Economic Linkages Inferred from News Stories and the Predictability of Stock Returns

Anna Scherbina* Bernd Schlusche†

UC Davis *Federal Reserve Board*

First draft: December 3, 2013

This draft: September 16, 2015

ABSTRACT

We show that news stories contain information about economic linkages between firms and document that information diffuses slowly across linked stocks. Specifically, we identify linked stocks from co-mentions in news stories and find that the average return of the stocks linked to a particular stock predicts that stock's return in the following month. Content analysis of common news stories reveals many types of firm linkages that have not been previously documented. Importantly, the cross-predictability in returns remains even after firm pairs with customer-supplier ties are removed. We also show that information can flow from smaller to larger stocks. Our results indicate that slow reaction to the news of linked stocks is consistent with limited attention and slow processing of complex information.

JEL classification: G10, G12, G14, G17

Keywords: News Media, Soft Information, Linked Stocks, Information Leadership, Lead-Lag Effect, Complex Information, Limited Attention, Market Efficiency

*Address: Graduate School of Management, University of California, Davis, One Shields Avenue, Davis, CA 95616. E-mail: ascherbina@ucdavis.edu. Phone: (530) 754-8076.

†Address: Board of Governors of the Federal Reserve System, 20th Street and Constitution Avenue NW, Washington, DC 20551. E-mail: bernd.schlusche@frb.gov. Phone: (202) 452-2591.

We are grateful to seminar participants at Pennsylvania State University, Tulane University, Michigan State University, Temple University, Johns Hopkins University, the Federal Reserve Board, Georgetown University, Acadian Asset Management, State Street Global Advisors, the Citi Quant Conference, the Financial Research Association early ideas session, Western Finance Association Annual Meetings, the 7th Annual Florida State University SunTrust Beach Conference, and the European Finance Association Meetings as well as our discussants, Oguzhan Ozbas (WFA), Umit Gurun (SunTrust Conference), and Jordi Mondria (EFA), for helpful comments. We are grateful to Thompson Reuters for making the TRNA dataset available for our research. The analysis and conclusions set forth are those of the authors alone and do not indicate concurrence by the Board of Governors or other members of the research staff.

Economic Linkages Inferred from News Stories and the Predictability of Stock Returns

ABSTRACT

We show that news stories contain information about economic linkages between firms and document that information diffuses slowly across linked stocks. Specifically, we identify linked stocks from co-mentions in news stories and find that the average return of the stocks linked to a particular stock predicts that stock's return in the following month. Content analysis of common news stories reveals many types of firm linkages that have not been previously documented. Importantly, the cross-predictability in returns remains even after firm pairs with customer-supplier ties are removed. We also show that information can flow from smaller to larger stocks. Our results indicate that slow reaction to the news of linked stocks is consistent with limited attention and slow processing of complex information.

JEL classification: G10, G12, G14, G17

Keywords: News Stories, Journalists, Soft Information, Linked Stocks, Information Leadership, Lead-Lag Effect, Limited Attention, Market Efficiency

I. Introduction

A corporate news event often affects not only the firm at the center of that event but also a number of other companies in similar circumstances. On April 17, 2013, the Supreme Court handed a 9-0 victory to Royal Dutch Petroleum (RDP) in the Kiobel v. Royal Dutch Petroleum case. RDP was sued by 12 Nigerian citizens, who claimed that the firm cooperated with the Nigerian government in the 1990s to brutally crush the resistance to oil development in the country. The case attracted considerable attention, in part, because it was about whether foreign citizens may seek compensation in U.S. courts for human-rights violations committed by firms outside the United States. About 150 such lawsuits have been filed in the United States in the past three decades, and these lawsuits were considered to be costly to settle and damaging to firm's reputations.[1] On September 30, 2012, Reuters reported on the firms filing amicus briefs in support of RDP and speculated that "a ruling against Kiobel could wipe out lawsuits pending against companies such as Exxon Mobil Corp, Rio Tinto Plc and Nestle, which are accused by private plaintiffs of helping governments violate human rights in Indonesia, Papua New Guinea and Ivory Coast, respectively."[2] Other business publications concurred that in the case of RDP's victory virtually all similar lawsuits would be dismissed.[3]

Unlike the firm at the center of an unfolding event, such RDP in the example above, the other firms also affected by the news are less immediately obvious. However, journalists following a particular set of firms and/or events often provide a broad analysis of the affected parties. Further helping to uncover firm linkages, Regulation Fair Disclosure requires that firms report significant events and business deals that have the potential to affect their stock

[1] *The Economist*, "The Shell game ends," April 20, 2013.

[2] Indeed, on August 8, 2012 (and on February 3, 2013, due to the Court's decision to hear additional arguments), a number of firms, such as Caterpillar, ConocoPhillips, General Electric, Honeywell International, IBM, and Monsanto Company (these firms filed a joint brief), Coca-Cola, KBR, Engility Corporation, Chevron, and Rio Tinto Group filed amicus briefs in support of RDP. A subset of these firms wrote in their amicus briefs that they had similar ongoing lawsuits.

[3] See, e.g., Forbes.com, "Supreme Court Observations: Kiobel v. Royal Dutch Petroleum & the Future of Alien Tort Litigation," April 18, 2013.

price to a broad audience and with a minimal delay, and press releases are a recommended way to comply. Many corporate press releases are then re-printed in the financial press. Hence, news coverage is likely to contain soft information about economic linkages between firms. Consider the explosion of the Deepwater Horizon drilling rig in the Gulf of Mexico on April 20, 2010. British Petroleum (BP) immediately took the brunt of the blame for the resulting oil spill as the leaser of the drilling rig and majority owner of the exploration rights. However, the ongoing coverage of BP and its operations may have revealed other potential culprits, such as Transocean Ltd. (the operator of the rig), Halliburton Energy Services Inc. (the cementer of the potentially faulty well), and Cameron International Corp. (the supplier of the blow-out preventers, which were intended to prevent such an oil spill) even before these firms were sued for their role in the disaster.

In this paper, we show that information diffuses slowly across linked firms and that economic linkages between firms identified through co-mentions in news stories are useful in predicting the stock returns.[4] If a firm experiences a news shock, its stock price will react quickly, but the linked stocks that are also affected by the news may react with a delay due to slow processing of complex information and limited investor attention. Indeed, the amount of corporate news that hits the market every day is immense. Based on a nearly complete set of corporate press releases for the period between April 2006 and August 2009, Neuhierl, Scherbina, and Schlusche (2013) document that, on an average day, a total of 281 valuation-relevant news items are released by all U.S.-based firms, which alone may explain, to a large extent, why information diffuses slowly. That study additionally shows that a large fraction of the news releases is of non-routine nature whose impact may be hard to quantify. Of course, it is even more difficult to quantify the impact a firm's news on its economically-linked firms. We show that the attention of sophisticated investors helps ensure a quicker price reaction to the news of linked stocks. Moreover, frequent co-mentions in news stories,

[4]Throughout the paper, we use the terms "linked stocks," "economically-linked stocks," and "connected stocks" interchangeably.

which make investors more aware of firm linkages, ensure quicker information diffusion. Thus, limited investor attention and slow processing of complex information both contribute to slow information diffusion across linked stocks.

Our approach to identifying stock linkages is novel in that we exploit news media's soft information to uncover linkages between firms, and we show that the information available ex-ante from the news coverage can be used to exploit the lead-lag relation in the returns of linked stocks. We begin by identifying, prior to each month t, all stocks $j = 1, \ldots, J^i$ that were co-mentioned with stock i in a news story in the Thomson-Reuters News Analytics (TRNA) dataset in the preceding three (or six or 12) months. If such linkages exist for stock i, we proceed to compute its "linked-stock signal" as the equal-weighted average of the prior month's returns of stocks J^i. We hypothesize that, in period t, stock i will move in the direction of its linked-stock signal. Our portfolio results show that the linked-stock signal possesses reliable predictive ability for the return of linked stocks. In a series of cross-sectional regressions, we confirm the predictive ability of the linked-stock signal and provide further evidence that the linked-stock signal is a novel predictor of future stock returns and that its predictive ability is not subsumed by previously known return predictors both at the firm and industry levels.

Our setting allows us to provide new insights into the process of information diffusion in the stock market. Interestingly, even though the coverage in our sample of news stories is heavily tilted towards larger firms, we find significant delays with which stock prices incorporate new information from their linked firms. Nevertheless, markets seem to be reasonably efficient. First, we find that the speed of information diffusion has increased over time. Second, despite failing to incorporate new information quickly, prices may be approximately efficient and lie within no-arbitrage bounds around the fair value (see, for example, Lo (2004)). Our analysis of break-even trading costs for simple trading strategies designed to exploit the lead-lag effect documented here shows that, even though prices fail to incorporate the relevant information from economically-linked stocks quickly, trading on this information is costly due to high portfolio turnover. Significant after-trading-cost profits can be achieved only by sophisticated traders.

Our paper contributes to the literature on lead-lag effects in stock returns. Studies in this literature typically use ex-ante stock characteristics to explain the lead-lag return patterns. Lo and MacKinlay (1990)—the paper that started this literature—show that returns of large firms tend to predict returns of small firms.[5] Subsequent studies identify various measures of investor attention that are associated with information leadership. Specifically, Brennan, Jegadeesh, and Swaminathan (1993), Badrinath, Kale, and Noe (1995), and Chordia and Swaminathan (2000) use analyst coverage, institutional ownership, and trading volume, respectively, to show that common information diffuses slowly from stocks with high levels of investor attention to those with low levels. More recently, Cohen and Lou (2012) show that information diffuses slowly from single-segment firms to multi-industry conglomerates and Menzly and Ozbas (2010) and Cohen and Frazzini (2008) find that information travels slowly among firms linked by the supply chain.[6] In contrast, Scherbina and Schlusche (2013) do not use ex-ante available stock characteristics or information to identify leader stocks. Rather, the identification is based on the statistical ability of leader stocks to Granger-cause the returns of their followers.

The approach here is most similar to Scherbina and Schlusche (2013) in that we do not rely on ex-ante available stock characteristics or required disclosures to identify lead-lag relations. By analyzing the content of common news, we are able to identify additional types of economic linkages between firms, such as business partnerships, potential M&A deals, similar legal and

[5] Hou (2007) shows that this effect works within industries, i.e., large stocks lead small stocks within the same industry. This effect is most recently confirmed by DeMiguel, Nogales, and Uppal (2014); furthermore, the authors argue that it is important to consider cross-covariances in stock returns when constructing mean-variance-efficient portfolios.

[6] More recent work documents excessive contemporaneous return correlations among stocks with common institutional ownership (Anton and Polk (2014)), common analyst coverage (Israelsen (2013) and Muslu, Rebello, and Xu (2014)), and textually similar financial reports and news stories (Hoberg and Phillips (2012) and Box (2014)), and attributes the excessive correlations to similarity in trading or the information environment.

regulatory exposures, labor and production commonalities, and so on. Even after we remove firm pairs tied by customer-supplier links, the cross-predictability in returns remains. While most topic categories are too small to analyze them individually, we document the cross-predictability in returns between firms linked exclusively by partnerships and investment banking prospects, by similar legal developments, and by operational similarities.

Identifying firm linkages from news offer additional advantages. First, this methodology allows to fill the gaps in data availability induced by limited disclosure (for example, firms are required by the SEC to report the identity only of a customer that comprises more than 10% of a firm's consolidated sales revenues, and hence smaller customers will be missing from the dataset). Second, it allows to uncover transitory leaders (for example, the RDP's leadership should cease shortly after the Supreme Court has reached its decision, and RDP is no longer be co-mentioned with firms in a similar legal situation), whereas lead-lag relationships are assumed to be somewhat permanent in the prior literature. Third, our methodology permits within-industry bets, while in the lead-lag literature intra-industry bets are, for the most part, precluded. Finally, similarly to Cohen and Lou (2012) and Scherbina and Schlusche (2013), we are able to show that smaller stocks can lead returns of larger stocks.

