Journal of Banking and Finance 91 (2018) 49–69

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

# Qualitative similarity and stock price comovement

Travis Box[1]

*School of Business Administration, University of Mississippi, 340 Holman, P.O. Box 1848, University, MS 38677-1848, United States*

ARTICLE INFO

*Article history:*
Received 29 November 2016
Accepted 14 April 2018
Available online 21 April 2018

*JEL classification:*
C33
D81
D83
G12
G14

*Keywords:*
Information consumption
Information-driven comovement
Textual analysis
Correlation

ABSTRACT

I introduce a method for gauging the qualitative similarity of firm-specific information based on linguistic commonality in newswire text. I show that this new qualitative similarity measure predicts future cross-firm return correlation even after accounting for the pair's contemporaneous price comovement, common exposures to systematic risk, firm liquidity, price, index membership, text volume, headquarters location, product similarity, shared mutual fund or institutional ownership, common analyst following and newswire co-mentions. I also demonstrate that content produced solely by journalists cannot predict an economically meaningful portion of future comovement. Out-of-sample tests confirm that knowledge of qualitative similarity can also reduce portfolio risk.

The primary determinant of equity portfolio risk is the likelihood that pairs of stock prices will rise and fall together. Both in research and in practice, our expectations regarding future price comovement, and our appraisals of portfolio risk, have relied on lengthy series of historical stock returns. Yet, the mechanism generating these returns depends on a flow of firm-specific information that changes throughout time. As new sources of opportunity and uncertainty are revealed to the market, the links between distant historical prices and future stock price comovement become weaker. Therefore, accurate predictions of comovement must also consider the similarity of contemporaneous information flows across firms. I develop a proxy for this similarity and test whether this new measure can improve predictions of future stock price comovement.

The field of finance is replete with simple quantitative descriptors designed to identify similarities in firm characteristics. However, contemporaneous changes in the flow of information are not reflected in these quantitative measures until firms announce earnings or publish financial reports. To identify contemporaneous changes in firm similarity, investors must rely on softer, more qualitative, sources of information. In real-time, this content is often delivered to the market through financial newswires. These services act as information conduits by compressing a vast array of firm-specific material into a digestible sequence that investors can use to make portfolio decisions. This paper examines whether the qualitative information circulated on one such newswire, the Reuters Integrated Data Network, can predict how future equity payoffs are correlated across firms.

During each six-month period from 2003 to 2013, I measure the similarity of firm-specific newswire text written about different companies. I propose that the contemporaneous information flows for two firms are qualitatively similar if there is commonality in their newswire text. In support of this hypothesis, I find that the qualitative similarity of the newswire items written about a firm pair predicts their stock return correlation during the following six-month period. Furthermore, this new measure of qualitative similarity can predict future price comovement even after accounting for the pair's contemporaneous return correlation. Thus, qualitative similarity describes similarity in information flows that cannot be inferred from historical stock prices.

Prior literature has recognized that characterisitics such as firm beta (Ledoit and Wolf 2003), size (Pindyck and Rotemberg 1993), book-to-market (Bekaert et al., 2009), momentum (Asness et al., 2013) and industry (Campbell et al., 2001; Irvine and Pontiff, 2009;

*E-mail address:* tbox@bus.olemiss.edu

[1] I thank the University of Mississippi for their research support. I am grateful for the helpful comments of Eric Kelley, Volkan Muslu, Adam Kleinbaum, Andrew Lynch, Ryan Davis, Robert Van Ness, Seong Byun, Angela Box and all the participants at the Ole Miss Finance Department Seminar, the Midwest Finance Association Annual Meeting, and the Arrowstreet Capital, LP Practitioner Seminar. I thank Paul Tetlock from Columbia University and Richard Brown and Maciek Pomalecki from Thomson Reuters Machine Readable News for data considerations. All mistakes in this article are my own.

https://doi.org/10.1016/j.jbankfin.2018.04.010
0378-4266/© 2018 Elsevier B.V. All rights reserved.

Brandt et al., 2010) proxy for common sources of systematic variance that generate price comovement between firms. The literature also offers many alternative explanations for stock price comovement that are based on some type of market friction. Specifically, cross-sectional variation in information diffusion (Barberis et al., 2005), as well as the categorical trading of assets (Barberis and Schleifer 2003), have been shown to cause higher levels of stock price comovement. To ensure that qualitative similarity does not proxy for one of these other documented sources of return correlation, I show that my measure's predictability remains after controlling for similarities in exposure to systematic risk as well as firm liquidity, price, index membership, text volume and headquarters location. Thus, commonality in the information flow across firms predicts return correlation that cannot be accounted for with standard asset pricing models and alternative explanations for stock price comovement

The newswire text appearing on the Reuters IDN originates from a variety of sources and sources and spans a broad range of topics. To better understand how the contemporaneous flow of information predicts future stock price comovement, I divide my sample of newswire text along two dimensions. First, I consider whether the qualitative similarity of text produced by journalists is more or less informative than content generated by the firms themselves. Second, I determine if the relation of interest depends on whether the newswire content describes the financial results of the firm.

Most of the text circulated on the Reuters Integrated Data Network is generated by the firms themselves in the form of press releases and legal disclosures. However, I give special attention to content written by journalists because a great deal of prior literature focuses specifically on the role of text produced by the traditional press (see Barber and Loeffler, 1993; Tetlock, 2007; Fang and Peress, 2009; Tetlock et al., 2008; Tetlock, 2011; Peress, 2014). If, as suggested by Ahern and Sosyura (2014), journalist-produced content merely summarizes primary sources that are written by the firms, then a thorough examination of their output will not provide much in the way of meaningful insights. Accordingly, I find only weak evidence that the qualitative similarity of newswire text produced by the financial press can predict an economically meaningful portion of future cross-firm comovement.

Pindyck and Rotemberg (1993) propose that the comovements of individual stock prices should depend only on expectations about future earnings. However, when I divide my newswire sample by topic, I find that text related to corporate financial results is a weaker predictor of future return correlation. Nevertheless, the predictive performance of newswire content focused on results does not imply that these stories contain less information. If most of the information revealed in these newswire items is communicated through a numerical value, such as an earnings level, then the text accompanying this release might contain less important qualitative information. In either case, the most notable result from this analysis is that truncation, whether between journalist and firm or earnings and non-earnings, leads to a significant loss of qualitative information relative to the full sample.

Next, a series of closely related projects attempt to quantify qualitative information produced by either the firms themselves or by some other information producer. With an eye toward predicting return correlation, Israelsen (2015) and Muslu et al. (2014) study common analyst coverage and Anton and Polk (2014) look at shared ownership among actively managed mutual funds. While Hoberg and Phillips (2010a, 2010b) do not forecast stock price comovement directly, they propose a text-based measure of product differentiation that should be well suited to the task. Tetlock (2007) and Tetlock et al. (2008) show that tonal measures of news text, like pessimism, can predict stock prices and accounting earnings. Finally, Scherbina and Schlusche (2015) identify cross-firm predicatability in stock returns for companies that are mentioned together in certain types of news stories.

Additional tests confirm that newswire text from the Reuters Integrated Data Network contains at least some information about future return correlations that is orthogonal to the other sources of qualitative information highlighted by these related projects. These tests also demonstrate that the Hoberg and Phillips (2010a, 2010b) product similarity measures are strong predictors of future comovement. However, the results are less encouraging with regards to the remaining sources of qualitative information. First, I find little evidence that textual tone contributes positively to future return correlation. Furthermore, the variables measuring shared mutual fund or institutional ownership, common analyst following and newswire co-mentions appear to be correlated with persistent firm-pair panel effects. It is not surprising that such connections are persistent enough to be subsumed by panel effects if specific analysts and reporters follow, or institutions and mutual funds hold, firms with similar characteristics.

In recent years, another growing body of research has examined return predictability arising from interfirm linkages and investor inattention. Cohen and Frazzini (2008) propose that stock prices do not promptly incorporate news about economically related firms when investors are subject to information constraints. In support of their hypothesis, they find evidence of return predictability across groups of firms that are linked through customer-supplier relationships. Likewise, Menzly and Ozbas (2010) find that stocks in economically related supplier and customer industries cross-predict each other's returns. Also, Cohen and Lou (2012) posit that limited information processing capacity, not just inattention, can lead to a significant delay in the impounding of information into asset prices. They demonstrate that the returns of stand-alone firms predict the returns of more complex conglomerate firms that conduct some their business in the same industry. Finally, Cao et al. (2016) find evidence of return predictability between firms engaging in strategic alliances.

To confirm that qualitative similarity is not a proxy for these relationships, I perform my analysis on subsets of firms that are less likely to have identifiable economic linkages. Thus, I remove all observations for firm-pairs mentioned in the same newswire item and all firm-pairs that are in the same industry according to the Hoberg and Phillips (2010a, 2010b) product similarity measure. I also remove all firm-pairs with "second-tier" linkages whereby two firms are not linked directly but are linked through their respective direct linkages to some other firm. These filters remove combinations where the companies sell a similar product, announce a strategic partnership or have a public supply chain relationship at any point during my sample period. My results demonstrate that qualitative similarity is still able to predict future comovement between firm pairs that lack these types of direct or indirect economic linkages.

