K., W. WANG, AND L. YU (2016): "TENET: Tail-event Driven NETwork Risk," *Journal of Econometrics*, 192(2), 499–513.

HAUTSCH, N., J. SCHAUMBURG, AND M. SCHIENLE (2014): "Financial Network Systemic Risk Contributions*," *Review of Finance*, 19(2), 685–738.

HERSKOVIC, B. (2018): "Networks In Production: Asset Pricing Implications," *The Journal of Finance*, 73(4), 1785–1818.

43

HERSKOVIC, B., B. KELLY, H. LUSTIG, AND S. VAN NIEUWERBURGH (2020): "Firm Volatility In Granular Networks," *Journal of Political Economy*, 128(11), 4097–4162.

JACKSON, M. O. (2010): *Social And Economic Networks*. Princeton university press.

JEGADEESH, N., AND S. TITMAN (1993): "Returns To Buying Winners And Selling Losers: Implications For Stock Market Efficiency," *The Journal of Finance*, 48(1), 65–91.

SCHERBINA, A., AND B. SCHLUSCHE (2015): "Economic Linkages Inferred From News Stories And The Predictability Of Stock Returns," *Available at SSRN 2363436*.

SCHWENKLER, G., AND H. ZHENG (2019): "The Network Of Firms Implied By The News," *Boston University Questrom School of Business Research Paper*, (3320859).

YANG, Z., AND Y. ZHOU (2017): "Quantitative Easing And Volatility Spillovers Across Countries And Asset Classes," *Management Science*, 63(2), 333–354.

ZHANG, J. L., W. K. HÄRDLE, C. Y. CHEN, AND E. BOMMES (2016): "Distillation of News Flow Into Analysis of Stock Reactions," *Journal of Business & Economic Statistics*, 34(4), 547–563.

ZHU, X., W. WANG, H. WANG, AND W. K. HÄRDLE (2019): "Network Quantile Autoregression," *Journal of Econometrics*, 212(1), 345–358.

Appendix A.

Algorithm 1: Algorithm to Identify Firms

    Input   : A list of articles $[article_i]$, $i = 0, 1, 2, \cdots, N\text{-}1$;
    Output: A list of lead-follower pairs $resList = [pair_j]$, $i = 0, 1, 2, \cdots, M\text{-}1$;

1 **Function** `pairsExtractor`($article_i$):
      Input   : A list of article entry $article_i = [titleStr, contentStr, timeStr]$

2   Initialization $pairsList = [\;]$;

3   `/* One might need to clean the text e.g special characters.`   `*/`

4   `/* Refer to the code online for more details.`   `*/`

5   `cleanText`($titleStr$); `cleanText`($contentStr$);

6 $lead = $ `S1`($titleStr$); `// S1 Strategy`

7   **if** `length`($lead$) $== 0$ **then**

8 $lead = $ `S2`($titleStr$); `// S2 Strategy`

9   **if** `length`($lead == 0$) **then**

10 $lead = $ `S3`($titleStr$); `// S3 Strategy`

11   **end if**

12   **end if**

13   **if** `length`($lead$)> 1 **then**

14 $lead = [\;]$; `// We only identify single-lead articles`

15   **end if**

16 $followersList = $ `S1`($contentStr$) $+$ `S2`($contentStr$) `// S1, S2 Strategy`

17   **if** `length`($lead$)$!=0$ & `length`($followersList$)$!=0$ **then**

18   **for** $i = 0$ **to** `length`($followersList$) **do**

19 $pairsList$.`append`(($lead, followersList[i], timeStr$))

20   **end for**

21   **end if**

22   **return** $pairsList$

23 **End Function**

24 **Parallel Processing** $article_i$, $i = 0$ **to** $N\text{-}1$

25 $resList$.`extend`(`pairsExtractor`($article_i$));

26 **End Parallel**

27 **return** $newList$

Algorithm 2: Mapping Firm Names to n-gram Name Segments

Input : A list of full names of all the firms $[name_i]$, $i = 0, 1, 2, \cdots, n$-1;
A list of suffix and redundant strings $[s_m]$, $m = 0, 1, 2, \cdots, M$-1;
Threshold values η_1 and η_2
Output: A list of lists of the segments for each firm: $segmentsList = [[segList_i]]$,
        $i = 0, 1, 2, \cdots, n$-1