We also contribute to the relatively new strand of literature that investigates the news media's role in financial markets. This literature examines the extent to which qualitative information in news articles—in particular, pessimism in the language—is associated with stock price reactions.[7] The literature also shows that news coverage helps improve price efficiency (see, e.g., Peress (2008) and Peress (2014)) and reduce firms' cost of capital (e.g., Fang and Peress (2009)). In this study, we shed light on a different aspect of news media's

[7] Tetlock (2007) focuses on news articles about the broad stock market and documents that the linguistic tone predicts aggregate stock prices. Similarly, Tetlock, Saar-Tsechansky, and Macskassy (2008) find that content in firm-specific news articles is correlated with the returns of individual stocks. Following a different identification strategy, Dougal, Engelberg, Garcia, and Parsons (2012) show that the identity of columnists writing for the *Wall Street Journal* significantly predicts returns of the Dow Jones Industrial Average. Engelberg and Parsons (2011) document that trading behavior of investors in a number of local markets in response to the same event differs depending on local media coverage.

role in financial markets: We provide evidence that news stories contain soft information that helps identify economic linkages between firms.

One possible extension is to check whether our findings can help address the "R^2 puzzle," famously articulated by Roll (1988). That paper investigates whether traditional asset pricing models can explain reasonably well daily stock price movements of 96 large firms and finds that the traditional pricing factors used in return regressions, which may include industry controls, produce only low R^2s; furthermore, the R^2s computed for the subsample of no-news days are only slightly higher than the R^2s computed for the subsample of news days, on which one would expect large idiosyncratic price movements.[8] Building on this analysis, Boudoukh, Feldman, Kogan, and Richardson (2013) are able to increase CAPM R^2 differences between news and no-news days by defining as news days only those days with truly price-relevant firm-level news, which the authors identify with textual analysis. Our results suggest that the difference in R^2s between no-news and news days can be further increased by adding to the set of news days those days on which significant news developments occurred for linked firms. Importantly, linked firms frequently belong to different industries, and hence a firm's industry factor, which is used in this literature to account for industry-wide developments, is unable to fully capture all the relevant news of other firms.

The paper is organized as follows: Section II describes the data. Section III explains our methodology to identify linked stocks. Section IV documents the ability of linked stocks to cross-predict each others' returns. Section V concludes.

II. Data

Stock-specific data in this paper are obtained from the daily stock file of the Center for Research in Security Prices (CRSP) and include all NYSE-, Amex-, and Nasdaq-traded

[8]News days are defined as days on which the firm appeared in the Dow-Jones news service or in *The Wall Street Journal*.

stocks in the dataset. We adjust stock returns for delisting in order to avoid survivorship bias (Shumway (1997)) as follows: When a stock is delisted, we use the delisting return from CRSP, if available. Otherwise, we assume the delisting return to be -100%, unless the reason for delisting is coded as 500 (reason unavailable), 520 (went to OTC), 551-573, 580 (various reasons), 574 (bankruptcy), or 584 (does not meet exchange financial guidelines). In these cases, we assume that the delisting return is -30%. Accounting data are obtained from the merged CRSP/Computstat database.

Industry classifications and daily factor returns are obtained from Kenneth French's web site.[9] Throughout the paper, we use 38 industry classifications, but our results are nearly unchanged when using 12 industry classifications instead. Table A1 in the Appendix presents, for our sample, the average fractions of firms in each industry. The industry classified as "Irrigation Systems" drops out of our sample after data restrictions explained below are imposed, reducing the number of industries to 37. Additionally, whenever portfolio sorts are performed within industries or when leaders are required to belong to a different industry than their followers, we drop stocks in the industry identified as "Other" because of the implied heterogeneity (however, as can be seen from the table, this industry has very few stocks).

Finally, the news data are available from the Thompson-Reuters News Analytics (TRNA) dataset for the period April 1996 through December 2012. The dataset links news stories to TRNA's firm IDs (which we link to ticker symbols and then to permno's) and assigns news topic codes to each news item. Each distinct news story about a firm is labeled with a unique primary news access code (PNAC). The TRNA dataset provides news headlines and news sources, as well as a variety of quantitative scores for the news items (such as the number of words and sentences in the story, news newness scores, etc.). One score of interest to us is the sentiment score that takes on values +1, -1, or 0, indicating whether a story is positive,

[9] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_48_ind_port.html.

negative, or neutral. This score is roughly based on the predominance of positive/negative words in the news text. We will rely on the sentiment score when deciding whether or not the co-mentioned stocks are competitors. (A more detailed description of the TRNA dataset is provided in Scherbina and Schlusche (2013).)

For the purposes of this analysis, we discard news stories that could randomly combine pairs of stocks. Specifically, we remove stories about technical analysis, large price movements, index changes, bond credit ratings, listings and delistings of equity, shareholder meetings, analyst recommendations, investment fund news, and trade order imbalances. Some large news stories are transmitted in parts, and we consider only the complete story in order not to double count the same news.[10] With these conditions imposed, the total number of news items in our sample decreases by roughly one-third. When these conditions are not imposed and all news are kept, more unrelated stock pairs are erroneously identified through common news coverage as economically-linked stocks. However, this simply adds noise to the estimates of the linked-stock signal and does not significantly affect the results.

The tables and figures presented throughout the paper cover the returns period from July 1996 to January 2013, unless stated otherwise, as the initial three (six or 12) months are used to identify stocks linked through common news stories, and the linked-stock returns are forecasted one month ahead.

III. Identifying Linked Stocks

A group of linked stocks with respect to stock i in period t is comprised of all publicly traded stocks $j = 1, \ldots, J^i$ that were mentioned in at least one common news story with stock i in the Thomson-Reuters News Analytics (TRNA) dataset in a trailing three (six or 12) month

[10]Specifically, we remove stories with topic codes INSI, STX, HOT, INDX, AAA, LIST1, USC, MEVN, RCH, FUND, and DBT. We also discard stories with item genre 'Imbalance' (with different capitalization possibilities). Finally, as per Reuters' suggestion, we delete news observations for which the variable "more_news" takes on values 'M' or 'm' and for which the variable "update_sz" is greater than 8500 in order to only consider complete news stories rather than the pieces in which they were transmitted.

period, excluding the five trading days prior to the end of period $t-1$. We exclude these days to ensure that a released piece of news has become available to the broad market and to allow investors sufficient time to identify linked stocks before forming signals.[11]

Our methodology for identifying linked stocks is illustrated in Figure 1. In the examples provided in the figure, stock A is linked to stocks B, C, and D, and stock B is linked to stocks A and E. The dashed and dotted ellipses indicate the groups of stocks linked to stocks A and B, respectively, which are used to compute signals for these two stocks.

Table I provides descriptive statistics of the TRNA dataset as well as descriptions of the stocks linked by common news stories. As shown in Panel A, the TRNA dataset contains a total of almost 5.5 million unique news stories. This sample is reduced by about a third after removing news that would randomly group stocks together (as described in the Data section). Of the remaining news sample, just over 14% mention more than one firm. As reported in Panel B of the table, this subset of news mentions 2.78 firms, on average. The median number of firms mentioned is 2.

Despite removing what we consider to be irrelevant stories for our purposes, we are still concerned that firms may be grouped together in a story by coincidence, such as when reporting on market conditions or unrelated events (e.g., contemporaneous earnings announcements). Therefore, we further narrow our news sample, to the news that co-mention exactly two firms. The sample is further reduced, to 331,232 unique news stories.

As we discuss in more detail in Section IV, our return predictability signal hinges on the presumption that after a shock to a stock's price the prices of stocks linked to that stock move in the same direction. Hence, in order to minimize the probability of linked stocks to move in opposite directions, we eliminate competitor stock pairs in two steps. First, we argue that two firms are competitors if the difference between their assigned sentiment scores for a particular news story in the TRNA dataset is two (i.e., firms are considered competitors if a news story about them is interpreted as "positive" for one firm and "negative" for the

[11] Throughout the paper, we refer to this period as the "identification period."

other.). This algorithm eliminates about 10% of the stories that mention exactly two firms. In addition, we discard news stories that contain a variation of the words "competitor" and "rival" in the headline, as these stories are likely to cover competing firms. As a validity check, in untabulated results we confirm that the words "competitor" and "rival" in the headline are associated with more than twice as high a probability of the firms receiving opposite sentiment scores. This final step reduces our dataset to 299,060 unique news stories.

Panels C through G of the table produce additional descriptive statistics on the final dataset. Panel C shows that the majority of firm pairs, 64.48% and 62.20%, are co-mentioned only once during the 3- and 6-months identification window, respectively. Panel D shows that there is a higher than random chance that both stocks in the pair are in the same industry. Panel E shows that only 10.51% (15.12%) of stock-months are linked when 3-month (6-month) identification window is used. Moreover, since the news media tends to cover more extensively larger stocks with higher trading volume, linked stocks tend to have substantially higher market capitalization than unlinked stocks. Consistently, Panel F shows that the number of linked stocks goes up as a stock's market capitalization increases.

Panel G shows that economic linkages between stocks show significant persistence over time (the length of the identification window used for this table is 3 months). Specifically, the probability that a particular pair of linked stocks in month τ continues to be identified as such in month $\tau + 12$ is about 22.86%. Even five years later, this probability is 18.43%. For stock pairs that had at least two common news stories over identification period and that are thus presumably more closely linked, the persistence is even stronger, with the probability of linkage in month $\tau + 12$ equal 42.76%.

IV. Return Predictability

Having identified linked stocks, we next need to determine the direction of the information flow. Stocks with an above-normal turnover likely had experienced a news event. Stocks with turnover within the normal range likely have not. We assume that return leaders are those stocks with the relatively high turnover in month t and followers those in the normal turnover range in month t, and that information flows from thus-identified leaders to their followers. In determining turnover thresholds for leaders and followers, we have to balance the considerations of the precision of our identification against reducing the sample size too much.

In each period t, we compute for each stock i, for which turnover in period $t-1$ was less than its 75th-percentile turnover over the trailing 12-month window, the aggregate linked-stock signal, $Signal_t^i$, simply as the weighted average of the lagged returns of stocks $j = 1, \ldots, J^i$ that are linked to stock i and for which turnover in period $t-1$ was above the median turnover over the trailing 12-month window.

$$Signal_t^i = \sum_{j=1}^{J_t^i} \omega_j Ret_{t-1}^j, \tag{1}$$

where Ret_{t-1}^j is stock j's return in period $t-1$, excluding the last five trading days of that period to mitigate concerns regarding non-synchronous trading and to allow investors sufficient time to compute signals, and ω_j is the weight assigned to stock j's return. In the baseline specification, signaling stocks' returns are equal-weighted, in which case $\omega_j = 1/J^i$ (see the examples in Figure 1).

In robustness checks, we consider alternative weighting schemes for the linked-stock returns. In the first scheme, the signaling stocks' returns are weighted by the number of common news stories that linked stock pairs in the preceding 3 months. The second scheme is value-weighting, using the linked stocks' market capitalization in period $t-2$.

For the stocks that receive signals, we require that they be priced at or above \$5 per share and that they traded on the last day of period t. Moreover, we limit the set of stocks receiving signals to common shares of U.S.-incorporated firms by considering only stocks with share codes 10 or 11.

A. Baseline specification

Each month t, we identify linked-stock pairs as of five trading days prior to the end of calendar month $t-1$, and compute linked-stock signals using the returns in month $t-1$ (excluding the last five trading days). Quintile portfolios are formed *within* each industry in the beginning of month t and held for one month. Then, same-number portfolios are combined ensuring that each industry has the same representation in each portfolio 1 through 5.[12] (Forming portfolios within industries enables us to show that we are not simply picking up the industry momentum effect or the large-small stock lead-lag effect within industries.) The process is repeated in the subsequent month. The timeline used for the portfolio formation is illustrated in Figure 2.

Table II reports various return measures—excess return, market alpha, and three-, four-, and five-factor alphas—for equal-weighted portfolios (Panel A) and value-weighted portfolios (Panel B) as well as return differentials between the extreme quintile portfolios.[13] For reference, Panels A and B of Table A2 present the factor loadings for equal- and value-weighted portfolios, respectively. Note that the factor loadings are about unchanged across various quintile portfolios.[14]

[12]If in a given month there are fewer than five eligible stocks in an industry, that industry/month observation is discarded.

[13]The reported t-statistics are adjusted for autocorrelation in returns using the methodology in Newey and West (1987), and, for each specification, the number of lags is determined as the third root of the number of observations in the time series.