Finally, to evaluate the economic significance of the relation between qualitative similarity and future stock price comovement, I test whether forecasts of rolling correlations can reduce the out-of-sample volatility of an equity portfolio. In general, I find that portfolios based on forecasted correlations have dramatically lower standard deviations than passive strategies such as market- and equal-weighted portfolios. Furthermore, out-of-sample correlation forecasts benefit when qualitative similarity is included in the regression specification. Ultimately, my results indicate that investors may reduce the out-of-sample volatility their portfolios by incorporating the qualitative similarity of firm-specific information into their covariance predictions.

## 1. Newswire data and linguistic methodology

The firm universe for this study consists of all domestic common stocks trading on the NYSE, NASDAQ and Amex exchanges with CRSP share codes 10 or 11. I calculate the NYSE price and size decile breakboints each six-month period from January 2003 to December 2013 based on the price and shares outstanding for the final trading day of the previous interval. Firms falling in the smallest price or size decile for a particular time period are removed from the sample where the average lowest breakpoints across all intervals are $7.89 and $259 million, respectively. The resulting sample contains an average of 1982 firms at the beginning of each period with 2723 unique firms appearing in at least one interval.

### *1.1. Thomson Reuters Newsscope Archive*

The newswire text comes from the Thomson Reuters NewsScope Archive, a historical database of *Reuters News* and select third party content. The Archive is derived from the Reuters Integrated Data Network (IDN) newswire feed and consists of the message stream which communicates text to client workstations. Newswire stories are transmitted across the IDN in smaller pieces called "takes." Each observation in the archive represents a take, and multiple takes with a common id number can be combined to recreate a story. In addition to the raw story text, each observation contains a field listing all of the tickers for the firms mentioned in the take.

A variety of additional filters are necessary for the construction of an appropriate firm-specific text corpus. The process described above results in a collection of newswire stories that mention a firm from the universe at least once in the text. However, just because a firm is mentioned in a particular take does not mean that the majority of the text is relevant. Thomson Reuters also provides a related product known as News Analytics containing proprietary scores for, among other things, the relevance of a take to each of the firms mentioned. This relevance measure is a real valued number bound between 0 and 1 describing the applicability of the take to the firm in question. A take is only retained for a firm if the relevance score is at least 0.5.[1]

On average, the sample contains 639 takes, representing 513 unique stories and 396 unique firms, each trading day across all distributions. Fig. 1 graphs the number of takes, stories, and firms included in the sample each trading day for the year 2013, though all years have a similar pattern. The most obvious feature of the time series is the effect of earnings season on the flow of company-specific news, recognizable by the four distinct peaks throughout the year. This pattern implies that newswire content is likely to contain information about firm fundamentals.

### *1.2. Term-document matrix*

Overall, the textual analysis used for this study most closely resembles the techniques described in Hoberg and Phillips (2010a, 2010b). The basic object of my analysis is the term-document matrix, a mathematical representation of the frequency of terms that occur in a collection of documents. The intuition behind this methodology is as follows: if the frequency of words used in the takes about different firms is similar, then the qualitative information contained in those stories is also similar. As an example, if the takes about two firms use words like "interest," "debt," and "default," it may be the case that both firms are having some difficulty accessing capital. Even if these firms are in entirely different industries and have entirely different market capitalizations, a newswire subscriber might expect some covariance in their future payoffs relative to firms whose newswire text does not mention these words.

In a term-document matrix, columns correspond to the documents (firms) in the collection and rows correspond to the terms (words). For each six-month period, all takes related to a specific firm are aggregated into one master firm document. The frequencies with which terms appear in this document are recorded as integers in a firm's term-document vector. Combining these vectors for all sample firms produces the term-document matrix for the period.[2] The field of linguistics refers to this type of analysis, dissecting a document by examining only word frequencies, as the bag-of-words model (Bilisoly, 2008). Because any random permutation of the text produces the same frequencies as the original version, word order is irrelevant. While this permutation removes information from the text, it allows for a tractable comparison of the content related to different firms.[3]

To choose the appropriate formation period, term-document matrices are constructed using 1-, 3-, 6- and 12-month spans. Fig. 2 shows the number of firms that would be included in the sample if the formation period ended on the date described by the horizontal axis. For 1-month formation periods, the number of firms in the matrix is greatly affected by the earnings season. The figure implies that many firms are only mentioned on the newswire around the time of their earnings releases, so any formation period that did not span these events would have an excessively volatile sample size. The 3-, 6- and 12-month formation periods remove the effect of earnings season from the data, and the six-month interval is chosen to strike a balance that would allow for observing discrete changes in information flows while still including newswires pertaining to the broadest universe of firms.

### *1.3. Qualitative similarity*

The term-document matrix itself can be thought of as the raw quantitative data for my analysis. However, to compare the information flows across different firms, the similarity of their newswire content must be computed explicitly. Hoberg and Phillips (2010a, 2010b) construct a measure of document similarity that compares the occurrence of unique words between term-document vectors of firms $i$ and $j$. Following their methodology, the elements of the term-document vectors $\mathbf{f}_{it}$ and $\mathbf{f}_{jt}$ consist only of 1's and 0's to indicate whether or not a firm document contains a particular word. Thus, their measure of document similarity is

[1] Several other filters are applied to the newswire text. The news archive draws on stories written from all over the world in many different languages, but only stories written in English are retained. All of the stories related only to exchange order imbalances, identified in the News Analytics database with the genre type "IMBALANCE," are also filtered from the sample. The News Analytics database reports the number of linked articles in a particular time period in order to gauge the novelty of the content being reported. Takes having a linguistic fingerprint similar to any other newswire items appearing in the previous 12 h are omitted from the sample. I also remove takes for which the variable "more_news" takes on values 'M' or 'm' and for which the variable "update_sz" is greater than 8500.

[2] When constructing the term-document matrix, all letters are changed to lower case, summary information about the authors is removed, and all tickers and numbers are deleted. Punctuation is removed with the exception of dashes between words and apostrophes between conjunctions. This should preserve the appropriate interpretation for tokens like "on-the-run" and "aren't." Finally, the individual words in own firm names, as listed in the CRSP Names History file, are removed from each firm's document to avoid arbitrary associations that are only caused by these words.

[3] The raw term-document matrix may possess some undesirable qualities that hinder a comparison between firms based on information content. For example, function words like "that," "this" and "is" are frequent, but add little to the information content of the text. The most common method of dealing with these function words is by simply removing them with a stop list. The list used in this study is included in the PERL Lingua module available for download on CPAN. After the function words are removed, the term-document matrices contain an average of 52,487 rows, or unique words, each period when constructed using all attributions.

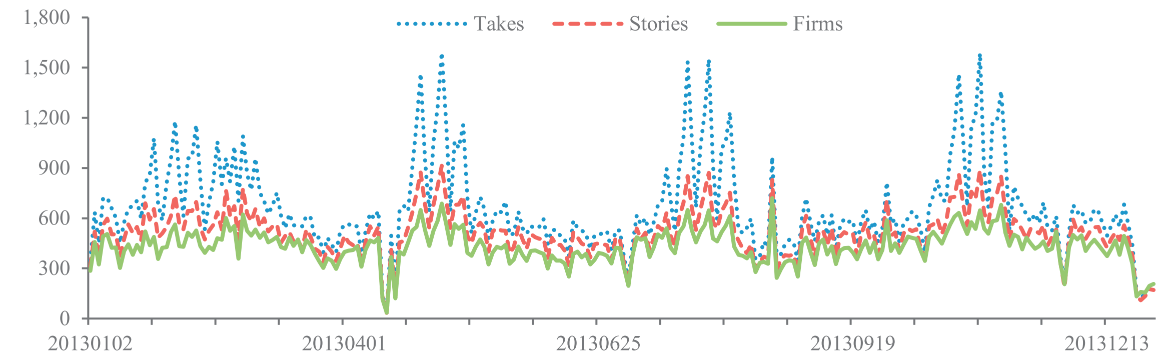

**Fig. 1.** Daily frequency of takes, stories, and firms during 2013. The daily frequency of individual news takes containing information relevant to a particular firm is pictured in blue. Multiple takes with the same matching identification numbers are used to form stories, and the daily frequency of unique stories is pictured in red. The number of individual firms mentioned in these stories each day is pictured in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

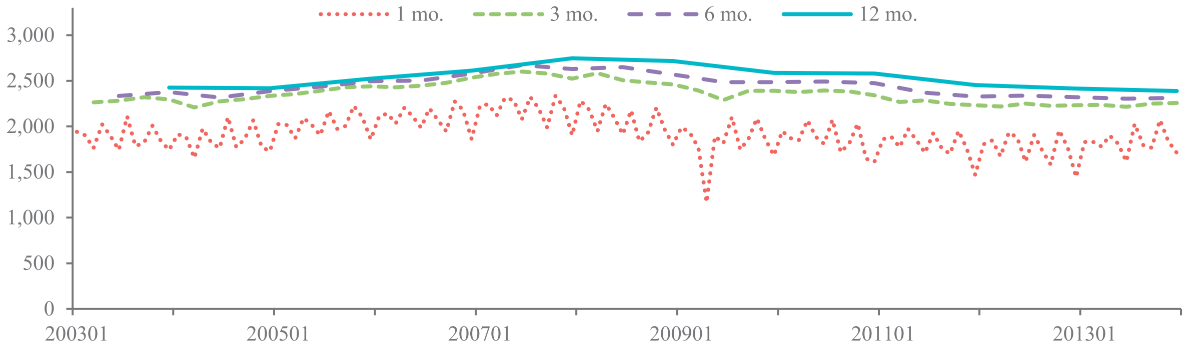

**Fig. 2.** Average document similarity variable over time. After compiling all relevant takes from the Thomson Reuters NewsScope Archive, the number of unique firms appearing in term-document matrices formed over 1-, 3-, 6-, and 12-month horizons from 2003 to 2013 are pictured below.