1 Initialization $segmentsList = [\ ]$;
2 `// Construct the n-gram name segments for each firm`
3 **for** $i = 0$ **to** *n-1* **do**
4   Remove special characters from $name_i$;
5   `// split` $name_i$ `into` K_i`-1 segments by space`
6 $segList_i = [seg_k]$, $k = 0, 1, 2 \cdots, K_i$-1 ;
7   **for** $k = 1$ **to** K_i*-1* **do**
8   `// replace each` seg_k `with n-gram string`
9 $seg_k \leftarrow segList_i[0, k].toString$;
10   **end for**
11 $segmentsList.append(segList_i)$
12 **end for**
13 `// Remove the common segments for different firms`
14 **for** $i = 0$ **to** $n - 2$ **do**
15   **for** $j = i$ **to** *n-1* **do**
16 $interList_{i,j} = $ `intersect`$(segList_i, segList_j)$;
17   **for** $p = 0$ **to** `length`$(interList_{i,j})$*-1* **do**
18 $segList_i \leftarrow segList_i.remove(interList[p])$;
19 $segList_j \leftarrow segList_j.remove(interList[p])$;
20   **end for**
21   **end for**
22 **end for**
23 `// Further reduce the complexity`
24 **for** $i = 0$ **to** *n-1* **do**
25   **if** `length`$(segList_i) > \eta_1$ **then**
26 $segmentsList.remove(segList_i)$ `// Remove the segments list of firm` i
27   **end if**
28   **if** `length`$(segList_i) > \eta_2$ **then**
29 $segList_i.remove(seg_0)$ `// Remove the first segment of firm` i
30   **end if**
31 **end for**
32 **return** $segmentsList$

Appendix B.

Table 10
Infeasible Portfolios of Sorting Contemporaneous *lead return* Sector Decomposition $\mathcal{LR}(\omega_{ij,t-365:t}^{w/c})$

Rank	Mean	SR	%MV	B/M	Liquidty	FF3 α	FF3 R^2	FF5 α	FF5 R^2
$\mathcal{LR}(\omega_{ij,t-365:t}^{w})$									
1	−90.35	−9.72	18.60	44.21	1.03	−92.98	15.26	−92.96	15.68
						(−26.81)		(−27.00)	
2	−27.87	−3.76	20.54	40.55	0.86	−30.4	17.69	−30.5	18.12
						(−11.71)		(−11.81)	
3	3.73	0.46	20.65	39.19	0.84	1.24	18.47	1.15	18.90
						(0.50)		(0.47)	
4	33.86	4.44	20.69	39.70	0.86	31.36	18.12	31.29	18.67
						(12.10)		(12.18)	
5	90.38	10.30	19.51	43.10	0.99	88.14	16.56	88.03	17.23
						(26.37)		(26.62)	
$\mathcal{LR}(\omega_{ij,t-365:t}^{c})$									
1	−0.18	−0.08	14.47	42.69	0.92	−2.65	18.12	−2.75	18.67
						(−1.00)		(−1.04)	
2	0.98	0.07	21.68	41.01	0.93	−1.64	20.29	−1.72	20.79
						(−0.63)		(−0.67)	
3	3.10	0.35	25.26	40.71	0.92	0.64	18.06	0.59	18.50
						(0.25)		(0.23)	
4	3.89	0.44	22.85	40.45	0.92	1.43	17.63	1.38	18.09
						(0.54)		(0.53)	
5	5.70	0.69	15.74	41.81	0.92	3.33	18.41	3.27	18.94
						(1.28)		(1.27)	

The upper panel reports the performance of the equal-weighted portfolios sorted by within-sector *lead return* $\mathcal{LR}(\omega_{ij,t-365:t}^{w})$, and the bottom panel contains the performance of the equal-weighted portfolios sorted by cross-sector *lead return* $\mathcal{LR}(\omega_{ij,t-365:t}^{c})$. Both news networks are formed with a 1-year rolling window. All the portfolios are infeasible. Column (1)-(5) are average simple daily return in bps; average annualized sharpe-ratio, average market value proportion in percentage, average book-to-market ratio in percentage, average volume-to-share ratio in percentage. The last 4 columns report the alpha, R^2, and robust Newy-West t-statistics in parentheses against Fama-French 3-factor and Fama-French 5-factor, respectively.