[14]Burt and Hrdlicka (2015) argue that the tests of the return predictability of economically linked stocks should be conducted based on four-factor alphas rather than raw returns. The authors argue that sorting on alphas as opposed to raw returns will produce a more precise measure of abnormal returns, which could improve the return predictability. (It may also decrease the observed return predictability if excess return pick up similar factor loadings of the economically linked stocks.) We have tried the method suggested by the authors. In the first step, we have calculated daily four-factor alphas over a trailing 12-month window (when a stock had less than 6 months of daily return observations, in order to keep the same sample size, we have set the alpha to the excess return). Next, we have added up the daily alphas of a signaling stock in month $t-1$, having skipped the last five trading days and proceeded to calculate linked-stock signals and

The results indicate that signals from linked stocks have significant predictive ability for stock returns for both equal- and value-weighted portfolios. Portfolios that are comprised of stocks with low linked-stock signals earn low returns, while portfolios containing stocks with high signals earn high returns. Portfolio returns increase gradually in magnitude with the linked-stock signal. Return differentials between quintiles 5 and 1, reported in the bottom row of the table, are positive and statistically significant. We consider these differentials to be economically meaningful as the excess returns produced by the long-short portfolios are over 9% annually for both weighting schemes.

As shown in the second column, stocks in the extreme quintile portfolios tend to be linked to fewer stocks than those in the other quintile portfolios. That is, the predictive ability of the link-stock signal decreases with the number of linked stocks a firm has.

Importantly, if the returns of linked stocks were not driven by information but, perhaps, by correlated trading, then the returns in the month following portfolio formation would reverse and the performance of the long-short portfolios in subsequent months would be negative. However, we do not find any return reversal. In fact, the average monthly raw return differential between month $t+2$ and $t+6$ after the month of portfolio formation is slightly positive but insignificant for both equal- and value-weighted portfolios.[15]

Moreover, results from a subsample analysis, presented in Panel C of Table II, shows that the return differentials have declined over time. While the five-factor alphas of the return differentials for the period from July 1996 to December 2003 are 1.45% and 1.19% per month for the equal- and value-weighted portfolio, respectively, they drop to 0.35% and 0.37% per

form portfolios as in our main specification. Since we have already conditioned on turnover and adjusted the subsequent returns for factor loading, calculating linked-stock signals based on alphas as opposed to raw returns does not significantly change the results. In particular, the five-factor monthly alpha of the equal-weighted long-short portfolio increases by 0.09% from 0.82% to 0.91% and the corresponding t-statistic increases from 3.82 to 4.11. The five-factor alpha of the return differential on the value-weighted portfolio decreases slightly, by 0.01%, from 0.78% to 0.77%, with the t-statistic decreasing from 3.27 to 3.16. For the simplicity of replication, and because the results are very similar, we will report the results based on the linked-stock signals calculated from raw returns throughout the paper.

[15]This monthly raw return differential is 0.10% (t-statistic=1.02) and the four-factor alpha is 0.02% (t-statistic=0.20) for equal-weighted portfolios. For value-weighted portfolios, these numbers are 0.14% (t-statistic=0.91) and 0.10% (t-statistic=0.75), respectively.

month, respectively, for the period from January 2004 to January 2013. This finding is consistent with market participants becoming more adapt at processing complex information about firms' interconnections in a timely fashion.

To illustrate the profitability of a trading strategy based on the linked-stock signal, we plot, in Figure 3, the cumulative return of an investment of one dollar over the period from July 1996 to January 2013. A portfolio with returns equal those of the equal-weighted long-short portfolio would have been worth \$5.38; for the value-weighted long-short portfolio, this number would have been \$4.65. For reference, a buy-and-hold investor in the S&P 500 index would have accumulated a profit of only one dollar over the same period.

B. Alternative specifications and robustness checks

In the following, we test whether the return predictability is sensitive to variations in the methodology used to form portfolios and to identify linked stocks. The results for various alternative specifications are reported in Table III. As a first robustness check, we modify the weighting scheme used for the calculation of linked-stock signals. In Panel A, signaling stocks' returns are weighted by the number of common news stories that linked stock pairs in the preceding three months, and, in Panel B, linked-stock returns are value-weighted using the stocks' market capitalization in month $t-2$. Weighting linked-stock signals by the number of common stories produces substantially lower and less significant excess returns than equal-weighting them, as in the baseline specification of Table II. Frequent co-mentions should increase market participants' awareness of the economic linkage between these stocks, and, hence, information should be transmitted more quickly between the stocks. Similarly, for value-weighted portfolios, return differentials significantly decrease relative to the baseline specification when linked stocks' returns are value-weighted, implying that signals from large stocks get impounded more quickly than with a one-month delay.

In Panels C and D, we slightly change the identification of linked stocks and of leaders and followers. Specifically, in Panel C, we no longer discard news stories we believe to cover competing firms. As expected, return differentials for both weighting schemes are smaller than in the baseline specification. The reason is that a news shock to a firm could have the opposite-sign effect on the competitor's return. In Panel D, we no longer impose the turnover condition to identify leaders and followers between linked stocks. As a result, the return differentials are reduced, especially for value-weighted portfolios. Hence, our turnover restrictions indeed help properly identify the direction of the information flow between linked firms.

In Panels E and F, we broaden our sample of common news stories. Instead of identifying firm linkages exclusively from news stories that co-mention exactly 2 firms, in Panel E, we allow all news stories that co-mention up to 5 firms, and in Panel F, up to 10 firms. Because it is more likely that stocks are co-mentioned by coincidence rather than because they are truly linked, the return differentials are lower than in the baseline specification. This is especially the case for equal-weighted portfolios and for the more extreme Panel F. Since the news media devote less space and effort to smaller firms, it is likely that some smaller firms now appear in the news simply by coincidence. These results confirm the previously expressed concern that news stories that cover a larger set of stocks convey less valuable information about linkages between stocks than those that focus exclusively on one stock pair.

Next, we increase the length of our identification window. Instead of a 3-month window of the baseline specification, we use a 6-month window in Panel G and a 12-month window in Panel H (as before, we end the widow five trading days before the month-end). While the 6-month window works slightly better than the 3-month window, the 12-month window works worse. An identification window that is too short will miss some important linkages. But when the identification window is too long, the linkages may have either already dissolved or become well known to market participants, ensuring a quicker information transmission.

In the last four panels, Panels L - O, we introduce restrictions on the stocks eligible to be included in the set of linked stocks that is used to calculate the linked-stock signal. It is now important to have a longer identification window since a 3-month window of the baseline specification will cause a great reduction in the sample size. Therefore, we also consider the results with a 6-month identification window. In Panels L and M, we limit the set of linked stocks used to compute the signal to those that belong to a different industry. The portfolio return differentials noticeable decrease noticeably, which indicates that the news media is particularly successful in identifying linked stocks that belong to the same industry, even within our rather granular industry classifications, and that inferring economic linkages from stories that connect stocks from different industries produces more noisy results. In Panels N and O, the set of signaling stocks is limited to those that are smaller than the stock receiving the signal. Although the observed return differentials for value-weighted portfolios decrease relative to the baseline specification, they are closer to the baseline results for equal-weighted portfolios. Hence, in contrast to the findings in the lead-lag literature that large stocks lead the returns of smaller stocks, we show that, in our setting, small stocks can lead returns of larger stocks.

C. The content of common news stories

In this section, we check whether additional types of economic linkages, other than those previously documented in the literature (e.g., customer-supplier relations), give rise to the cross-predictability in stock returns. We classify news into topics by using key words for news headlines and topic codes assigned by Thompson-Reuters. The algorithm to classify news stories into topic codes, which is described in detail in Appendix A1, consists of three steps. In the first step, we parse the headlines in our news dataset of common news stories into words and rank these words, after removing articles, preposition, and conjunctions, by the number of headlines in which they appear. In the second step, we assign each word to one of 14 news topics. In the last step, based on these mappings and on topic codes assigned by Thompson-Reuters, we classify news stories into news topics.

All news categories are made to be mutually exclusive, and topic order in Appendix A1 represents the hierarchy of the categories. If a news story can be classified into more than one topic, it is classified into the topic closer to the top of the list. Moreover, if a firm pair was co-mentioned in more than one news story during the identification period, it is classified into the topic closer to the top of the list. The order of the topics was made based on how much each topic was studied in the prior literature and on how narrowly defined it can be. Following our approach, we are able to classify over 60% of the news stories in our sample (that is, the sample of news stories that cover exactly two firms that are non-competitors).

Legal is the largest topic category, with 11.09% of news stories falling into that topic. Also common are stories about natural resources, geopolitical developments and regulation, these categories containing 10.01%, 9.33%, and 7.61% of the news stories in our sample. Of note, news stories about customer-supplier relationships, which have been investigated extensively in previous literature, are not prevalent in our dataset, representing only about 4.41% of our news sample.

In order to show that firm linkages other than customer-supplier relations can also give rise to the cross-predictability in returns, in this section, we firm pairs that were assigned to the "Supply chain" topic at any time over the trailing 3-year window. Finally, in order to ensure that we do not inadvertently miss customer-suppliers links between firm pairs, we discard all unclassified news stories.

Since we have to discard many news stories, which greatly reduces our sample, we increase the identification window to six months. Panels A - C, of Table IV report portfolio returns when nature of firm linkages is limited as described in Panel headings. In Panel A, we only keep all firm pairs that remain after the restrictions described above are implemented (specifically, all the news that co-mention the firm pairs are assigned to a defined topic and the firm pair had no customer-supplier ties in the past three years). The number of stocks in the portfolio decreases by roughly 11% relative to that of Panel G of Table III (there are about 90 stocks per portfolio in that specification). Though the return differentials are a little lower than in that table, they are economically meaningful and statistically significant. These results indicate that economic linkages other than customer-supplier linkages can also lead to the cross-predictability in returns.

Next, we investigate narrower categories of firm linkages. To ensure a sufficient number of observations, in some specifications, we group several news topics into broader categories and perform the portfolio analysis for these broader topics. In Panel B, we group together firm pairs belonging in the category "Partnerships" and "M&A activity." These categories fit together because they indicate similar business fortunes. The number of stocks per portfolio declines by about 80% relative to that of Panel G of Table III. The return differentials are not very statistically significant, in part because of the higher return volatility.

In Panel C, we only consider only firm pairs belonging to the category "Legal." This is the most populous category, and we do not combine it with other categories. The number of stocks per portfolio is about 79% lower than in Panel G of Table III. The equal-weighted return differentials are substantially larger in magnitude than in Panel G of Table III, however, the statistical significance of these return differentials is lower because of the high return volatility due to small sample size. The value-weighted return differentials are about 65% in magnitude of the value-weighted return differentials in Panel G of Table III but are not statistically significant. These results suggest that legal similarities induce significant cross-predictability in returns.

Panel D, we group together the remaining topic categories and classify firm linkages under the common heading "Operational similarities." The return differentials are now significant for the equal-weighted but insignificant for the value-weighted portfolios.

All told, these results show that firm linkages that lead to cross-sectional return predictability are best uncovered by investigating a wide variety of news. That is, information diffuses slowly not only between stocks linked through previously documented types of linkages, such as linkages of firms along the supply chain, but also between stocks linked by similar legal exposures, business and operational similarities, and so on. Since we are able to produce significant return differentials focusing deliberately on news about fundamentals, it appears that the lead-lag effect that we document is driven by fundamental business ties between firms as opposed to common investor sentiment.

D. Placebo tests

In order to rule out the possibility that an omitted stock characteristic drives the cross-predictability of stock returns, we conduct a placebo test of the cross-predictability of returns. Since Panel G of Table I shows that stocks that are mentioned in a common news story tend to be co-mentioned in the subsequent five years, we check whether the cross-predictability in return existed *before* rather than after a stock pair was first co-mentioned in the news, speculating that the the cross-predictability in returns should not yet exist at that time. Since it may take some time for journalists to notice and report on a firm linkage after it has already occurred, we move each linked stock pair far enough back. Specifically, we move a stock pair three years back from their first co-mention in the news. If both stocks in he linked pair traded at that time, we proceed to compute linked-stock signals and form portfolios exactly how we did in the main specification of Table II.

As we expected, the portfolio return differentials are no longer significant. Specifically, the five-factor alpha of the equal-weighted portfolio is equal -0.21% (with the t-statistic=-0.66), and the five-factor alpha of the value-weighted portfolio is equal -0.14% (with the t-statistic=-0.34). The placebo results, therefore, confirm our conjecture that it is the soft information embedded in the news co-mentions rather than some stock-specific characteristics that explain why linked stocks cross-predict each other's returns.