**Table 1**
Word count deciles and average document similarity. The document similarity variable $\widetilde{WireSim}_{ijt}$ is the cosine similarity between the firm vectors $i$ and $j$ in the term-document matrix for period $t$ constructed from text appearing on the Reuters Integrated Data Network. For each period in the sample, firms with some relevant text are classified into deciles based on total word counts. The variable $\widetilde{WireSim}_{ijt}$ represents the average document similarity between firms appearing in the same word count deciles as $i$ and $j$ during period $t$. .

| Decile | Lower word count ⟨—————————————————————⟩ Higher word count | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 0.17 | | | | | | | | | |
| 2 | 0.18 | 0.22 | | | | | | | | |
| 3 | 0.18 | 0.23 | 0.25 | | | | | | | |
| 4 | 0.17 | 0.23 | 0.25 | 0.26 | | | | | | |
| 5 | 0.17 | 0.23 | 0.26 | 0.27 | 0.28 | | | | | |
| 6 | 0.17 | 0.23 | 0.26 | 0.27 | 0.29 | 0.30 | | | | |
| 7 | 0.16 | 0.23 | 0.26 | 0.28 | 0.29 | 0.30 | 0.31 | | | |
| 8 | 0.16 | 0.22 | 0.26 | 0.28 | 0.29 | 0.31 | 0.32 | 0.34 | | |
| 9 | 0.15 | 0.22 | 0.25 | 0.27 | 0.29 | 0.31 | 0.33 | 0.35 | 0.37 | |
| 10 | 0.13 | 0.20 | 0.23 | 0.26 | 0.28 | 0.30 | 0.32 | 0.35 | 0.38 | 0.42 |

calculated as:

$$\widetilde{WireSim}_{ijt} = \cos\theta_{ijt} = \frac{\mathbf{f}_{it}^{\mathrm{T}}\mathbf{f}_{jt}}{|\mathbf{f}_{it}||\mathbf{f}_{jt}|} \tag{1}$$

The angle $\theta_{ijt}$, and thus the cosine of the angle, between the term-document vectors of two firms is greater when many of the same words appear in both vectors. If the text written about a pair of firms contains none of the same words, the pairwise cosine similarity $\widetilde{WireSim}_{ijt}$ will be 0. If both documents have identical word lists, the cosine similarity will be 1.

Table 1 demonstrates how document similarity changes in response to individual firm text volume. In each six-month span, firms with some positive quantity of text appearing on the IDN are divided into deciles based on total word counts. The variable $\widetilde{WireSim}_{ijt}$ represents the average document similarity between firms in the same text volume deciles as $i$ and $j$ during period $t$. Table 1 reports the time series average of $\widetilde{WireSim}_{ijt}$ over the entire sample period. Moving vertically along the columns, document similarity decreases as the gap between word counts grows larger. This suggests that the qualitative information about firms with low text volumes may be truly dissimilar from that of higher volume firms. However, there is also evidence that document similarity increases as text volume grows. Moving along the diagonal of the matrix in either panel, document similarity increases monotonically with word count decile.

Given this mechanical relation between document similarity and text volume, I propose the following adjustment that should focus the measure on similarities in underlying information:

$$WireSim_{ijt} = \widetilde{WireSim}_{ijt} - \overline{\widetilde{WireSim}}_{ijt} \tag{2}$$

This new variable $WireSim_{ijt}$ removes any patterns in document similarity $\widetilde{WireSim}_{ijt}$ that are only related to average word count $\overline{\widetilde{WireSim}}_{ijt}$. This volume-based portion of document similarity should contribute little to the information content of the text. Thus, $WireSim_{ijt}$ should provide a clearer description of how the flow of information related to a firm-pair is qualitatively similar or dissimilar.

## 2. Estimation methodology

Understanding how the prices of various financial securities evolve in relation to each other has long been a goal of asset pricing researchers and practitioners alike. From simple linear factor models to complex arbitrage strategies, security returns are commonly explained in the context of their comovement with other assets. While prior research documents the existence of persistent comovement in stock returns, it has provided little explanation for how the underlying cross-firm relationships evolve over time, and even less explanation on how these evolutions are discovered by market participants. Such insight is needed because even when historical patterns in comovement are identified, minor innovations in the origins of individual asset prices can transform the covariance structure of the entire market. I will examine how the qualitative similarity of newswire text written about firms $i$ and $j$ is related to their future Pearson return correlation $\rho_{ijt+1}$. If this relation is positive, qualitative similarity may help predict how the future payoffs of two firms are correlated.

Most of the subsequent analysis will center on the following basic regression model:

$$\rho_{ijt+1} = \beta_0 + \beta_1 WireDum_{ijt} + \beta_2 TakeSim_{ijt} + \beta_3 WireSim_{ijt} + \sum_{k=4}^{K} \beta_k Control_{kijt} + \varepsilon_{ijt+1} \tag{3}$$

where $WireDum_{ijt}$ is a binary variable indicating that both firms had some positive volume of text during period $t$. This variable is necessary to differentiate when qualitative similarity is 0 because information about the two firms was unrelated, or because one of the firms did not have a positive text volume during the period. The variable $TakeSim_{ijt}$ is defined as follows:

$$TakeSim_{ijt} = N_{ijt}^{take} / \sqrt{N_{it}^{take} N_{jt}^{take}} \tag{4}$$

where $N_{ijt}^{take}$ is the number of takes that mention both firms $i$ and $j$ in a period $t$, and $N_{it}^{take}$ and $N_{jt}^{take}$ are the number of takes mentioning firms $i$ and $j$, respectively. $TakeSim_{ijt}$ is included to account for situations where qualitative similarity is high because two firms are frequently mentioned in the same take.[4] If both firms are mentioned together in every take, $TakeSim_{ijt}$ will be 1, and if they are never mentioned in the same take, $TakeSim_{ijt}$ will be zero. The additional control variables $Control_{kijt}$ are discussed along with the presentation of my empirical results.

As written, the disturbances estimated from Eq. (3) contain some unfavorable structure. Like most panel datasets, all the observations occurring in time period $t+1$ should be related to each other because of immeasurable common factors generating their stock returns. Also, Eq. (3) attempts to measure the change in future return correlation that would result from a hypothetical change in contemporaneous qualitative similarity. It is possible that contemporaneous changes in qualitative similarity are responses to changes in return correlation earlier in the same period. Therefore, the specification should also account for the current period's, and possibly even earlier periods', observations of pairwise return correlation. Next, all the estimated return correlations have a value bound between $-1$ and 1, but the error term $\varepsilon_{ijt+1}$ is assumed to be distributed over a range of $-\infty$ to $\infty$. To improve the accuracy of the coefficient standard errors, the Fisher transformation is applied to the correlation estimates:

$$z_{ijt} = \frac{1}{2} \ln \frac{1+\rho_{ijt}}{1-\rho_{ijt}} \tag{5}$$

Taken together, these concerns motivate the following model with transformed and lagged dependent variables and time series fixed effects $\alpha_{t+1}$:

$$z_{ijt+1} = \sum_{s=0}^{S} \phi_s z_{ijt-s} + \beta_0 + \beta_1 WireDum_{ijt} + \beta_2 TakeSim_{ijt} + \beta_3 WireSim_{ijt} + \sum_{k=4}^{K} \beta_k Control_{kijt} + \alpha_{t+1} + \varepsilon_{ijt+1} \tag{6}$$