Table 11
Infeasible Portfolios of Sorting Contemporaneous *lead return* Signs of Return Decomposition $\mathcal{LR}^{+/-}(\omega_{ij,t-365:t})$

Rank	Mean	SR	%MV	B/M	Liquidty	FF3 α	FF3 R^2	FF5 α	FF5 R^2
$\mathcal{LR}^{+}(\omega_{ij,t-365:t})$									
1	−54.11	−7.63	12.85	43.72	0.91	−56.18	16.13	−56.22	16.56
						(−21.61)		(−21.67)	
2	−31.36	−4.22	20.19	40.12	0.88	−33.93	18.94	−33.95	19.38
						(−13.25)		(−13.35)	
3	−5.13	−0.72	24.18	39.64	0.87	−7.80	18.43	−7.88	18.87
						(−2.99)		(−3.04)	
4	23.37	2.85	23.63	40.22	0.89	20.78	17.40	20.69	17.88
						(7.51)		(7.53)	
5	77.31	8.40	19.15	42.99	1.03	74.81	17.32	74.71	17.97
						(21.99)		(22.23)	
$\mathcal{LR}^{-}(\omega_{ij,t-365:t})$									
1	55.67	7.87	12.72	42.79	0.90	53.73	16.82	53.68	17.41
						(20.38)		(20.58)	
2	39.01	5.10	20.63	39.60	0.88	36.59	18.57	36.50	19.20
						(13.80)		(13.91)	
3	13.04	1.64	24.52	39.50	0.87	10.57	18.05	10.47	18.51
						(4.04)		(4.04)	
4	−15.80	−2.00	23.63	40.95	0.89	−18.57	18.68	−18.68	19.16
						(−6.73)		(−6.82)	
5	−74.98	−8.14	18.50	43.88	1.04	−77.75	15.91	−77.76	16.33
						(−23.37)		(−23.52)	

The upper panel reports the performance of the equal-weighted portfolios sorted by positive *lead return* $\mathcal{LR}^{+}(\omega_{ij,t-365:t})$, and the bottom panel contains the performance of the equal-weighted portfolios sorted by negative *lead return* $\mathcal{LR}^{-}(\omega_{ij,t-365:t})$. Both news networks are formed with a 1-year rolling window. All the portfolios are infeasible. Column (1)-(5) are average simple daily return in bps; average annualized sharpe-ratio, average market value proportion in percentage, average book-to-market ratio in percentage, average volume-to-share ratio in percentage. The last 4 columns report the alpha, R^2, and robust Newy-West t-statistics in parentheses against Fama-French 3-factor and Fama-French 5-factor, respectively.

Table 12

Estimation of the Impacts from Sector and Sign of Returns Decomposed Network Price Proxies

Depend. Var.	$\mathcal{LR}^+(\mathcal{W}_T^w)$	$\mathcal{LR}^+(\mathcal{W}_T^c)$	$\mathcal{LR}^-(\mathcal{W}_T^w)$	$\mathcal{LR}^-(\mathcal{W}_T^c)$
r_{t+1}	0.031	-0.005	0.007	0.015
	(4.701)	(-0.968)	(1.155)	(2.339)

A 1-year time window for news network $T = [t - 365 : t]$. Estimation the impacts of *lead return* decomped by sector and sign of return on 1-day ahead stock returns. Coefficients of *lead return* variables are estimated by fixed effect panel model, and t-statistics reported in parentheses with the standard errors corrected by clustering on individual and time. Control variables include market value $\log(MV)$, book-market ratio B/M, and turnover rate *volume/share* and 1-day lagged return r_t. All the financial statement related variables are 2 quarters lagged.

Table 13
Portfolios Sorted by the Preceding *Network Degree* $\mathcal{A}$
(Equal-Weighted, Keep $\mathcal{A}_{i,t} = 0$)

						FF3		FF5	
Rank	Mean	SR	%MV	B/M	Liquidty	α	R^2	α	R^2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	0.48	0.21	8.66	44.55	0.89	−0.77	91.52	−0.73	92.33
						(−2.81)		(−2.77)	
2	0.58	0.27	10.12	42.57	0.88	−0.78	91.71	−0.75	92.33
						(−3.02)		(−2.97)	
3	0.64	0.31	12.85	42.40	0.90	−0.74	92.25	−0.71	92.71
						(−3.26)		(−3.29)	
4	0.51	0.22	18.97	39.84	0.91	−0.80	88.40	−0.74	89.74
						(−2.91)		(−2.80)	
5	0.98	0.57	49.40	36.56	1.00	−0.37	92.12	−0.35	92.99
						(−1.93)		(−1.78)	
5-1	0.51	0.82				0.40	43.07	0.38	43.68
						(2.07)		(1.97)	
Mkt	0.64	0.31				−0.69	92.56	−0.65	93.34
						(−3.03)		(−2.97)	