E. Cross-sectional regressions

In this section, we run a number of Fama and MacBeth (1973) cross-sectional regressions to provide further evidence of the previously documented predictive ability of the linked-stock signal. The regression setting allows us to control for other factors that are known to forecast returns. We show that these control variables do not subsume the linked-stock signal's predictive ability, which confirms that we have identified an independent source of return predicability.[16]

The results of the cross-sectional regressions are reported in Table V. In all but the first specification, we include, in addition to the linked-stock signal, the stock's lagged return and the lagged return of the stock's industry as cross-sectional return predictors. In all specifications but the third, we also include size, book-to-market, and the stock return over the past six months, ending one month prior, from $t-6$ to $t-1$. Specifications (7)-(9), in addition, contain interaction terms between the linked-stock signal and additional control variables meant to capture the attention of sophisticated market players; these regression specifications also include as controls the variable included in the interaction term.

The regression coefficient on the linked-stock signal is highly significant in all specifications. The estimated coefficient ranges from 0.018 to 0.036 in magnitude, somewhat lower than the coefficients on the lagged industry return. However, as in the portfolio specifications, the linked-stock signal does not include the last five trading days in the month t, but the lagged industry return does. Despite this disadvantage, the coefficients on the linked-stock signal are more statistically significant than the coefficients on the lagged industry return or

[16]The control variables are described in detail in Appendix A2.

the lagged own return in all specifications. The significance level never falls below the 1% p-value.

In specification (7), we interact the linked-stock signal with a dummy that equals one if institutional ownership of a stock is above the median value for the cross-section of stocks at the most recent quarter-end, which is a direct measure of the level of *sophisticated* investor attention. The results indicate that high levels of institutional ownership are associated with weaker price reactions in month $t+1$ to the linked-stock signal in month t, suggesting perhaps that institutional ownership helps ensure that prices incorporate relevant information without a delay of one month.

In specifications (8) and (9), we use less direct proxies for the level of *sophisticated* investor attention, that is, analyst coverage and size, and interact these proxies with the linked-stock signal. As before, we include dummy variables indicating if these variable are above the median in the cross-section in a given month. The interactions between the linked-stock signal and the analyst coverage dummy is insignificant, although the interaction term is of the expected sign. The interaction between the linked-stock signal and the size dummy is statistically significant at the 10% level.

Overall, the results confirm the robustness of the linked-stock signal as a return predictor and show that the slow reaction is likely due to slow processing of complex information, as the presence of institutional investors appears to speed up information transmission.

F. Break-even trading costs

Thus far, we have abstracted from trading costs. Even though we have documented significantly positive monthly excess returns, high trading costs associated with long-short portfolios may render these simple trading strategies unprofitable. In the following, we estimate break-even trading costs that would set the post-transaction-cost trading profits equal to zero. We assume that trading costs are incurred when a stock is bought or sold and that these costs are identical across stocks and independent of the amount traded (though in reality trading costs are lower for more liquid stocks and increase with the number of shares traded). Table A4 in the Appendix shows that portfolio assignments exhibit slight persistence over a one-month period; hence, portfolio turnover is slightly lower than 100% in any given trading period. All else equal, this introduces a slight advantage for value-weighted trading strategies since they do not incur rebalancing costs. However, value-weighted returns are typically lower than equal-weighted returns. On net, our estimate of break-even trading costs, shown below, is higher for value-weighted portfolios.[17]

Our estimates of break-even trading costs are expressed in units of return, i.e., as a percentage cost per dollar of a stock traded. In order to conserve space, we estimate break-even trading costs only for the baseline trading strategies, in which stocks are held in portfolios for one period, linked-stock signals are computed in the previous period, and linked-stock signals are equal-weighted.

We obtain the following estimates of break-even trading costs. For the baseline trading strategy, corresponding to Table II, our estimates of break-even trading costs are 0.22% for equal-weighted portfolios and 0.20% for value-weighted portfolios.To put these estimates in perspective, using the TAQ dataset for the period from January 1983 to August 2001, Sadka and Scherbina (2007) estimate an average effective spread of 0.25% for a typical stock and a typical trade, which is a somewhat higher than our estimates of break-even trading costs. Hedge funds are more skilled at minimizing trading costs than an average trader in the TAQ dataset, and their trading costs may easily fall below our estimated break-even values.

The relatively low estimates of break-even trading costs indicate that the strategies designed to profit from linked-stock news can support only small investment amounts since large amounts would entail large price impacts. Therefore, the profits that one could realize from trading on this strategy are rather low, which explains why the strategy's on-paper

[17] In reality, value-weighted strategies, of course, involve lower trading costs because larger amounts are traded in large stocks, which are typically more liquid.

profitability has persisted through the years. These results suggest that in order to trade on the linked-stock signals, one has to be skilled at minimizing trading costs. For an average investor, prices appear to lie within no-arbitrage bounds around the fair value imposed by trading costs. Yet, even if trading on linked-stock signals is difficult in practice, these signals should still be valuable inputs in portfolio allocation decisions.

G. Firms connected by second-tier links

It is possible that firms may cross-predict each other's stock returns even though they are not directly linked via joint news stories (e.g., firms A and E in Figure 1–Firm B is linked to firm A and firm E is linked to firm B, but firms A and E are not directly linked). Whether or not it is useful to collect information on such second-tier linkages for the purpose of return predictions is an empirical question. It is quite possible that the information from second-tier links is muted enough to not sufficiently contribute to the return predictability.

In order to check whether stocks connected through second-tier linkages still cross-predict each other's returns, we identify all stock pairs that are connected through second-tier linkages during the identification window (e.g., connected through having common first-tier linked stocks) but that, at the same time, are not directly connected. We form a predictive signal for each stock by calculating equal-weighted returns from all second-tier linkages. We form portfolios based on these signals as in the baseline specification, but we in addition try 6- and 12-month identification windows.

The results are reported in Table VI. The number of stocks in the second-tier linked portfolios is about one-third larger than that for the portfolios constructed based on first-tier links. This is because there are more second-tier links than first-tier links, and more linked stocks will meet the criteria to be included in the set of signaling stocks (return leaders). The tables show that the information from the second-tier linked stocks seems to be important smaller stocks. For larger stocks, there is less useful information available from second-tier linkages; for these stocks the news coverage is more comprehensive and stock connections that are sufficiently strong to influence each other's returns would be reflected in the existence of joint news stories. These results indicate that the news coverage of smaller stocks is sparse and that additional useful information on firm linkages can be collected by including the links of links.

V. Conclusion

This paper documents that economically-linked stocks have the ability to predict each other's returns. And, while in practice it may be challenging to uncover economic linkages between firms, we show that such linkages can be identified through media coverage. Specifically, we argue that a pair of stocks is economically linked if the two stocks appeared together in a news story over a trailing identification window. Portfolio sorts based on the linked-stock signal, computed as the average return of all stocks linked to a given stock, produce statistically significant and economically meaningful excess returns. We also show that the predictive ability of the linked-stock signal is robust to various firm- and industry-level controls. Moreover, the attention of sophisticated investors and frequent co-mentions in the news both reduce the cross-predictability in returns. Therefore, the slow information diffusion is likely due to the slow processing of complex information and to limited investor attention.

Our methodology allows us to identify lead-lag relationships without relying on ex-ante stock characteristics. As a result, we are able to identify return leaders that are smaller than their followers and to detect short-lived lead-lag relationships. Moreover, our approach allows for within-industry bets. Such bets are generally precluded in the lead-lag literature as the signals in that literature tend to be correlated within industries.

The additional advantage of identifying firm linkages from co-mentions in the news is that one can study the content of common news stories to learn about the nature of firm linkages. We show that existence of a variety of linkages between firms. Importantly, not only customer-supplier links, but also other types of links can give rise to the cross-predictability of stock returns. Legal similarities seem to be a particularly important type of linkage in our sample.

Our estimations show that it is difficult to earn significant profits from trading on the lead-lag effect in linked-stocks' returns documented here. Therefore, while the market is not perfectly efficient in processing all relevant information, it may be approximately efficient with prices lying within trading-cost bounds around the fair value.

Finally, our results could help address the R^2 puzzle, which states that firm-specific news explains only a small portion of stocks' idiosyncratic volatility. Our results suggest that the set of news days on which one should expect to see large idiosyncratic movements should also include the news days of economically-linked firms.

References

Anton, Miguel, and Christopher Polk, 2014, Connected stocks, *Journal of Finance* 69, 1099–1127.

Badrinath, S. G., Jayant R. Kale, and Thomas H. Noe, 1995, Of shepherds, sheep, and the cross-autocorrelations in equity returns, *Review of Financial Studies* 8, 401–30.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan, and Matthew Richardson, 2013, Which news moves stock prices? a textual analysis, Working paper.

Box, Travis, 2014, Stock price comovement and the market for information, Working paper.

Brennan, Michael J, Narasimhan Jegadeesh, and Bhaskaran Swaminathan, 1993, Investment analysis and the adjustment of stock prices to common information, *Review of Financial Studies* 6, 799–824.

Burt, Aaron, and Christopher Hrdlicka, 2015, Correcting the bias in network-based tests of price discovery, Working paper.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Chordia, Tarun, and Bhaskaran Swaminathan, 2000, Trading volume and cross-autocorrelations in stock returns, *Journal of Finance* 55, 913–935.

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

Cohen, Lauren, and Dong Lou, 2012, Complicated firms, *Journal of Financial Economics* 104, 383–400.

DeMiguel, Victor, Francisco J. Nogales, and Raman Uppal, 2014, Stock return serial dependence and out-of-sample portfolio performance, *Review of Financial Studies* 27, 1031–1073.

Dougal, Casey, Joseph Engelberg, Diego Garcia, and Christopher Parsons, 2012, Journalists and the stock market, *Review of Financial Studies* 25, 639–679.

Engelberg, Joseph, and Christopher Parsons, 2011, The causal impact of media in financial markets, *Journal of Finance* 66, 67–97.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 46, 427–466.

——— , 1993, Common risk factors in the returns of stocks and bonds, *Journal of Financial Economics* 33, 3–56.

——— , 2000, Characteristics, covariances, and average returns: 1929-1997, *Journal of Finance* 55, 389–406.

Fama, Eugene F., and James MacBeth, 1973, Risk, return and equilibrium: Empirical tests, *Journal of Political Economy* 51, 55–84.

Fang, Lily, and Joel Peress, 2009, Media coverage and the cross-section of stock returns, *Journal of Finance* 64, 2023–2052.

Hoberg, Gerard, and Gordon M. Phillips, 2012, Industry choice and product language, Working paper.

Hou, Kewei, 2007, Industry information diffusion and the lead-lag effect in stock returns, *Review of Financial Studies* 20, 1113–1138.

Israelsen, Ryan D., 2013, Does common analyst coverage explain excess comovement?, Working paper.

Lo, Andrew W ., 2004, The adaptive markets hypothesis, *Journal of Portfolio Management* 30, 15–29.

Lo, Andrew W, and A Craig MacKinlay, 1990, When are contrarian profits due to stock market overreaction?, *Review of Financial Studies* 3, 175–205.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *Journal of Finance* 66, 35–65.

Menzly, Lior, and Oguzhan Ozbas, 2010, Market segmentation and cross-predictability of returns, *Journal of Finance* 65, 1555–1580.

Muslu, Volkan, Michael Rebello, and Yexiao Xu, 2014, Sell-side analyst research and stock comovement, *Journal of Accounting Research* 52, 911–954.

Neuhierl, Andreas, Anna Scherbina, and Bernd Schlusche, 2013, Market reaction to corporate press releases, *Journal of Financial and Quantitative Analysis* 48, 1207–1240.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive-definite, heteroscedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Pástor, Ľuboš, and Robert F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642–685.

Peress, Joel, 2008, Media coverage and investors attention to earnings announcements, Working paper.

——— , 2014, The media and the diffusion of information in financial markets: Evidence from newspaper strikes, *Journal of Finance* 69, 2007–2043.

Roll, Richard, 1988, R^2, *Journal of Finance* 43, 541–566.

Scherbina, Anna, and Bernd Schlusche, 2013, Cross-firm information flows and the predictability of stock returns, Working paper.

Shumway, Tyler, 1997, The delisting bias in CRSP data, *Journal of Finance* 52, 327–340.

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

——— , Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437–1467.

Appendix

A1. Classifying news into topics

We classify news into topics by utilizing key words in headlines and topic codes assigned to news stories by Thompson-Reuters. Here, we work with the news sample used in our baseline analysis. Hence, in addition to removing news stories that may co-mention firms by chance, we only consider only news stories that mention exactly two firms and remove stories for which we consider the firms to be competitors.