The transformed pairwise return correlation $z_{ijt+1}$ at time $t+1$ for firms $i$ and $j$ is also related to the transformed return correlation of the same firm-pair at all other points in time due to shared, but unobservable, characteristics. The cross-sectional disturbances are also likely to have structure induced by individual, but unobservable, firm characteristics. The addition of firm-pair and firm-specific panel effects to the specification should correct for the omitted variable bias associated with these relationships:

$$z_{ijt+1} = \sum_{s=0}^{S} \phi_s z_{ijt-s} + \beta_0 + \beta_1 WireDum_{ijt} + \beta_2 TakeSim_{ijt} + \beta_3 WireSim_{ijt} + \sum_{k=4}^{K} \beta_k Control_{kijt} + \alpha_{t+1} + \gamma_{i\wedge j} + \delta_{i\vee j} + \varepsilon_{ijt+1} \tag{7}$$

where $\gamma_{i\wedge j}$ is a panel effect for a unique pair of firms $i$ and $j$, and $\delta_{i\vee j}$ is a panel effect for each individual firm $i$ or $j$.[5] Unfortunately, OLS estimation of Eq. (7) would still be biased and inconsistent. Because the variables $z_{ijt+1}$ and $z_{ijt}$ would both be functions of the firm-pair, $\gamma_{i\wedge j}$, and firm-specific, $\delta_{i\vee j}$, panel effects, those parameters would be mechanically correlated with the disturbances. Therefore, I proceed with the dynamic panel estimator (henceforth DPE) proposed by Arellano and Bover (1995) and Blundell and Bond (1998).[6]

Not only can the approach described by Eqs. (6) and (7) help to identify the determinants of future return correlation, practioners should enjoy the limited data requirements neccesary to generate accurate predictions. To forecast the next period's return correlation between two firms, only a few years worth of return observations are required to generate reliable estimates. Thus, the expected correlation of a new firm or asset class could be included in the development of a trading strategy relatively quickly, instead of waiting several years or decades for the data neccesary to estimate a consistent sample covariance matrix (DeMiguel et al., 2009).

[4] Scherbina and Schlusche (2015) argue that economically linked stocks cross-predict each other's returns and that economic linkages can be identifed through media coverage.

[5] Box and Shang (2018) use a similar specification to measure the type of qualitative information that is consumed and incorporated into asset prices.

[6] Wintoki et al. (2012) and Box et al. (2018) use a similar dynamic panel estimator to mitigate endogeneity in an empirical corporate finance setting.

## 3. Predicting comovement

Pearson correlations $\rho_{ijt}$, and their Fisher transformations $z_{ijt}$, are calculated from daily and ten-day cumulative returns in excess of the risk-free rate for each six-month period in the sample; the first ending in June of 2003 and the last ending in June of 2014. Because Eqs. (6) and (7) contain lagged dependent variables, only firm-pairs with at least six consecutive return correlation observations are retained. The resulting sample contains 43,076,139 firm-pair-period observations that include 3,146,459 unique firm-pairs.

The sheer size of this panel makes the estimation of Eqs. (6) and (7) computationally infeasible. When estimating Eq. (6), subsequently referred to as the OLS approach, 1,500,000 firm-pairs are randomly selected from the initial universe of 3,146,459, with all of the time series observations from those firm-pairs included in the estimation. Some firm-pairs might only exist for a few periods in the beginning or end of the time series, and others might have usable observations over the entire sample period. This means that the number of eligible time series observations that a firm-pair may have does not affect the likelihood of its inclusion in the final sample, which ultimately contains 19,750,851 firm-pair-period observations.

When viewed in terms of individual firm prices and newswire content, this sampling methodology still makes use of all available firm-specific information on the newswire and in the CRSP price data. For the results reported below, the final OLS sample includes individual price and newswire text for all of 2723 firms that stay in the sample at least 6 periods. Thus, the final estimation includes firms of all different sizes, ages and, most importantly, newswire text volumes.

The computational demands of the Arellano and Bover (1995) and Blundell and Bond (1998) estimation procedure are much greater due to the dimensions of the instrument matrix required for efficient parameter estimation. For Eq. (7), 150,000 firm-pairs are randomly selected from the initial universe of 3,146,459. The resulting sample contains 1,367,394 firm-pair-period observations. As before, the sample used for the DPE still contains price and newswire text for all possible 2723 firms that have data available for at least 6 periods.

A series of related projects also study the determinants of return correlation, but use a sample of firm-pairs truncated by individual firm characteristics. Israelsen (2015) and Muslu et al. (2014) examine the effect of correlated analyst coverage on yearly stock-price comovement. Anton and Polk (2014) show that the degree of shared ownership among actively managed mutual funds forecasts cross-sectional variation in return covariance. In an effort to present results that are easily comparable with these existing studies, I perform my analysis on a truncated sample that only includes companies from the five largest NYSE size deciles. For this truncated panel, the lowest size breakpoint, averaged across all six-month periods, rises from $259 million to $2.7 billion, leaving a total of 7,373,461 firm-pair-period observations and 767,307 unique firm-pairs. This truncated sample contains an average of only 824 firms at the beginning of each quarter and 1355 unique firms over all time periods.

Based on Eqs. (6) and (7), I examine the degree to which commonality in the contemporaneous information flows of two firms predicts their future return correlation. Table 2 reports summary statistics for all of the regression variables included in estimates of Eqs. (6) and (7). Correlations are calculated from daily returns in Panel A of Table 2, whereas Panel B reports estimates that are based on ten-day cumulative returns. The average daily and ten-day cumulative return correlation $\rho_{ijt}$ across all firm-pairs and all six-month periods is roughly 28% and 26%, respectively, when the sample consists of the nine largest size deciles, and 31% and 29% when the sample contains only larger firms.

Across the broad sample, roughly 89% of the firm-pair-periods consist of two companies with some positive quantity of text broadcast over the IDN. When the universe is constrained only to larger firms, 95% of the firm-pairs consist of companies that both have positive text volumes. In either case, my measure of qualitative similarity, $WireSim_{ijt}$, can be calculated for most firm-pairs. Table 2 also demonstrates that firms are rarely mentioned in the same take. Across the 43,076,139 firm-pair-periods in my sample, $TakeSim_{ijt}$ is greater than zero for only 44,389 observations.

Previous research has identified a number of firm-specific characteristics that are associated with systematic comovement. To account for comovement that is related to firm size, market capitalizations are calculated on the final trading day of each six-month span, and firms are assigned to NYSE size deciles for the following period $t$. Following Muslu et al. (2014) and Chen et al. (2012), the dummy variable $SizeDum_{ijt}$ is included in my regression specifications to indicate whether or not two firms are in the same size decile. According to Table 2, only 11% of observations consist of firm-pairs where $SizeDum_{ijt} = 1$, so the limited scope of this variable may overlook some of the complexity in the market correlation structure. To control for comovement between firms of different sizes, I calculate the daily market-weighted average return for each NYSE size decile portfolio. The correlation, $SizeCorr_{ijt}$, between these portfolios is used to predict firm-pair return correlation in the following period. For instance, if firm $i$ is in NYSE size decile 3 and firm $j$ is in decile 7, the period $t$ correlation between the size portfolios 3 and 7 will be used to predict the correlation between firms $i$ and $j$ during the following period $t+1$.[7] Controls related to other firm characteristics, $BetaCorr_{ijt}$, $BetaDum_{ijt}$, $Bk/MktCorr_{ijt}$, $Bk/MktDum_{ijt}$, $MomCorr_{ijt}$, $MomDum_{ijt}$, $IndCorr_{ijt}$ and $IndDum_{ijt}$, are calculated using similar portfolio return correlations.

All specifications reported in Table 3 include untabulated fixed effects for each six-month period. The OLS specifications have standard errors clustered by firm-pair, both individual firms and time using the Cameron et al. (2011) multi-way clustering procedure. DPE results are generated by the two-step estimator with Windmeijer (2005) bias-corrected robust variance-covariance estimates of the model parameters. The second order test for serial correlation (p-values reported) was suggested by Arellano and Bond (1991) to detect any pattern in the differenced time series residuals of the individual cross-sections. Additional lagged dependent variables $\{z_{ijt-1},\ldots, z_{ijt-4}\}$ and lagged systematic variables, $\{BetaDum_{ijt-1},\ldots, BetaDum_{ijt-4}, BetaCorr_{ijt-1},\ldots, BetaCorr_{ijt-4}, \text{etc.}\}$ are included as untabulated controls in Eq. (7) to remove any evidence of serial correlation in the first-differenced residuals and validate the moment conditions of the DPE.

Table 3 reports estimates of Eq. (6) based on both the broad and truncated samples of firms. However, due to the lag structure in the DPE procedure, it is not possible to remove a firm for just one year because it temporarily falls below the sixth NYSE size decile break point. Many firms are not always large or always small, so the DPE can only be estimated on the broad sample where all firm-pair lags are available for inclusion in the instrument matrix.

Regardless of sample truncation or return frequency, the coefficient on my measure of qualitative similarity, $WireSim_{ijt}$, is positive and significant in Table 3 for both the OLS and the DPE methodologies. Thus, the similarity of newswire text between two firms can predict a significant portion of future price comovement even after controlling for contemporaneous return correlation. Furthermore, the significance of the coefficients in the presence of firm-

[7] The portfolio correlation variables are set to 0 whenever both firms are members of the same category. Thus, $SizeCorr_{ijt}$ will only have a value different from 0 whenever the value of $SizeDum_{ijt}$ is equal to 0, and vice versa. The former describes how market model betas influence return correlation between categories (dissimilar firms), and the latter describes correlation within categories (similar firms).