The table reports the performance of the equal-weighted portfolios sorted by *network degree*, $\mathcal{A}(\omega_{ij,t-30:t})$, computed with news linkages from the preceding month. Firms with zero network degree are kept in the portfolios. Columns (1)-(5) are average simple daily return in percentage; the average annualized sharpe-ratio, average market value proportion in percentage, average book-to-market ratio in percentage, average volume-to-share ratio in percentage. The last 4 columns report the alpha, R^2, and robust Newy-West t-statistics in parentheses with respect to the Fama-French 3-factor and the Fama-French 5-factor models, respectively.

Table 14
Table 14
Portfolios Sorted by the Preceding *Network Degree* $\mathcal{A}$
(Value-Weighted, Drop $\mathcal{A}_{i,t} = 0$)

Rank	Mean	SR	%MV	B/M	Liquidty	FF3		FF5	
						α	R^2	α	R^2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	0.47	0.22	6.76	36.63	0.73	−0.84	90.38	−0.82	90.8
						(−3.42)		(−3.41)	
2	0.57	0.31	8.87	34.76	0.70	−0.72	92.99	−0.7	93.27
						(−4.09)		(−4.11)	
3	0.65	0.38	13.05	34.16	0.67	−0.49	88.71	−0.48	88.79
						(−2.63)		(−2.65)	
4	0.56	0.31	19.02	30.28	0.63	−0.72	88.13	−0.7	88.47
						(−3.2)		(−3.06)	
5	1.11	0.72	52.31	32.47	0.65	−0.26	95.39	−0.25	95.64
						(−2.25)		(−2.19)	
5-1	0.64	0.95				0.58	26.43	0.56	27.46
						(2.53)		(2.54)	
Mkt	0.85	0.52				−0.53	95.19	−0.52	95.47
						(−4.25)		(−4.08)	

The table reports the performance of the value-weighted portfolios sorted by *network degree*, $\mathcal{A}(\omega_{ij,t-30:t})$, computed with news linkages from the preceding month. Firms with zero network degree are removed before forming the portfolios. Columns (1)-(5) are average simple daily return in percentage; the average annualized sharpe-ratio, average market value proportion in percentage, average book-to-market ratio in percentage, average volume-to-share ratio in percentage. The last 4 columns report the alpha, R^2, and robust Newy-West t-statistics in parentheses with respect to the Fama-French 3-factor and the Fama-French 5-factor models, respectively.

Appendix C.