The algorithm to classify news stories into topic codes consists of three steps: The first step is the parsing of headlines into key words and ranking these words. To that end, we parse all headlines into words and discard articles, preposition, and conjunctions from the set of words. We then rank the remaining words by the frequency of appearance in *unique* headlines and keep all words that appeared in at least 100 headlines. Of note, the set of words includes variations in the spelling of words, as well as words that share a common root with words in the set. As a reference, the word “update,” the most frequently used word across headlines, appears in 97,276 headlines and the word “layoffs” appears in 100 headlines and is therefore among the least frequently used words we consider.

In the second step, we do our best to assign each word to one of eleven news topics, the selection of which is guided by the topic list proposed by Neuhierl, Scherbina, and Schlusche (2013). For example, we deem the word “update” impossible to be assigned to a topic code, but assign the word “layoff” to the topic “labor, production and infrastructure.” Moreover, we augment our list of words which determine the topic code by the key words used in Neuhierl, Scherbina, and Schlusche (2013) for the relevant topics; for the “legal” topic we augment our word list with key words from the online legal word list, having removed finance terms (Loughran and McDonald (2011)).[18]

[18] `http://www.enchantedlearning.com/wordlist/legal.shtml`

In the last step, based on these mappings and on topic codes assigned by Thompson-Reuters, we classify news stories into eleven news topics. Table A3 contains a comprehensive list of key words and TRNA topic codes used to classify headlines into various news topics. The assignment of news stories into news topics is unambiguous in that the assignment occurs in the classification order listed below. Specifically, if a headline is not classified into the topic "Supply chain," we check whether the contained key words and the TRNA topic code meet the requirements for being classified into the topic "Legal;" if the headline is not classified into the topic "Legal," we proceed and check whether it fits into the topic "Partnership," and so on. If a headline cannot be classified, it falls in the category "Other." The classifications are as follows (percentages of news stories classified in each category are in parentheses):

1. **Supply chain (4.41%).** News involving customers and suppliers.
   EXAMPLE: Headline: "ExxonMobil says it is the exclusive supplier for 33 Caterpillar(reg) lubricant" (Firms: Exxon Mobil and Caterpillar; PNAC: n0322NATN1; Date 22MAR1997).

2. **Partnerships (3.66%).** News about corporate partnerships, strategic alliances, and licensing deals.
   EXAMPLE: Headline: "Unilever says 50-50 venture with Pepsico <PEP.N> to expand Lipton Tea distribution" (Firms: Unilever and Pepsico; PNAC: nAAT003618; Date: 30JUL2004).

3. **M&A activity (6.89%).** News about mergers, acquisitions, spin-offs, carve-outs, etc., as well as ongoing negotiations and speculations for potential M&A activity.
   EXAMPLE: Headline: "Aegon <AEGN.AS> says in talks to buy Global Preferred's stake in Bermuda unit" (Firms: Aegon and Aerogen; PNAC: nAAT003618; Date: 30JUL2004).

4. **Legal (11.09%).** News relating to legal issues relating to accounting fraud, labor, production and environmental issues, class action lawsuits, issues with the SEC, criminal investigations, etc.
   EXAMPLE: Headline: "FOCUS-Agribusiness put on defensive by legal threat" (Firms: E.I. DuPont de Nemours & Co. and Astrazeneca Plc; PNAC: n5439663; Date: 13SEP1999).

5. **Regulation (7.61%).** News about government and agency regulations, as well as political actions affecting corporations.
   EXAMPLE: Headline: "U.S. Senate panel targets offshore profits and taxes" (Firms: Microsoft and Hewlett Packard; PNAC: nL1E8KIGMB; Date: 18SEP2012).

6. **Labor, production, and infrastructure (2.89%).** News about the firm's labor force, products, operations, and infrastructure.

EXAMPLE: Headline: “Kiev local favouritism seen spurring Motorola exit” (Firms: Motorola Solutions and Royal KPN; PNAC: nFLLB41CON; Date: 01APR1997).

7. **Executive compensation and corporate governance (0.01%).** News about executive compensation and firms’ corporate governance.
   EXAMPLE: Headline: “WPP <WPP.L> shareholders back executive bonus plan” (Firms: Omnicom Group and Interpublic Group Cos; PNAC: nLAT001047; Date: 02SEP1999).

8. **Management news (0.05%).** News about changes in top management: promotions, retirements, firings, managers changing firms.
   EXAMPLE: Headline: “H&R Block announces management change” (Firms: H&R Block and Fiserv; PNAC: nWEN3804; Date: 07NOV2005).

9. **Common customer (0.32%).** News about common customers, typically, U.S. federal and state governments and the military and other countries’ government projects and militaries.
   EXAMPLE: Headline: “GE and Juniper Networks to develop family of rugged, secure network appliances for military vehicles <GE.N><JNPR.O>” (Firms: General Electric and Juniper Networks; PNAC: nMKW18241a; Date: 07NOV2011).

10. **Financing (0.69%).** News about financing, cross-investments, firms leasing each other’s assets and extending loans to each other.
    EXAMPLE: Headline: “TowerJazz and GE Capital sign definitive asset based loan agreement to provide up to 4 billion yen credit line (approximately $50 million)” (Firms: General Electric and Tower Semiconductor Ltd; PNAC: nWNAB9038; Date: 09MAY2012).

11. **Natural resources (10.01%).** News that mention raw materials used in production inputs, as well as natural and environmental disasters affecting firms’ operations.
    EXAMPLE: Headline: “RPT-India Panna-Mukta fields, shut by explosion, likely to resume oil output by next week - sources” (Firms: Bunge Ltd and InterOil Corp; PNAC: nDEL001551; Date: 12JUN2008).

12. **Energy (2.54%).** News that mention energy inputs.
    EXAMPLE: Headline: “Gas strike threatens to shut Sri Lanka industry” (Firms: Royal Dutch Petroleum and Sheldahl Co; PNAC: nCOL001356; Date: 07MAY1997).

13. **Technology (2.96%).** News that mention various production/operation technologies.
    EXAMPLE: Headline: “ATM security flaws could be a jackpot for hackers” (Firms: Diebold Inc and NCR Corp; PNAC: nN25138100; Date: 25JUN2010).

14. **Geopolitical (9.33%).** News about firms’ foreign operations and geopolitical events, regional conflicts, sovereign policies, etc., affecting these operations.
    EXAMPLE: Headline: “FEATURE - Unrest grows in Nigerian oilfields” (Firms: Royal Dutch Petroleum and Sheldahl Co; PNAC: nLA57708; Date: 07OCT1998).

15. **Other (37.54%).** News that could not be classified into any of the above categories.

A2. Variable definitions and estimations

This appendix provides detailed descriptions of the variables used in our cross-sectional regressions. Unless specified otherwise, all variables are calculated at the month end.

Previous month's industry return *(Ind.* Ret_{t-1}*)* is defined as the value-weighted industry return over the previous month.

Size *(Size).* A stock's size is defined as the product of the price per share and the number of shares outstanding, expressed in thousands of dollars and measured at the end of the previous month.

Book-to-market ratio *(Book/Market).* Following Fama and French (1992, 1993, 2000), the book-to-market equity ratio is computed at the end of June of each year as the book value of stockholders' equity, plus deferred taxes and investment tax credit (if available), minus the book value of preferred stock, scaled by the market value of equity. Depending on availability, we use the redemption, liquidation, or par value (in that order) to estimate the book value of preferred stock for the last fiscal-year end. The market value of equity is the product of share price and the number of shares outstanding at the end of December of the previous fiscal year.

Previous month's return *(*Ret_{t-1}*).* This short-term reversal predictor is defined as the stock return over the previous month.

Turnover *(Turn)* is the average monthly turnover, computed over the past 12 months, and measured at the end of the previous month.

Institutional Ownership *(Inst. Ownership)* is defined as the percentage of total shares outstanding owned by institutions, computed using the data in the Institutional Holdings (13F) dataset and measured at the end of the previous month.

Analyst Coverage *(Analyst Cov.)* is defined as the number of analysts issuing annual earnings forecasts for the current fiscal year, computed using the I/B/E/S dataset and measured at the end of the previous month.

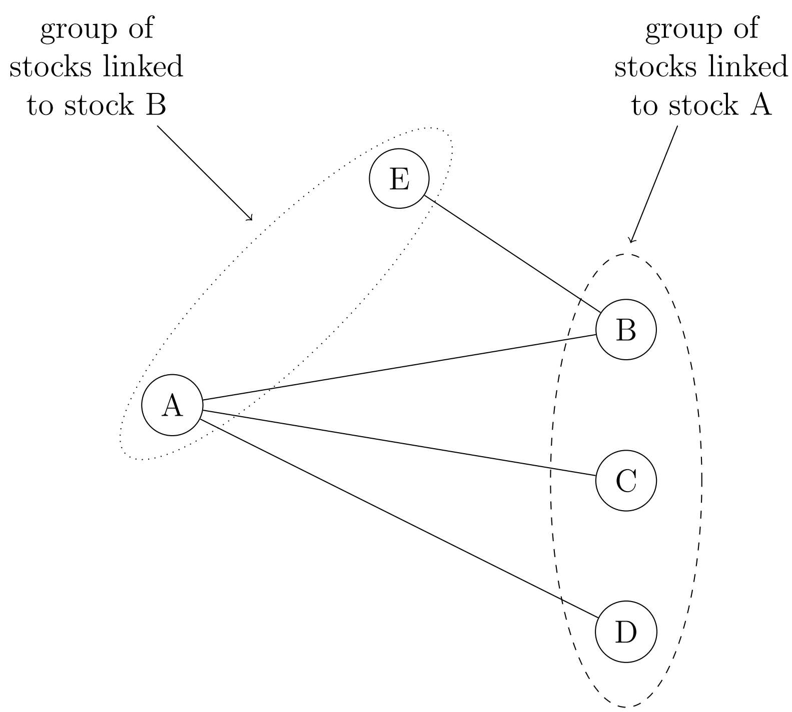

Figure 1. Examples of Linked Stocks. This figure gives two examples of groups of linked stocks and illustrates the computation of linked-stock signals. The dashed ellipse contains the stocks whose returns are used to compute the linked-stock signal for stock A, and the dotted ellipse, the stocks whose returns are used to compute the signal for stock B.

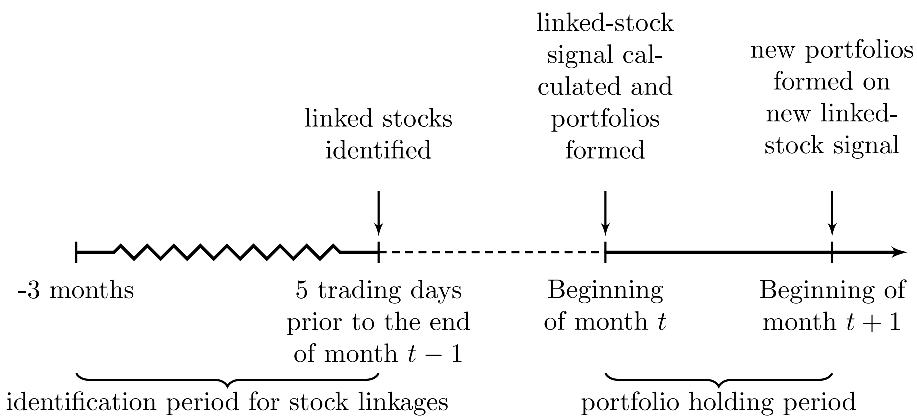

Figure 2. Timeline for Analysis. This figure presents the timeline for the identification of stock linkages, the calculation of predictive return signals, and the formation of portfolios. The zigzag line indicates a long period of time and the dashed line a short period of time relative to the period between months t and $t+1$.