**Table 2**
Summary statistics for comovement prediction regressions.
This table presents summary statistics for the variables appearing in Eqs. (6) and (7) and estimated in Table 3. $\rho_{ijt}$ is the Pearson return correlation between firms $i$ and $j$ during period $t$, and $z_{ijt}$ is the Fisher transformation of $\rho_{ijt}$. Correlations are calculated from daily returns in Panel A, whereas Panel B reports summary statistics that are based on ten-day cumulative returns. The binary variable $WireDum_{ijt}$ is set to 1 whenever both firms have some positive number of total words transmitted across the Reuters Integrated Data Network. $TakeSim_{ijt}$ is equal to $N^{take}_{ijt}/\sqrt{N^{take}_{it}N^{take}_{jt}}$ where $N^{take}_{ij}$ is the number of takes that mention both firms in period $t$, and $N^{take}_{it}$ and $N^{take}_{jt}$ are the number of takes mentioning firms $i$ and $j$. Likewise, $\widetilde{WireSim}_{ijt}$ is a mesure of document similarity based on text transmitted across the Reuters Integrated Data Network. $WireSim_{ijt}$ is a measure of qualitative similarity defined in Eq. (2). A description for all other variable calculations is provided in the surrounding text and in Panel A of Table A-1.

| | Broad sample | | | | | | Larger firms | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | P1 | P5 | P50 | P95 | P99 | Mean | P1 | P5 | P50 | P95 | P99 |
| Panel A: Correlations from daily returns | | | | | | | | | | | | |
| $\rho_{ijt}$ | 0.283 | −0.07 | 0.02 | 0.27 | 0.59 | 0.72 | 0.307 | −0.07 | 0.03 | 0.29 | 0.63 | 0.75 |
| $z_{ijt}$ | 0.304 | −0.07 | 0.02 | 0.28 | 0.68 | 0.91 | 0.333 | −0.07 | 0.03 | 0.30 | 0.75 | 0.98 |
| $BetaDum_{ijt}$ | 0.106 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.109 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $BetaCorr_{ijt}$ | 0.733 | 0.00 | 0.00 | 0.86 | 0.97 | 0.98 | 0.727 | 0.00 | 0.00 | 0.85 | 0.97 | 0.98 |
| $SizeDum_{ijt}$ | 0.110 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.200 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $SizeCorr_{ijt}$ | 0.836 | 0.00 | 0.00 | 0.95 | 0.99 | 0.99 | 0.763 | 0.00 | 0.00 | 0.95 | 0.99 | 0.99 |
| $Bk/MktDum_{ijt}$ | 0.107 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.120 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $Bk/MktCorr_{ijt}$ | 0.796 | 0.00 | 0.00 | 0.90 | 0.97 | 0.98 | 0.787 | 0.00 | 0.00 | 0.90 | 0.97 | 0.98 |
| $MomDum_{ijt}$ | 0.103 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.107 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $MomCorr_{ijt}$ | 0.770 | 0.00 | 0.00 | 0.88 | 0.96 | 0.98 | 0.774 | 0.00 | 0.00 | 0.89 | 0.97 | 0.98 |
| $IndDum_{ijt}$ | 0.042 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.040 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| $IndCorr_{ijt}$ | 0.623 | 0.00 | 0.04 | 0.68 | 0.89 | 0.93 | 0.613 | 0.00 | 0.03 | 0.67 | 0.89 | 0.93 |
| $WireDum_{ijt}$ | 0.888 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.949 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| $TakeSim_{ijt}$ | 0.00005 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00018 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $\widetilde{WireSim}_{ijt}$ | 0.233 | 0.00 | 0.00 | 0.25 | 0.38 | 0.43 | 0.288 | 0.00 | 0.00 | 0.30 | 0.42 | 0.47 |
| $WireSim_{ijt}$ | 0.0033 | −0.091 | −0.058 | 0.000 | 0.067 | 0.111 | 0.0080 | −0.076 | −0.049 | 0.004 | 0.071 | 0.121 |
| Panel B: Correlations from 10-day cumulative returns | | | | | | | | | | | | |
| $\rho_{ijt}$ | 0.264 | −0.40 | −0.21 | 0.28 | 0.70 | 0.81 | 0.289 | −0.40 | −0.20 | 0.30 | 0.74 | 0.84 |
| $z_{ijt}$ | 0.298 | −0.42 | −0.21 | 0.28 | 0.86 | 1.13 | 0.332 | −0.42 | −0.21 | 0.31 | 0.94 | 1.23 |
| $BetaCorr_{ijt}$ | 0.705 | −0.13 | 0.00 | 0.84 | 0.97 | 0.99 | 0.691 | −0.13 | 0.00 | 0.82 | 0.97 | 0.99 |
| $SizeCorr_{ijt}$ | 0.830 | 0.00 | 0.00 | 0.95 | 0.99 | 1.00 | 0.761 | 0.00 | 0.00 | 0.96 | 0.99 | 1.00 |
| $Bk/MktCorr_{ijt}$ | 0.779 | 0.00 | 0.00 | 0.89 | 0.98 | 0.99 | 0.768 | 0.00 | 0.00 | 0.89 | 0.98 | 0.99 |
| $MomCorr_{ijt}$ | 0.728 | 0.00 | 0.00 | 0.84 | 0.97 | 0.99 | 0.734 | 0.00 | 0.00 | 0.85 | 0.97 | 0.99 |
| $IndCorr_{ijt}$ | 0.584 | −0.21 | 0.00 | 0.66 | 0.92 | 0.96 | 0.571 | −0.23 | 0.00 | 0.64 | 0.92 | 0.95 |

pair panel effects demonstrates that this method for quantifying the companies' qualitative information identifies a predictor of future stock price correlation that is not present in the historical time series of returns.

While it is not possible to calculate a reliable goodness of fit measure in the DPE specifications, the adjusted R-squared in the OLS results implies that roughly half (one-quarter) of the variance of future daily (ten-day cumulative) return correlation is accounted for with only contemporaneous observations of the dependent variable, the qualitative similarity of newswire text and the systematic control variables. When firm-pair panel effects and four additional systematic lags are added, the explanatory power is likely to be even higher. The large $t$-stats reported in Table 3 are consistent with the level of explanatory power and observation counts that reach into the millions.[8] The magnitude of the $p$-values from the second order tests for serial correlation are below 2 in all specifications, implying that there is no evidence of persistence in the differenced residuals.

Almost all of the included systematic controls have coefficients that are consistently positive and significant in Table 3. Most notable, in terms of magnitude, are the coefficients on variables related to size, in the OLS specifications, and book-to-market, in the DPE specifications. However, nothing predicts future daily return correlation better than contemporaneous daily return correlation. Thus, the pairwise associations observable in realized daily returns are still more useful predictors of comovement than qualitative similarity or any of the systematic forces commonly used to explain stock returns.

For most of the included variables, the magnitudes of the coefficients are smaller when the DPE is used instead of OLS. Unobserved heterogeneity that drives persistent stock price comovement between a firm-pair will be captured by the model's panel effects. The reduction in coefficient magnitudes reflect the degree of collinearity between the included regressors and these unobservable characteristics. Because most of the tabulated coefficients are changed by the inclusion of panel effects, it safe to assume that the firm-pair, $\gamma_{i\wedge j}$, and firm-specific, $\delta_{i\vee j}$, panel effects are not all equal to 0, and that Eq. (6) may be misspecified.

$TakeSim_{ijt}$ accounts for situations where qualitative similarity is only high because the same newswire takes contribute to the documents of different firms. If two companies are always mentioned in the same take, their term-document matrices will be identical because all the text written about them would be from the same sources. Though their returns may be highly correlated, any positive relation that is observed between their information flows and their future stock price comovement would not be useful for predicting a similar relation between firms that were never mentioned in the same take. The negative and significant coefficient on $TakeSim_{ijt}$ in the DPE specifications implies that the portion of $WireSim_{ijt}$ related to newswire co-mentions does not improve predictions of comovement after accounting for persistent unobservable heterogeneity between firms and firm-pairs.

Table 3 also describes the performance of the document similarity variable $\widetilde{WireSim}_{ijt}$ before subtracting off the average doc-

[8] Table 6 of Muslu et al. (2014) reports parameter estimates from similar OLS specifications that also have large t-stats, however, the explanatory power of their regressions is much lower. All of the OLS regression standard errors were estimated with the Kleinbaum et al. (2013) clus_nway Stata ado file. I would like to thank Volkan Muslu and Adam Kleinbaum for providing helpful comments on the OLS estimation procedure.