- Title: "3 Catalysts for Apple Stock in 2021"
- Timestamp: "2020-12-30 12:00:42"
- Content: "The NASDAQ-100 Technology Sector index has more than doubled the return of the S&P 500 index over both the last one- and five-year periods, but Apple (NASDAQ: AAPL) has remained among the cream of the crop, surging roughly 84% so far in 2020 and about 405% over the last five years. Those heady gains put Apple's stock price at a premium valuation of 34 times forward earnings estimates, which looks expensive compared with a forward price-to-earnings multiple of 24 for the S&P 500. However, Apple is currently entering one of its strongest product cycles in years, which could keep the business humming along and justify the stock's premium. Here are three growth catalysts to watch in 2021. Apple enters the new year after delivering a strong earnings report for the fiscal fourth quarter. While iPhone revenue was slightly down in fiscal 2020, sales of non-iPhone products grew 30% year over year despite supply constraints on iPad, Mac, and Apple Watch. That momentum should continue into the first half of calendar 2021 with recent product releases. The new AirPods Max headphones have been sold out since their initial release in December. Other recent launches should provide sales momentum in the short term, including the HomePod mini, Apple Watch Series 6 and Watch SE, two new iPad models, and new MacBooks featuring Apple's internally developed M1 processor. Early signs are pointing to strong sales of iPhone 12. Even though 5G network coverage is spotty, customers seem to be scooping up the new iPhone at a record clip, which is great news since the iPhone generates half of Apple's total revenue. Apple is planning to increase its production of the iPhone 12 by 30% in the first half of 2021, according to a report from Nikkei Asia that cited a source from a key Apple supplier. This would put Apple on pace to have its biggest year for iPhone sales since the record-breaking shipments of the iPhone 6 in 2015. Apple made a bold move earlier this year by announcing a two-year transition to use internally developed processors for its line of Macs, dropping Intel chips in the process. The M1 chip offers several features that will significantly enhance the user experience on Mac and could lead to further market share gains in the PC market for Apple. For years, Apple has been working to bridge the user experience across its operating system for Mac and iOS. The M1 chip will take this a step further by allowing Mac users to run iPhone and iPad apps. This could be huge for Mac sales over the long term. By taking control over the development of its own chips, Apple can better plan its product road map and tailor future versions of the M1 for specific user experiences, such as enhanced image processing, security, and other cutting-edge features and technologies. Apple will likely unveil more benefits of the M1 chip over time, but for 2021, the significant boost to battery life and ability to run iOS apps directly on Mac should be enough to encourage more sales of MacBooks. Despite double-digit percentage growth from subscription services, sales of hardware products still make up nearly 80% of Apple's total revenue. But with services growing 16% year over year in the last quarter, this $53 billion annual business could reach $100 billion in the next five years. After reaching an installed base of 1.5 billion active devices earlier this year, Apple reported that its installed base hit another record high in the fiscal fourth quarter. New services – including Apple TV+, Arcade, News+, and Apple Card – are attracting more users, and Apple is still adding new services and content to drive further growth. Apple TV+ continues to add new streaming content, which will be crucial to persuade users coming off their free trials that it's worth paying the relatively low monthly fee of $4.99. The recent launch of Apple Fitness+ is yet another service with a lot of potential. Interactive fitness was already a fast-growing market before 2020, but it got an extra kick during the pandemic with more people looking for alternative workout solutions at home. Nike and Peloton Interactive have reported high engagement levels with their respective training apps lately. Apple's massive installed base of users should win a decent share of this booming market. Apple is on the verge of a major upgrade cycle across all its products. Moreover, this upgrade cycle could intensify as COVID-19 vaccines become available, encouraging more people to visit Apple stores as the year progresses. With these growth catalysts on the horizon for this top tech stock, Apple is well-positioned to outperform in 2021. 10 stocks we like better than AppleWhen investing geniuses David and Tom Gardner have a stock tip, it can pay to listen. After all, the newsletter they have run for over a decade, Motley Fool Stock Advisor, has tripled the market. David and Tom just revealed what they believe are the ten best stocks for investors to buy right now... and Apple wasn't one of them! That's right – they think these 10 stocks are even better buys. See the 10 stocks Stock Advisor returns as of November 20, 2020 John Ballard owns shares of Apple, Nike, and Peloton Interactive. The Motley Fool owns shares of and recommends Apple, Nike, and Peloton Interactive. The Motley Fool recommends Intel. The Motley Fool has a disclosure policy."

Figure 12: Sample News on Apple, Intel, Nike, Peloton

The sample news that is used in the demonstration of the algorithm flowchart in Figure 2. Apple is identified in the headline, while Intel and Nike are identified in the content. Peloton is not identified as it is not in the S&P500 index.

- Title: "Dow Jones Industrial Average Rises 9 Points Despite Apple Stock Slump"

- Content: "MoneyMorning.com Report - For Jan. 5, 2016, here's the top stock market news and stocks to watch based on today's market moves. Dow Jones
  ...
  The Dow received a boost from DuPont Co. (DD), which gained 1.9%. The top stock news came from Apple Inc. (Nasdaq: AAPL), which saw shares slump 2.5% after the Nikkei Asian Review reported the company will slash its iPhone production by roughly 30% this quarter
  ..."

- Identified Firm Tickers: MON,TWTR,FB,CVX,XOM,F,FSLR,DIS,GS,AMZN,DD,GM

Figure 13: Sample News of Market Summary Report

This sample news demonstrates the possibility of market report news having an iconic firm in the headline.

IRTG 1792 Discussion Paper Series 2021

For a complete list of Discussion Papers published, please visit
http://irtg1792.hu-berlin.de.

IRTG 1792 Discussion Paper Series 2021

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.