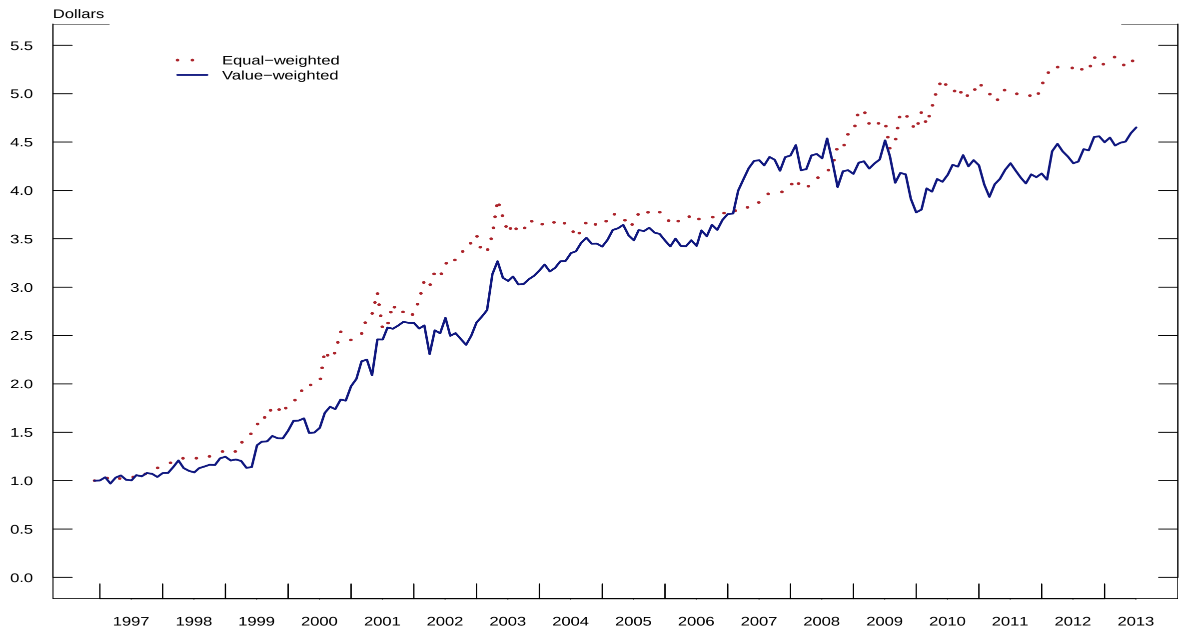

Figure 3. Cumulative returns. The charts plot, for equal- and value-weighted portfolios, the value of $1 invested in the beginning of the period at the return earned on a zero-investment strategy of buying stocks in the top and selling short stocks in the bottom linked-stock-signal quintile. Linked stocks are identified as those co-mentioned in news stories and linked-stock signals are calculated as the prior month's equal-weighted average return of all stocks linked to a particular stock. Quintile portfolios are formed monthly. The time period is July 1996 - January 2013.

Table I
Descriptive statistics

This table provides descriptive statistics for the TRNA dataset and for the news sample used in the main tests of the paper. It also presents various descriptions of the stocks linked by common news stories. The CRSP universe comprises stocks with share codes 10 or 11. In Panels B through G, the statistics are calculated based on Sample 3 of Panel A. Panel G presents probabilities that a stock pair classified as linked by common news in week τ (a stock pair is classified as linked if (1) there was a common news story written about exclusively these two stocks in the previous 3 months and (2) the stocks are identified as non-competitors) will remain linked in week $\tau + t$, given that both stocks are still present in the CRSP dataset at that time. The sample period is July 1996–December 2012.

Panel A: The news dataset

Total number of unique news stories	5,455,605
# of stories remaining after stories about trade order imbalances, index changes, bond ratings news, analyst recommendation revisions, etc. removed	3,689,918
Of these:	
# of stories that mention more than one firm (Sample 1)	521,845
# of stories that mention exactly 2 firms (Sample 2)	331,232
Of the stories that mention exactly 2 firms:	
# of stories that are about non-competitors (Sample 3)	299,060

Panel B: Statistics on the number of firms being mentioned per story (based on Sample 1)

mean	median	75th percentile	95th percentile
2.78	2	3	5

Panel C: The distribution of the number of common news stories for linked pairs (computed in non-overlapping windows)

no. of common news	% of total 3-month window	% of total 6-month window
1	64.48%	62.20%
2	16.27%	16.29%
3	6.47%	6.51%
4	3.38%	3.67%
5	2.09%	2.31%
>5	7.30%	9.03%

Panel D: Fraction of linked firm pairs in the same industry

Industries defined with 38 industry classifications	40.86%
Industries defined with 12 industry classifications	49.06%

Panel E: Linked vs. unlinked firm-months in the CRSP universe

	Identification window used			
	3 months		6months	
	linked	unlinked	linked	unlinked
Fraction of CRSP universe	10.51%	89.49%	15.12%	84.88%
Average market capitalization ($, million)	13,776.48	1,119.51	10,869.02	944.07
Average CRSP size decile	7.88	5.21	7.68	5.10

Panel F: The average number of stock linkages by CRSP size decile)

CRSP size	Identification window used	
decile	3 months	6 months
1	0.07	0.13
2	0.08	0.15
3	0.10	0.19
4	0.13	0.24
5	0.18	0.32
6	0.22	0.38
7	0.27	0.48
8	0.34	0.63
9	0.56	1.02
10	2.47	4.32

Panel G: Persistence of linkages between stock pairs

Number of months in the future (t)	All stock pairs		Stock pairs with at least two common stories	
	common news count	prob.	common news count	prob.
3	2.44	27.10%	5.31	50.87%
6	2.43	24.57%	4.59	45.91%
9	2.44	23.34%	4.41	44.09%
12	2.45	22.86%	4.28	42.76%
...				
60	2.37	18.43%	5.02	33.83%

Table II
Abnormal portfolio returns based on the baseline linked-stock signal within 36 industries

This table presents monthly abnormal returns of portfolios sorted based on equal-weighted linked-stock signals. Each month, all stocks that were linked to other stocks are sorted into quintile portfolios within each of the 36 industries. Linked stocks are identified over the trailing window that starts three months and ends five trading days before the month in which portfolio returns are calculated. Portfolio returns are equal-weighted in Panel A and value-weighted in Panel B. The second column contains the average linked-stock signal. The third column shows the number of linked stocks to which an average stock in a particular quintile is linked. The fourth column reports the average portfolio return in excess of the risk-free rate; the fifth column reports the market alpha; the sixth column reports the alpha of the Fama and French (1993) three-factor model; the seventh column reports the alpha of the four-factor model that also includes the Carhart (1997) momentum factor; and the eighth column reports the alpha of the five-factor model that in addition includes the liquidity factor of Pástor and Stambaugh (2003). The last row shows the return differential between the highest- and lowest-signal portfolios. Newey-West-adjusted t-statistics are reported in parentheses.

Panel A: Equal-weighted portfolios

Quintile	Linked-stock signal	Number of linked stocks	Excess return	Market alpha	3-factor alpha	4-factor alpha	5-factor alpha
1	-12.93%	1.45	0.13%	-0.42%	-0.53%	-0.45%	-0.46%
			(0.27)	(-2.31)	(-3.21)	(-2.76)	(-2.74)
2	-4.16%	1.95	0.46%	-0.06%	-0.20%	-0.11%	-0.11%
			(1.01)	(-0.44)	(-1.71)	(-0.98)	(-0.95)
3	0.36%	2.21	0.48%	-0.03%	-0.15%	-0.08%	-0.09%
			(1.07)	(-0.18)	(-1.18)	(-0.65)	(-0.67)
4	4.97%	2.04	0.59%	0.03%	-0.07%	-0.03%	-0.02%
			(1.31)	(0.22)	(-0.47)	(-0.22)	(-0.15)
5	16.29%	1.50	1.01%	0.45%	0.31%	0.35%	0.36%
			(2.14)	(2.42)	(2.26)	(2.50)	(2.51)
5-1			0.88%	0.87%	0.84%	0.80%	0.82%
			(4.65)	(4.57)	(4.06)	(3.87)	(3.82)

Panel B: Value-weighted portfolios

Quintile	Linked-stock signal	Number of linked stocks	Excess return	Market alpha	3-factor alpha	4-factor alpha	5-factor alpha
1	-11.11%	2.62	0.03%	-0.43%	-0.42%	-0.42%	-0.42%
			(0.06)	(-2.07)	(-2.02)	(-1.99)	(-2.00)
2	-3.38%	4.59	0.49%	0.02%	-0.01%	0.01%	-0.01%
			(1.16)	(0.12)	(-0.04)	(0.06)	(-0.07)
3	0.46%	5.54	0.54%	0.06%	0.14%	0.14%	0.11%
			(1.28)	(0.38)	(0.88)	(0.86)	(0.71)
4	4.49%	4.96	0.43%	-0.04%	0.03%	0.01%	0.02%
			(1.05)	(-0.21)	(0.21)	(0.07)	(0.11)
5	13.86%	2.97	0.87%	0.39%	0.40%	0.37%	0.36%
			(1.95)	(2.25)	(2.30)	(2.22)	(2.19)
5-1			0.84%	0.82%	0.81%	0.79%	0.78%
			(3.57)	(3.53)	(3.41)	(3.31)	(3.27)

Panel C: Subperiod analysis

July 1996–December 2003 subperiod				
	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	-0.13%	-0.74%	-0.12%	-0.54%
	(-0.19)	(-2.37)	(-0.16)	(-1.30)
...				
5	1.40%	0.71%	1.33%	0.65%
	(2.15)	(2.51)	(2.18)	(2.27)
5-1	1.52%	1.45%	1.44%	1.19%
	(4.94)	(3.55)	(3.57)	(2.25)

January 2004–January 2013 subperiod				
	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.29%	-0.30%	0.13%	-0.32%
	(0.42)	(-2.32)	(0.20)	(-1.69)
...				
5	0.66%	0.05%	0.49%	0.05%
	(0.98)	(0.41)	(0.76)	(0.27)
5-1	0.37%	0.35%	0.36%	0.37%
	(2.66)	(2.37)	(1.66)	(1.72)

Table III
Alternative specifications and robustness checks

This table presents monthly abnormal returns of portfolios sorted based on linked-stock signals. In the baseline specification, signals from linked stocks are equal-weighted and portfolios are formed within 36 industries. Variations to this baseline specification are described in each panel heading. Each panel reports excess returns and five-factor alphas for equal- and value-weighted portfolios, as well as the return differentials between the highest- and lowest-signal portfolios in the last row. Newey-West-adjusted t-statistics are reported in parentheses.

Alternative signal weighting schemes

Panel A: Linked-stock returns are weighted by the number of common news stories in equation (1)

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.15%	-0.41%	0.27%	-0.16%
	(0.32)	(-2.64)	(0.54)	(-0.83)
...				
5	0.97%	0.33%	0.81%	0.27%
	(2.08)	(2.37)	(1.71)	(1.27)
5-1	0.82%	0.74%	0.53%	0.43%
	(4.55)	(3.76)	(1.96)	(1.57)

Panel B: Linked-stock returns are value-weighted by market capitalization in month $t-2$

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.17%	-0.42%	0.58%	0.10%
	(0.35)	(-2.79)	(1.22)	(0.52)
...				
5	1.04%	0.39%	1.05%	0.48%
	(2.24)	(2.91)	(2.35)	(2.18)
5-1	0.87%	0.81%	0.48%	0.38%
	(4.99)	(4.54)	(2.13)	(1.80)

Relax leader-follower selection restrictions

Panel C: Do not remove competitors

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.15%	-0.44%	0.34%	-0.10%
	(0.32)	(-2.78)	(0.72)	(-0.51)
...				
5	0.97%	0.31%	0.88%	0.41%
	(2.03)	(2.11)	(1.97)	(2.22)
5-1	0.82%	0.76%	0.54%	0.52%
	(3.91)	(3.35)	(2.05)	(1.76)

Panel D: Do not impose turnover conditions on leaders and followers

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.27%	-0.61%	0.68%	-0.00%
	(0.50)	(-4.14)	(1.41)	(-0.02)
...				
5	0.97%	0.21%	1.00%	0.45%
	(1.84)	(1.66)	(1.84)	(2.41)
5-1	0.70%	0.82%	0.33%	0.46%
	(3.36)	(4.27)	(1.09)	(1.73)

Identify stock linkages from news stories that co-mention more than 2 firms

Panel E: Connections are computed from news stories that co-mention up to 5 firms

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.38%	-0.22%	0.41%	-0.16%
	(0.81)	(-1.47)	(0.92)	(-0.93)
...				
5	0.80%	0.12%	0.90%	0.39%
	(1.70)	(0.96)	(2.30)	(3.00)
5-1	0.43%	0.34%	0.49%	0.55%
	(1.98)	(1.69)	(2.80)	(2.81)