**Table 3**
Qualitative similarity and stock price comovement.
The dependent variable in all specifications is the Fisher transformation $z_{ijt+1}$ of the Pearson correlation $\rho_{ijt+1}$ calculated from the returns of firms $i$ and $j$ in excess of the risk-free rate for each six-month period $t+1$. Correlations are calculated from daily returns in Panel A, whereas Panel B reports estimates that are based on ten-day cumulative returns. The binary variable $WireDum_{ijt}$ is set to 1 whenever both firms have some positive number of total words transmitted across the Reuters Integrated Data Network. $TakeSim_{ijt}$, defined in Eq. (4), accounts for how often firm-pairs are mentioned in the same newswire take. Document similarity $\widetilde{WireSim}_{ijt}$ is based on text transmitted across the Reuters Integrated Data Network. $WireSim_{ijt}$ is a measure of qualitative similarity defined in Eq. (2). A description for all other included variable calculations is provided in the surrounding text and in Panel A of Table A-1. Eq. (6) is estimated with ordinary least squares and Eq. (7) is estimated with a dynamic panel estimation (DPE) methodology. Ordinary least squares standard errors are clustered by firm-pair, both individual firms and time using the Cameron et al., (2011) multi-way clustering procedure. DPE results are generated using the approach described in Arellano and Bover (1995) and Blundell and Bond (1998) with bias-corrected robust variance-covariance estimates of the model parameters. Coefficients marked * and ** are significant at the 5% and 1% level, respectively, and $t$-statistics are reported in parenthesis. All of the independent variables are used as predetermined instruments in the DPE specifications. "Systematic lags" refers to the total number of lags included in each specification for the variables $z_{ijt}$, $BetaDum_{ijt}$, $BetaCorr_{ijt}$, $SizeDum_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktDum_{ijt}$, $Bk/MktCorr_{ijt}$, $MomDum_{ijt}$, $MomCorr_{ijt}$, $IndDum_{ijt}$ and $IndCorr_{ijt}$.

| | Ordinary least squares | | | | | | Dynamic panel estimator | | |
|---|---|---|---|---|---|---|---|---|---|
| | Broad sample | | | Larger firms | | | Broad sample | | |
| Panel A: Correlations from daily returns | | | | | | | | | |
| $z_{ijt}$ | 0.411** | 0.410** | 0.408** | 0.456** | 0.455** | 0.453** | 0.232** | 0.230** | 0.231** |
| | (22.67) | (22.59) | (22.82) | (24.35) | (24.16) | (24.24) | (147.7) | (147.7) | (148.4) |
| $BetaDum_{ijt}$ | 0.0768** | 0.0765** | 0.0769** | 0.124** | 0.122** | 0.122** | 0.0319** | 0.0318** | 0.0313** |
| | (4.044) | (4.016) | (4.029) | (4.888) | (4.777) | (4.784) | (14.49) | (14.45) | (14.25) |
| $BetaCorr_{ijt}$ | 0.0824** | 0.0821** | 0.0825** | 0.134** | 0.132** | 0.131** | 0.0331** | 0.0326** | 0.0322** |
| | (3.965) | (3.940) | (3.956) | (4.737) | (4.629) | (4.635) | (13.60) | (13.41) | (13.27) |
| $SizeDum_{ijt}$ | 0.156** | 0.156** | 0.145** | −0.0377 | −0.0363 | −0.0555 | 0.0860** | 0.0864** | 0.0844** |
| | (4.453) | (4.389) | (4.130) | (− 0.849) | (− 0.824) | (− 1.255) | (9.944) | (10.00) | (9.770) |
| $SizeCorr_{ijt}$ | 0.155** | 0.155** | 0.144** | −0.0433 | −0.0413 | −0.0614 | 0.0854** | 0.0854** | 0.0835** |
| | (4.272) | (4.218) | (3.954) | (− 0.934) | (− 0.898) | (− 1.328) | (9.484) | (9.493) | (9.279) |
| $Bk/MktDum_{ijt}$ | 0.0884** | 0.0881** | 0.0863** | 0.0827* | 0.0826* | 0.0795* | 0.102** | 0.101** | 0.103** |
| | (3.771) | (3.740) | (3.641) | (2.214) | (2.192) | (2.110) | (24.43) | (24.26) | (24.67) |
| $Bk/MktCorr_{ijt}$ | 0.0967** | 0.0964** | 0.0944** | 0.0913* | 0.0912* | 0.0879* | 0.112** | 0.111** | 0.112** |
| | (3.806) | (3.777) | (3.681) | (2.276) | (2.253) | (2.171) | (24.29) | (24.07) | (24.48) |
| $MomDum_{ijt}$ | 0.0727** | 0.0728** | 0.0724** | 0.0635** | 0.0644** | 0.0643** | 0.0542** | 0.0539** | 0.0531** |
| | (3.822) | (3.826) | (3.837) | (2.937) | (2.941) | (2.986) | (27.35) | (27.26) | (26.84) |
| $MomCorr_{ijt}$ | 0.0790** | 0.0791** | 0.0787** | 0.0653* | 0.0662* | 0.0663* | 0.0560** | 0.0557** | 0.0547** |
| | (3.602) | (3.606) | (3.619) | (2.635) | (2.640) | (2.685) | (25.63) | (25.53) | (25.08) |
| $IndDum_{ijt}$ | 0.0990** | 0.0986** | 0.0938** | 0.117** | 0.115** | 0.110** | 0.0396** | 0.0420** | 0.0408** |
| | (5.792) | (5.722) | (5.435) | (5.487) | (5.363) | (5.132) | (6.745) | (7.160) | (6.966) |
| $IndCorr_{ijt}$ | 0.0582* | 0.0585* | 0.0561* | 0.0620* | 0.0621* | 0.0601* | −0.0694** | −0.0699** | −0.0689** |
| | (2.644) | (2.660) | (2.564) | (2.176) | (2.173) | (2.111) | (−33.30) | (−33.69) | (−33.19) |
| $WireDum_{ijt}$ | | 0.00142 | 0.00508 | | −0.00539 | 0.00841 | | 0.00172 | 0.00368** |
| | | (0.294) | (1.418) | | (−0.850) | (1.583) | | (1.367) | (3.379) |
| $TakeSim_{ijt}$ | | 0.344** | 0.241* | | 0.294** | 0.223* | | −0.275** | −0.298** |
| | | (3.064) | (2.229) | | (3.004) | (2.432) | | (−3.774) | (−4.145) |
| $\widetilde{WireSim}_{ijt}$ | | 0.0163 | | | 0.0497** | | | 0.00536 | |
| | | (0.857) | | | (3.478) | | | (1.446) | |
| $WireSim_{ijt}$ | | | 0.171** | | | 0.187** | | | 0.0404** |
| | | | (7.202) | | | (5.419) | | | (7.845) |
| Time fixed effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-pair and firm-specific panel effects | No | No | No | No | No | No | Yes | Yes | Yes |
| Systematic lags | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 5 | 5 |
| Adjusted $R$-squared | 0.523 | 0.523 | 0.524 | 0.585 | 0.585 | 0.586 | | | |
| AR(2) test | | | | | | | −0.563 | −0.261 | −0.328 |
| Observations | 19,750,851 | | | 7,373,461 | | | 1,367,394 | | |
| Panel B: Correlations from 10-day cumulative returns | | | | | | | | | |
| $z_{ijt}$ | 0.122** | 0.122** | 0.121** | 0.154** | 0.153** | 0.152** | 0.0588** | 0.0571** | 0.0576** |
| | (11.09) | (11.14) | (11.11) | (10.15) | (10.10) | (10.04) | (40.42) | (39.54) | (39.81) |
| $BetaDum_{ijt}$ | 0.0993** | 0.0994** | 0.0989** | 0.150** | 0.149** | 0.148** | 0.0498** | 0.0508** | 0.0499** |
| | (5.242) | (5.222) | (5.201) | (4.828) | (4.774) | (4.783) | (18.29) | (18.66) | (18.35) |
| $BetaCorr_{ijt}$ | 0.103** | 0.103** | 0.103** | 0.160** | 0.159** | 0.158** | 0.0464** | 0.0474** | 0.0465** |
| | (5.092) | (5.081) | (5.058) | (4.619) | (4.572) | (4.578) | (16.19) | (16.57) | (16.25) |
| $SizeDum_{ijt}$ | 0.173** | 0.169** | 0.161** | 0.0221 | 0.0279 | 0.00106 | 0.0939** | 0.0891** | 0.0930** |
| | (4.614) | (4.490) | (4.375) | (0.388) | (0.496) | (0.0185) | (12.51) | (11.88) | (12.41) |
| $SizeCorr_{ijt}$ | 0.172** | 0.169** | 0.160** | 0.0169 | 0.0241 | −0.00404 | 0.0860** | 0.0806** | 0.0849** |
| | (4.473) | (4.366) | (4.238) | (0.284) | (0.411) | (−0.0671) | (10.89) | (10.22) | (10.76) |
| $Bk/MktDum_{ijt}$ | 0.0532* | 0.0529* | 0.0515* | 0.0509* | 0.0504* | 0.0480* | 0.0386** | 0.0384** | 0.0388** |
| | (2.705) | (2.643) | (2.634) | (2.318) | (2.273) | (2.188) | (8.025) | (7.991) | (8.072) |
| $Bk/MktCorr_{ijt}$ | 0.0585* | 0.0584* | 0.0568* | 0.0581* | 0.0576* | 0.0550* | 0.0386** | 0.0381** | 0.0384** |
| | (2.724) | (2.669) | (2.660) | (2.407) | (2.363) | (2.277) | (7.083) | (7.010) | (7.066) |
| $MomDum_{ijt}$ | 0.0748** | 0.0748** | 0.0740** | 0.0724** | 0.0728** | 0.0720** | 0.0340** | 0.0350** | 0.0336** |
| | (4.054) | (4.057) | (4.042) | (3.783) | (3.746) | (3.738) | (10.98) | (11.34) | (10.86) |
| $MomCorr_{ijt}$ | 0.0812** | 0.0812** | 0.0804** | 0.0742** | 0.0746** | 0.0740** | 0.0347** | 0.0357** | 0.0340** |
| | (3.828) | (3.831) | (3.819) | (3.397) | (3.365) | (3.367) | (10.19) | (10.50) | (9.997) |
| $IndDum_{ijt}$ | 0.136** | 0.135** | 0.129** | 0.185** | 0.181** | 0.174** | 0.196** | 0.192** | 0.193** |
| | (9.852) | (9.451) | (9.071) | (11.14) | (10.74) | (10.40) | (15.30) | (15.08) | (15.14) |
| $IndCorr_{ijt}$ | 0.0682** | 0.0683** | 0.0662** | 0.0892** | 0.0888** | 0.0868** | −0.0196** | −0.0184** | −0.0191** |
| | (4.082) | (4.079) | (3.987) | (4.681) | (4.634) | (4.578) | (−8.789) | (−8.297) | (−8.568) |