Panel F: Linkages are computed from news stories that co-mention up to 10 firms

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.39%	-0.22%	0.34%	-0.21%
	(0.86)	(-1.38)	(0.80)	(-1.43)
...				
5	0.77%	0.10%	0.91%	0.40%
	(1.62)	(0.75)	(2.40)	(2.65)
5-1	0.38%	0.32%	0.57%	0.60%
	(1.79)	(1.50)	(3.09)	(2.96)

Increase the length of the trailing window to identify linked stocks

Panel G: Use the 6-month trailing window to identify stock linkages

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.25%	-0.34%	0.24%	-0.23%
	(0.53)	(-2.42)	(0.55)	(-1.37)
...				
5	1.07%	0.46%	1.06%	0.56%
	(2.36)	(3.05)	(2.42)	(3.25)
5-1	0.82%	0.80%	0.82%	0.79%
	(4.53)	(4.09)	(4.04)	(4.04)

Panel H: Use the 12-month trailing window to identify stock linkages

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.38%	-0.25%	0.46%	-0.08%
	(0.78)	(-2.10)	(1.09)	(-0.57)
...				
5	0.91%	0.29%	1.03%	0.51%
	(2.02)	(2.77)	(2.39)	(2.89)
5-1	0.53%	0.53%	0.57%	0.60%
	(3.67)	(3.70)	(2.62)	(2.88)

Signals are computed exclusively from linked stocks in a different industry

Panel L: 3-month trailing window is used to identify linkages

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.21%	-0.33%	0.55%	0.09%
	(0.39)	(-1.79)	(1.04)	(0.31)
...				
5	0.71%	0.08%	0.69%	0.26%
	(1.49)	(0.52)	(1.29)	(1.18)
5-1	0.50%	0.41%	0.14%	0.17%
	(2.37)	(1.84)	(0.46)	(0.53)

Panel M: 6-month trailing window is used to identify linkages

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.36%	-0.24%	0.65%	0.14%
	(0.71)	(-1.26)	(1.33)	(0.50)
...				
5	0.95%	0.39%	0.98%	0.52%
	(2.07)	(2.12)	(2.11)	(2.31)
5-1	0.59%	0.63%	0.33%	0.38%
	(2.69)	(2.56)	(1.26)	(1.35)

Signals are computed exclusively from smaller linked stocks

Panel N: 3-month trailing window is used to identify linkages

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.38%	-0.26%	0.52%	0.06%
	(0.76)	(-1.13)	(1.03)	(0.20)
...				
5	1.04%	0.40%	1.08%	0.59%
	(2.19)	(2.13)	(2.24)	(2.22)
5-1	0.66%	0.66%	0.56%	0.53%
	(2.47)	(2.30)	(1.88)	(1.56)

Panel O: 6-month trailing window is used to identify linkages

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.41%	-0.25%	0.40%	-0.17%
	(0.89)	(-1.42)	(0.91)	(-0.72)
...				
5	0.94%	0.38%	0.91%	0.38%
	(2.05)	(2.45)	(2.00)	(2.08)
5-1	0.52%	0.63%	0.51%	0.55%
	(2.31)	(2.61)	(2.16)	(1.94)

Table IV
Conditioning on the content of news stories connecting firm pairs, excluding firm pairs linked by customer-supplier ties

The set of linked stocks excludes all stock pairs that were linked by at least one story classified to be in the customer-supplier category in the prior three years. This table presents monthly abnormal returns of portfolios sorted based on linked-stock signals for different categories of news stories that are used to identify linked stocks. The classification method is described in the main text and in Appendix A1. Linked stocks are identified based on the news co-mentions in the trailing window that starts six months and ends five trading days before the start of the month in which portfolio returns are calculated. We require that any firm pair is classified into only one category and that none of these classifications fall into the excluded categories, as described in each panel header. Any unclassified common news stories are dropped from the analysis. We form quintile portfolios based on the previous month's linked-stock signal and compute portfolio excess returns, five-factor alphas, and return differentials between the extreme quintiles for equal- and value-weighted portfolios. Newey-West-adjusted t-statistics are reported in parentheses.

Panel A: All defined topics
(except for customer-supplier links)

	EW portfolios		VW portfolios	
Quintile	Excess return	4-factor alpha	Excess return	4-factor alpha
1	0.27%	-0.33%	0.28%	-0.22%
	(0.58)	(-2.38)	(0.66)	(-1.32)
...				
5	0.95%	0.33%	0.95%	0.45%
	(2.10)	(2.85)	(2.18)	(2.68)
5-1	0.69%	0.66%	0.67%	0.67%
	(4.33)	(3.97)	(3.34)	(3.24)

Panel B: Only partnership and M&A linkages

	EW portfolios		VW portfolios	
Quintile	Excess return	4-factor alpha	Excess return	4-factor alpha
1	0.24%	-0.25%	0.64%	0.35%
	(0.51)	(-0.92)	(1.51)	(1.09)
...				
5	0.87%	0.33%	0.94%	0.52%
	(2.10)	(1.61)	(1.75)	(1.52)
5-1	0.63%	0.58%	0.29%	0.17%
	(2.05)	(1.73)	(0.65)	(0.35)

Panel C: Only common legal issues†

	EW portfolios		VW portfolios	
Quintile	Excess return	4-factor alpha	Excess return	4-factor alpha
1	0.12%	-0.36%	0.44%	0.12%
	(0.22)	(-1.18)	(0.90)	(0.42)
...				
5	1.36%	0.97%	0.98%	0.63%
	(2.19)	(1.94)	(2.14)	(1.66)
5-1	1.23%	1.33%	0.54%	0.51%
	(2.00)	(2.08)	(1.00)	(0.94)

†Exclude firm pairs tied by partnership and M&A linkages.

Panel D: Only operational similarities‡

	EW portfolios		VW portfolios	
Quintile	Excess return	4-factor alpha	Excess return	4-factor alpha
1	0.29%	-0.29%	0.84%	0.42%
	(0.62)	(-1.93)	(2.04)	(1.84)
...				
5	0.99%	0.36%	1.08%	0.56%
	(2.29)	(2.60)	(2.51)	(2.93)
5-1	0.71%	0.66%	0.25%	0.14%
	(3.71)	(3.30)	(1.22)	(0.62)

‡All defined topics except partnership, M&A, and legal linkages.

Table V
Cross-sectional regressions

This table presents the results of Fama and MacBeth (1973) regressions of monthly stock returns in month t on a set of explanatory variables available as of the end of month $t-1$. The sample consists only of stocks that had linked stocks identified over a rolling window that starts in month $t-3$ and ends five trading days before the start of month t. Control variables include log(mktcap), book/market, and the average monthly stock return over the months $t-7$ to $t-2$, as well as the dummy variable included in the interaction term when interaction terms are included. All explanatory variables are described in Appendix A2. The median value of each variable is denoted as *med*. Newey-West-adjusted t-statistics are reported in parentheses. All coefficients are multiplied by 100. The sample period is October 1996 - January 2013.

Model	(1)	(2)	(3)	(4)	(5)[¶]	(6)[§]	(7)	(8)	(9)
Linked-Stock Signal$_{t-1}$	2.15^a	2.31^a	1.79^a	2.49^a	1.63^a	1.77^a	3.56^a	2.66^a	2.98^a
	(3.89)	(4.52)	(3.30)	(4.56)	(3.19)	(3.75)	(3.79)	(3.08)	(3.81)
Ret$_{t-1}$		-2.61^b	-2.43^b	-2.67^b	-1.91^a	-3.17^a	-2.47^b	-2.57^b	-2.63^b
		(-2.53)	(-2.14)	(-2.47)	(-2.65)	(-3.60)	(-2.24)	(-2.47)	(-2.52)
Industry Ret$_{t-1}$		4.83^c	6.39^b	3.91	5.93^a	4.31^c	5.08^c	4.54^c	4.91^c
		(1.85)	(2.54)	(1.46)	(2.62)	(1.85)	(1.93)	(1.71)	(1.87)
$\mathbb{1}\{Inst.Ownership > med\}\times$							-2.47^b		
Linked-Stock Signal$_{t-1}$							(-2.24)		
$\mathbb{1}\{AnalystCoverage > med\}\times$								-1.03	
Linked-Stock Signal$_{t-1}$								(-0.85)	
$\mathbb{1}\{Size > med\}\times$									-1.67^c
Linked-Stock Signal$_{t-1}$									(-1.75)
Controls	Yes	Yes	No	Yes[†]	Yes	Yes	Yes	Yes	Yes

^{a}, ^{b}, and ^{c} indicate significance at the 1%, 5%, and 10% levels, respectively.

[†]In addition to the controls in specifications labeled [†], controls include the signal from stocks that are identified as monthly return leaders in Scherbina and Schlusche (2013).

[¶]Do not limit the sample of followers based on their turnover at $t-1$.

[§]Identify linked stocks from a six-month rolling window.

Table VI

Portfolios constructed from signals of second-tier links

This table presents monthly abnormal returns of portfolios sorted based on the signals of stocks linked exclusively by second-tier links (or links of links). Specifically, we required that the stocks whose returns in month $t-1$ are included in the signal calculation are not linked to the follower directly by is linked to a stock that is co-mentioned with the follower stock. Moreover, we require that the stocks in the second tier of links are not directly linked to the follower during the identification period. If a stock A is linked to a stock B and stock B is linked to a stock C, and stock C is not directly linked to stock A, then stock C is a second-tier link for stock A. In Panels A, the identification window starts three months before, and in Panel B, the identification window starts six months before the beginning of month t, and both windows end 5 trading days before. The returns of the second tier of links are equal-weighted. In Panels B and D, they are weighted by the combined relevance score assigned to the stocks B and C in the corresponding news stories. Each panel reports excess returns and five-factor alphas for equal- and value-weighted portfolios, as well as the return differentials between the highest- and lowest-signal portfolios in the last row. Newey-West-adjusted t-statistics are reported in parentheses.

Panel A: Three-month identification window

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.28%	-0.34%	0.41%	-0.15%
	(0.61)	(-2.75)	(0.94)	(-0.84)
...				
5	0.66%	0.03%	0.69%	0.23%
	(1.37)	(0.25)	(1.57)	(1.30)
5-1	0.38%	0.36%	0.29%	0.38%
	(2.48)	(2.29)	(1.10)	(1.69)

Panel B: Six-month identification window

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.46%	-0.18%	0.42%	-0.19%
	(1.01)	(-1.71)	(0.98)	(-1.28)
...				
5	0.80%	0.15%	0.84%	0.32%
	(1.70)	(1.20)	(2.10)	(1.76)
5-1	0.33%	0.33%	0.43%	0.51%
	(2.15)	(2.10)	(1.79)	(2.24)

Panel C: 12-month identification window

	EW portfolios		VW portfolios	
Quintile	Excess return	5-factor alpha	Excess return	5-factor alpha
1	0.50%	-0.14%	0.59%	-0.02%
	(1.07)	(-1.43)	(1.37)	(-0.15)
...				
5	0.89%	0.24%	0.58%	0.06%
	(1.96)	(1.92)	(1.51)	(0.45)
5-1	0.39%	0.39%	-0.01%	0.08%
	(2.46)	(2.28)	(-0.05)	(0.42)

Table A1
Industries

This table presents the monthly average distribution of stocks across industries in our sample. The sample consists of common shares of U.S.-incorporated firms (stocks with share codes 10 or 11) that traded on the last day of the previous month and were priced above $5 per share in 2012 inflation-adjusted dollars. The averages are computed using only the months that have at least one stock observation in a given industry. The sample period is 1996–2012.

Stone, Clay and Glass Products	0.14%
Agriculture, Forestry, and Fishing	0.15%
Textile Mill Products	0.22%
Sanitary Services	0.26%
Lumber and Wood Products	0.28%
Tobacco Products	0.31%
Leather and Leader Products	0.38%
Furniture and Fixtures	0.49%
Glass	0.50%
Apparel and other Textile Products	0.66%
Paper and Allied Products	0.77%
Miscellaneous Manufacturing Industries	0.78%
Rubber and Miscellaneous Plastics Products	0.87%
Mining	0.94%
Construction	1.02%
Petroleum and Coal Products	1.09%
Fabricated Metal Products	1.16%
Public Administration	1.20%
Primary Metal Industries	1.55%
Printing and Publishing	1.70%
Telephone and Telegraph Communication	1.71%
Radio and Television Broadcasting	1.79%
Transportation Equipment	2.60%
Food and Kindred Products	2.81%
Wholesale	3.13%
Transportation	3.29%
Oil and Gas Extraction	3.38%
Instruments and Related Products	4.58%
Electric, Gas, and Water Supply	5.30%
Machinery, Except Electrical	5.80%
Retail Stores	7.01%
Electrical and Electronic Equipment	7.46%
Chemicals and Allied Products	8.72%
Finance, Insurance, and Real Estate	11.97%
Services	16.83%

Table A2
Factor loadings for baseline linked-stock signal portfolios

This table presents the five-factor model factor loadings for the baseline linked-stock portfolios. Each month, all stocks that were linked to other stocks are sorted into decile portfolios within each of the 36 industries. Panels A and B report the factor loadings for equal- and value-weighted portfolios, respectively, along with their abnormal returns and the corresponding R^2s. Newey-West-adjusted t-statistics are reported in parentheses.

Panel A: Factor loadings for equal-weighted portfolios

Decile	alpha	MKT	SMB	HML	UMD	LIQ	R^2
1	-0.46%	1.07	0.46	0.08	-0.11	-0.01	86.54%
	(-2.74)	(23.09)	(4.36)	(0.97)	(-1.67)	(-0.27)	
2	-0.11%	1.05	0.34	0.20	-0.13	0.01	91.88%
	(-0.95)	(27.26)	(10.39)	(4.86)	(-6.04)	(0.57)	
3	-0.09%	1.01	0.34	0.17	-0.10	-0.01	89.93%
	(-0.67)	(23.44)	(9.12)	(3.50)	(-3.29)	(-0.55)	
4	-0.02%	1.12	0.31	0.13	-0.05	0.03	89.09%
	(-0.15)	(38.44)	(6.93)	(2.87)	(-1.45)	(1.01)	
5	0.36%	1.09	0.49	0.17	-0.06	0.03	89.92%
	(2.51)	(32.52)	(10.06)	(3.25)	(-1.35)	(1.38)	
5-1	0.82%	0.02	0.03	0.09	0.05	0.04	-0.00%
	(3.82)	(0.34)	(0.20)	(0.79)	(0.56)	(1.18)	

Panel B: Factor loadings for value-weighted portfolios

Decile	alpha	MKT	SMB	HML	UMD	LIQ	R^2
1	-0.42%	0.98	0.02	-0.05	0.00	-0.02	68.61%
	(-2.00)	(19.42)	(0.26)	(-0.59)	(0.05)	(-0.35)	
2	-0.01%	1.05	-0.05	0.10	-0.03	-0.08	74.46%
	(-0.07)	(14.46)	(-0.54)	(1.26)	(-0.79)	(-2.02)	
3	0.11%	1.06	-0.20	-0.12	0.01	-0.07	82.43%
	(0.71)	(24.24)	(-3.94)	(-2.03)	(0.26)	(-2.06)	
4	0.02%	1.04	-0.26	-0.07	0.03	0.02	78.76%
	(0.11)	(23.02)	(-5.08)	(-1.41)	(0.63)	(0.56)	
5	0.36%	1.06	-0.03	0.00	0.04	-0.04	77.07%
	(2.19)	(24.58)	(-0.39)	(0.02)	(0.64)	(-1.47)	
5-1	0.78%	0.08	-0.04	0.05	0.03	-0.03	-1.57%
	(3.27)	(1.24)	(-0.52)	(0.49)	(0.64)	(-0.57)	

Table A3
Key words and topic codes for the classification of headlines

This table presents key words and the TRNA topic codes that are used to classify news headlines into topic codes. Key words and word combinations used are separated by commas. In the text, the words making up key word combinations can be separated by other words. In addition, the set of key words includes variations in the spelling of words as well as words that share a common root with words in the set.

Supply chain. Key words: customer, supplier, (agreement) to supply (sell to, buy from), supply to, buy from, supplier (customer) agreement (pact), multi-year agreement. TRNA topic codes: DEAL, DEAL1
Partnership. Key words: partner, joint, consortium, collaboration, contract, alliance, distribution, consortium, outsource, licensing, common project, team up, join forces, distribution agreement, licensing agreement, contractual agreement, sign (make, finalize, extend) deal (agreement, contract, partnership, collaboration), exclusive agreement, reach a deal, agree on a deal, reach agreement, work out deal, joint project, joint operations, craft agreement, finalize agreement, joint venture, form (create, start) venture (collaboration, joint project), joint agreement, mutual agreement, temporary agreement, strategic deal, strategic agreement, share profit, profit sharing, share revenue, revenue sharing. TRNA topic codes: ALLCE.
M&A activity. Key words: M&A, merger, acquisition, takeover, sell subsidiary, buyout, consolidation, combine, purchase, carve-out, spin-off, split-off, divestiture, bid on. TRNA topic codes: MGR, STK, DLTK, IPO, NG1, PS1, DVST.
Legal. Key words: scandal, patent, ruling, lawsuit, settlement, antitrust, probe, judge, infringe, fraud, bankruptcy, privatization, law, recall, arbitration, verdict, cartel, audit, misstatement, restatement, defect, copyright, compliance, fiduciary, probe, investigate, subpoena, court, trial, smuggle, justice, attorney, acquit, affidavit, justice, allegation, arrest, assault, bail, bailiff, bankruptcy, circumstantial evidence, crime, complainant, confess, constitution, contract, continuance, counsel, court, crime, cross-examination, custody, damages, decree, defendant, defense, deposition, disbarment, docket, due process, entrapment, escrow, ethics, evidence, examination, exonerate, expunge, felony, jury, grievance, guilty, habeas, corpus, hearing, hearsay, immunity, incarceration, incompetent, indictment, infraction, injunction, innocent, jail, judge, judiciary, jurisdiction, jurisprudence, justice, larceny, lawsuit, lawyer, legal, legislation, leniency, liable, lien, litigation, manslaughter, marshal, mediation, misdemeanor, mistrial, murder, negligence, oath, objection, ordinance, overrule, paralegal, pardon, parole, perjury, petition, plaintiff, plea, precedent, probable cause, hearing, prison, probate, probation, prosecute, redress, rejoinder resolution, search warrant, sentence, sequester, settlement, sheriff, statute, subpoena, judgment, suit, suppress, testimony, theft, tort, transcript, trial, trustee, usury, verdict, voir dire, waiver, witness, zoning, off-shore tax, SEC, DOJ. TRNA topic codes: CASE1, CLASS, MNGISS, MONOP, JUDIC, LAW, ACB, HRGT, SCAM1, FAKE1, JUDIC, FRAUD1, REGS, CRIM, BRIB, DAT, CIV, CLJ.
Regulation. Key words: regulation, senate, Obama, Clinton, Bush, Yellen, Bernanke, justice, CDC, GAAP, environmental, parliament, democrats, republicans, cabinet, treaty, commissioner, Rotterdam, referendum, FDIC, Pentagon, Homeland, House of representatives, government regulation, U.S. House, House bill, House majority, House minority, U.S. Senate, Senate majority, Senate minority. TRNA topic codes: POL, JOB, WASH, USDA, DEFOR, CEN, GFIN, HEA, WOM, CO2, AWLQ, PLCY, ENV, MCE, SDS, GFIN, FED, HREP, SEN, G20, G8, G7, MEVN.
Labor, production, and infrastructure. Key words: worker salary (pay, benefits), pay cut, layoff, wage increase, labor union, union negotiations, job (positions) cut (reduced, eliminated), factory (plant, store) shut (closed, eliminated), strike, production, product (child, consumer) safety, product defect (recall), USDA, FDA, subsidy, stimulus, manufacturing, development, infrastructure, launch, exit, downsize, revamp, expand, expansion plan, innovate, restructure, product launch, new product, new device, launch product, launch service, launch delivery, new service, plant/facility/operations shut down/closure, open/begin production; exclude "wage war." TRNA topic codes: RTM, WPAY, STAT, BKRT.
Executive compensation and corporate governance. Key words: executive(CEO) salary (bonus, pay, benefits, compensation, contract), corporate governance, governance issue (problem, failure), weak (poor) governance, poorly governed; exclude "governance services", "corporate governance firms."
Management developments. Key words: appoint, resign, demote, retire, election, management change (turnover), board turnover, change in management. TRNA topic codes: BOSS1.
Common customer. Key words: missile, military, prison, defense, olympics, bombardier, US Marines, US Navy, Air Force, contract with (sell (services) to, supply services to, customer agreement with) government (prison, jail). TRNA topic codes: DEF, DEFBUY.
Financing. Key words: lease, loan, lend, financing, cross-financing, cross-holdings, credit facility, subsidiary, to fund (provide) funding (credit, capital), credit line, to purchase stake, line of credit, secure credit, infuse (invest) capital (equity, cash, money); exclude "finance chief (head)." TRNA topic codes: LOA, SFIN.
Natural resources. Key words: names or metals and minerals used in production and manufacturing. TRNA topic codes: MIN, AGRI, ALU, AMCRU, ASCRU, ATMY, AUSCRU, BASMTL, BRGE, BRLY, BSMH, BUN, CANCRU, CBLT, CHR, CHS, CO2, COA, COC, COCOIL, COF, MIN, COFARA, COFROB, CONT, COR, COROIL, COT, CPPR, CRU, DAIR, DBULK, DIAM, DISTLL, EMACRU, FERR, FERT, GOL, GRA, H2O, HOIL, INDI, IRDM, IRN, JET, LATCRU, LEAD1, LITH, LIV, LNG, LPG1, MEAL MECRU, METL, MGS, MGSM, MINMTL, MLDM, MLK, MOG, NAP, NASCRU, NATU, NGL, NGS, NIOB, NKL, NRG, NSCRU, NSEA, NUC, OILS, OLVOIL, ORJ, PALL, PETC, PGM, PHOS, PLAS, PLAT, PNTOIL, POIL, POTH, PRCP, PREMTL, PROD PWR, RAPOIL, RAREE, RFO, RHDM, RHEN, RICE1, RLFT, RNW, RTNM, RUB, RUSCRU, SCRP, SEACRU, SFTS, SHFV, SLCN, SLK, SLVR, SNFOIL, SOIL, SORG, SOY1, SSTE, STE, SUG, TEA, TGSN, TIN1, TMBR, TMP, TNKR, TNTE, TTNM, URAN, USCRU, VNDM, WEA, WHT, WINE1, WND, WOO, ZNC, TWAVE, DFTS, QUAK.
Energy. Key words: pipeline, deepwater, sunpower, electricity, energy, gasoline, exploration, refinery, oil, gas, hydroelectric, solar, biofuel, ethanol. TRNA topic codes: DRIL, TRNSPT, OILG, OILI, EXPL, ENER, AFRCRU, BIOCEL, BIODSL, BIOETH, BIOF, BIOMS.
Technology. TRNA topic codes: LSCI, WWW, NSS, SPAC, GMO, ITEC, SCI.
Geopolitical. Key words: quake, tsunami, pandemic, forex, militant, province, cossack, Putin, Kremlin, unrest; names of world currencies; names of continents and large non-US geographical areas (e.g., Mediterranean), names of large islands and island chains, names of all world countries and their capitals, names of large non-US cities. TRNA topic codes: VIO, WAR, PIA, as well as topic codes for all countries and international regions.

Table A4
Portfolio transition probabilities

This table reports portfolio transition probabilities between two consecutive months, t and $t+1$, calculated only if the stock exists in the CRSP universe at both points in time. These transition probabilities correspond to Table II.

Portfolio in month t	Portfolio in month $t+1$					
	1	2	3	4	5	unassigned
1 (low signal)	0.14	0.13	0.11	0.12	0.13	0.38
2	0.10	0.15	0.15	0.15	0.11	0.33
3	0.09	0.15	0.17	0.16	0.10	0.33
4	0.09	0.15	0.16	0.15	0.10	0.34
5 (high signal)	0.11	0.12	0.13	0.13	0.12	0.38

Portfolio in month t	Portfolio in month $t+2$					
	1	2	3	4	5	unassigned
1 (low signal)	0.09	0.11	0.10	0.10	0.10	0.49
2	0.09	0.12	0.13	0.12	0.09	0.44
3	0.08	0.13	0.15	0.13	0.09	0.41
4	0.08	0.12	0.14	0.13	0.09	0.43
5 (high signal)	0.09	0.11	0.10	0.11	0.10	0.49