(*continued on next page*)

**Table 3** (*continued*)

| | Ordinary least squares | | | | | | Dynamic panel estimator | | |
|---|---|---|---|---|---|---|---|---|---|
| | Broad sample | | | Larger firms | | | Broad sample | | |
| $WireDum_{ijt}$ | | −0.0114 | 0.00638 | | −0.0156 | 0.00826 | | −0.00968** | −0.00403 |
| | | (− 1.298) | (1.509) | | (− 1.484) | (1.019) | | (−3.723) | (−1.802) |
| $TakeSim_{ijt}$ | | 0.704** | 0.596** | | 0.629** | 0.498** | | −0.507** | −0.505** |
| | | (4.225) | (3.720) | | (4.065) | (3.718) | | (−4.020) | (−3.994) |
| $\widetilde{WireSim}_{ijt}$ | | 0.0709 | | | 0.0863** | | | 0.0258** | |
| | | (1.964) | | | (3.233) | | | (3.425) | |
| $WireSim_{ijt}$ | | | 0.262** | | | 0.332** | | | 0.00891* |
| | | | (6.586) | | | (5.471) | | | (2.456) |
| Time fixed effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-pair and firm-specific panel effects | No | No | No | No | No | No | Yes | Yes | Yes |
| Systematic lags | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 5 | 5 |
| Adjusted $R$-squared | 0.247 | 0.247 | 0.248 | 0.330 | 0.330 | 0.331 | | | |
| AR(2) test | | | | | | | −0.504 | −0.300 | −0.475 |
| Observations | 19,750,851 | | | 7,373,461 | | | 1,367,394 | | |

ument similarities of firms sharing the same text volume decile. As predicted, volume-based document similarity $\widetilde{WireSim}_{ijt}$ contributes little to the information content of the text. Thus, my measure of qualitative similarity, $WireSim_{ijt}$, provides a clearer prediction of how the future payoffs of two firms are correlated. Overall, the results presented in Table 3 confirm that qualitative information about two firms helps predict their future price comovement.

## 4. Alternative explanations for stock price comovement

Prior literature offers many alternative explanations for stock price comovement. Table 5 explores whether the qualitative similarity of newswire text is just a proxy for other previously documented sources of return correlation. For the results reported in Table 5, the systematic controls $BetaCorr_{ijt}$, $BetaDum_{ijt}$, $SizeCorr_{ijt}$, $SizeDum_{ijt}$, $Bk/MktCorr_{ijt}$, $Bk/MktDum_{ijt}$, $MomCorr_{ijt}$, $MomDum_{ijt}$, $IndCorr_{ijt}$ and $IndDum_{ijt}$, in addition to the contemporaneous dependent variable $z_{ijt}$ will be retained in the regression specifications but untabulated to conserve space. The coefficient on $WireSim_{ijt}$ is positive and significant in every specification, regardless of sample truncation or estimation methodology. Furthermore, the magnitude of the economic impact does not change much across the specifications implying that the similarity of IDN text contains information about future excess return correlation that is orthogonal to the 12 additional controls described below.

For all of my analysis, I control for contemporaneous correlations calculated over six months of returns. The predictive power of qualitative similarity, relative to contemporaneous six-month correlations, may stem from the fact that some of the content used to measure qualitative similarity might have arrived later in the formation period. The variables $\rho_{ijt}^{1mo}$ and $\rho_{ijt}^{2mo}$ are pairwise Pearson correlations calculated from the last month and two months, respectively, of daily returns during period $t$. Summary statistics for these short-term correlations are reported in Table 4. Using only 20–40 trading days makes these correlation estimates highly susceptible to isolated price shocks, and, compared to the correlations calculated over six-month windows, $\rho_{ijt}^{1mo}$ and $\rho_{ijt}^{2mo}$ exhibit more variance. Despite the potential for measurement error, Table 5 provides evidence that short-term correlations can predict future stock price comovement, regardless of whether the sample is truncated or correlations are calculated from daily or ten-day cumulative returns.

The information diffusion view, proposed by Barberis et al. (2005) states that, due to some market friction, information is incorporated more quickly into the prices of some stocks than others. In this view, there is a common factor in the returns of stocks that incorporate information at similar rates. My tests of the information diffusion view are based on variables related to analyst following and liquidity. $AnaCorr_{ijt}$ and $AmiCorr_{ijt}$ are constructed using the decile portfolio method described in Section 3. The former divides all firms into NYSE deciles based on the number of unique analysts with an earnings prediction recorded in the I/B/E/S database during period $t$. The latter splits firms into deciles based on the liquidity ratio provided by Amihud et al. (1997). The binary variables $AnaDum_{ijt}$ and $AmiDum_{ijt}$ are set to 1 if both are in the same analyst following or Amihud ratio decile, respectively.

Firms with a larger analyst following and more liquid equity should have stock prices that adjust more quickly to relevant information. According to Table 5, coefficients for the variables based on liquidity are only positive and significant for the DPE specifications, while coefficients related to analyst following are not significant in any of the specifications. Thus, these results provide only weak support for the information diffusion view.

The category view, proposed by Barberis and Shleifer (2003), predicts that in order to simplify portfolio decisions, investors group assets into categories and then allocate funds at the level of these categories rather than at the individual asset level. As in Barberis et al. (2005), membership in the S&P 500 Index will be used to separate firms into categories that should have no fundamental relation with return comovement. Similary, membership in one of the S&P Value or Growth indices, as suggested by Boyer (2011), can also be used to separate firms into categories. Finally, Green and Hwang (2009) propose that investors might also categorize stocks based on price, so a variable accounting for differences in stock price is included.

Index membership is taken from the Compustat Index Constituents file, and all firms that are listed as members of a particular index on the last day of period $t$ are considered index members for that period. The dummy variables $SP500_{ijt}$, $SPVal_{ijt}$, and $SPGrw_{ijt}$ are set to 1 if both firms $i$ and $j$ are members of the S&P 500, S&P 1500 Value and S&P 1500 Growth indices, respectively. According to Table 4, only 5% of the firm-pairs in the broad sample contain two members of the S&P 500, as opposed to 29% in the truncated sample. For the S&P 1500 Value Index, the proportions are 21% and 34%, respectively, whereas the S&P 1500 Growth Index contributes two firm-pair members to the broad sample 15% of the time and two members in the most truncated sample 29% of the time.

Table 5 also reports the tests of the category hypothesis with the dummy variables for index membership and the price portfolio correlations included as regressors. After including my measure of qualitative similarity and the systematic variables, S&P 500 membership is only positive and significant in about half of the specifications. The coefficient on the binary variable $SPVal_{ijt}$ is positive and signficant in the OLS specfications, but negative in the

**Table 4**
Summary statistics for alternative explanations for stock price comovement.
A description for all variable calculations is provided in the surrounding text and in Panel B of Table A-1.

| | Broad sample | | | | | | Larger firms | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | P1 | P5 | P50 | P95 | P99 | Mean | P1 | P5 | P50 | P95 | P99 |
| Panel A: Correlations from daily returns | | | | | | | | | | | | |
| $\rho^{1mo}_{ijt}$ | 0.338 | −0.30 | −0.11 | 0.35 | 0.73 | 0.82 | 0.365 | −0.29 | −0.10 | 0.38 | 0.76 | 0.85 |
| $\rho^{2mo}_{ijt}$ | 0.328 | −0.17 | −0.03 | 0.33 | 0.69 | 0.80 | 0.354 | −0.16 | −0.02 | 0.35 | 0.73 | 0.83 |
| $AnaDum_{ijt}$ | 0.114 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.150 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $AnaCorr_{ijt}$ | 0.823 | 0.00 | 0.00 | 0.93 | 0.98 | 0.99 | 0.794 | 0.00 | 0.00 | 0.93 | 0.98 | 0.99 |
| $AmiDum_{ijt}$ | 0.102 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.171 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $AmiCorr_{ijt}$ | 0.840 | 0.00 | 0.00 | 0.94 | 0.99 | 0.99 | 0.785 | 0.00 | 0.00 | 0.95 | 0.99 | 0.99 |
| $SP500_{ijt}$ | 0.053 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.289 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $SPVal_{ijt}$ | 0.216 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.343 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $SPGrw_{ijt}$ | 0.154 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.286 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $PrcDum_{ijt}$ | 0.112 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.133 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $PrcCorr_{ijt}$ | 0.797 | 0.00 | 0.00 | 0.90 | 0.97 | 0.98 | 0.783 | 0.00 | 0.00 | 0.91 | 0.97 | 0.99 |
| $MSA_{ijt}$ | 0.028 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.033 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Panel B: Correlations from 10-day cumulative returns | | | | | | | | | | | | |
| $AnaCorr_{ijt}$ | 0.816 | 0.00 | 0.00 | 0.93 | 0.99 | 0.99 | 0.786 | 0.00 | 0.00 | 0.93 | 0.99 | 0.99 |
| $AmiCorr_{ijt}$ | 0.833 | 0.00 | 0.00 | 0.94 | 0.99 | 1.00 | 0.779 | 0.00 | 0.00 | 0.94 | 0.99 | 1.00 |
| $PrcCorr_{ijt}$ | 0.773 | 0.00 | 0.00 | 0.88 | 0.98 | 0.99 | 0.762 | 0.00 | 0.00 | 0.89 | 0.98 | 0.99 |

DPE specifications, implying that comovement attributed to membership in the S&P 1500 Value Index is correlated with persistent firm- and firm-pair unobservable heterogeneity. Coefficients on the binary variable $SPGrw_{ijt}$ are not positive and significant in any of the specifications.

NYSE deciles based on the closing stock price of the last trading day in period $t-1$ will be used to form price decile portfolios. Similar to the return-based variables created to test the traditional view, the return correlation between these price portfolios $PrcCorr_{ijt}$ is used to test the category view. The binary variable $PrcDum_{ijt}$ is set to 1 if both are in the price deciles. The coefficients on $PrcCorr_{ijt}$ and $PrcDum_{ijt}$ are only positive and significant in the broad sample DPE specifications. Thus, it is not clear whether investors treat stock price as a category that influences their trading.

Finally, Pirinsky and Wang (2006) suggest that a variety of factors converging around the geographic location of a companies headquarters could cause the stock prices of neighboring firms to comove. The county and state of a company's headquarters locations are taken from the CRSP/Compustat Merged Company Header History file, and merged with the list of Metropolitan Statistical Areas (MSA) defined by the Office of Management and Budget and reported on the Census Bureau's website. All observations with firm-pairs headquartered in the same MSA will have a value of 1 for the dummy variable $MSA_{ijt}$. Much like S&P 500 Index membership, the scope of $MSA_{ijt}$ is rather limited. Table 4 reports that firm-pairs share an MSA in less than 3% of my total sample observations. According to Table 5, the sign and significance of the headquarters location variable's coefficient changes sign across the OLS and DPE specifications. Thus, $MSA_{ijt}$ may also have some degree of collinearity with the unobservable panel effects.

## 5. Qualitative similarity of text across different sources and topics

The Thomson Reuters NewsScope Archive describes the attribution, or source, of each take. There are 12 attributions contributing takes to my sample, however, only *Reuters News* consists of content primarily produced by journalists. Other attributions, such as *Business Wire* or *PR Newswire*, are more likely to contain text generated by the firms themselves in the form of press releases and legal disclosures. To test whether content produced by journalists is more or less informative than text generated by firms, I calculate separate measures of qualitative similarity based on the attribution of the story. $WireSim^{rtrs}_{ijt}$ describes the qualitative similarity of content attributed only to *Reuters News*, whereas $WireSim^{firm}_{ijt}$ measures the qualitative similarity of text originating from all other sources.

Thomson Reuters News Analytics also provides a list of proprietary topic codes identifying the subject matter of each take. Table 6 provides descriptive information about those topic codes that appear most frequently in my sample. All of the 2,624,133 takes included my analysis are written in English (LEN), with journalists and firms contributing 1,539,884 and 1,084,249 takes, respectively. According to Fig. 1, the volume of company-specific content increases dramatically during earnings season. Similarly, Table 6 demonstrates that roughly 29% of all takes appearing on the Reuters IDN discuss corporate financial results (RES). To determine whether content related to earnings contains more information about future return correlation, I also separate my newswire sample based on whether News Analytics assigns the topic code RES to a particular take. $WireSim^{resyes}_{ijt}$ describes the qualitative similarity of takes discussing corporate financial results, while $WireSim^{resno}_{ijt}$ measures the similarity of those that do not.

Even though *Reuters News* contributes more takes to the IDN feed than all other sources combined, Table 7 indicates that journalists only follow a small subset of all firms in the market. For the broad sample, only 49% of firm-pair-periods consist of two companies with some positive quantity of text contributed by *Reuters News*, whereas 83% of firm-pairs have text attributed to other sources. Clearly, the scope of coverage by the traditional press is appreciably more limited than on the IDN as a whole. With regards to topic, the sample is more equitably divided. According to Table 7, 66% or firm-pair-periods consist of companies whose newswire text discusses corporate financial results, and 84% of firm-pair-periods have at least some newswire text that does not focus on earnings.

Table 8 presents estimates of Eqs. (6) and (7) where my measure of qualitative similarity, and the other text-based controls, are based specific types or sources of IDN content. Specifications with correlations based on ten-day cumulative returns are untabulated to conserve space, but the inferences are qualitatively similar. The systematic controls $BetaCorr_{ijt}$, $BetaDum_{ijt}$, $SizeCorr_{ijt}$, $SizeDum_{ijt}$, $Bk/MktCorr_{ijt}$, $Bk/MktDum_{ijt}$, $MomCorr_{ijt}$, $MomDum_{ijt}$, $IndCorr_{ijt}$, and $IndDum_{ijt}$ are retained in all regression specifications but untabulated to conserve space. Variables based on the alternative explanations for stock price comovement introduced in Section 4, $\rho^{1mo}_{ijt}$,

**Table 5**
Alternative explanations for stock price comovement.
The dependent variable in all specifications is the Fisher transformation $z_{ijt+1}$ of the Pearson correlation $\rho_{ijt+1}$ calculated from the returns of firms $i$ and $j$ in excess of the risk free rate for each six-month period $t+1$. The binary variable $WireDum_{ijt}$ is set to 1 whenever both firms have some positive number of total words transmitted across the Reuters Integrated Data Network. $TakeSim_{ijt}$, defined in Eq. (4), accounts for how often firm-pairs are mentioned in the same newswire take. $WireSim_{ijt}$ is a measure of qualitative similarity defined in Eq. (2). A description for all other included variable calculations is provided in the surrounding text and in Panel B of Table A-1. Eq. (6) is estimated with ordinary least squares and Eq. (7) is estimated with a dynamic panel estimation (DPE) methodology. Ordinary least squares standard errors are clustered by firm-pair, both individual firms and time using the Cameron et al. (2011) multi-way clustering procedure. DPE results are generated using the approach described in Arellano and Bover (1995) and Blundell and Bond (1998) with bias-corrected robust variance-covariance estimates of the model parameters. Coefficients marked $^{*}$ and $^{**}$ are significant at the 5% and 1% level, respectively, and $t$-statistics are reported in parenthesis. All of the independent variables are used as predetermined instruments in the DPE specifications. "Systematic lags" refers to the total number of lags included in each specification for the variables $z_{ijt}$, $BetaDum_{ijt}$, $BetaCorr_{ijt}$, $SizeDum_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktDum_{ijt}$, $Bk/MktCorr_{ijt}$, $MomDum_{ijt}$, $MomCorr_{ijt}$, $IndDum_{ijt}$ and $IndCorr_{ijt}$.

| | Correlations calculated from daily returns | | | | | | Correlations calculated from 10-day returns | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ordinary least squares | | | | Dynamic panel estimator | | Ordinary least squares | | | | Dynamic panel estimator | |
| | Broad sample | | Larger firms | | Broad sample | | Broad sample | | Larger firms | | Broad sample | |
| $\rho_{ijt}^{1mo}$ | 0.0201$^{**}$ | 0.0201$^{**}$ | 0.0215$^{*}$ | 0.0215$^{*}$ | 0.0179$^{**}$ | 0.0180$^{**}$ | 0.0275$^{*}$ | 0.0274$^{*}$ | 0.0263 | 0.0262 | 0.0243$^{**}$ | 0.0243$^{**}$ |
| | (2.993) | (2.986) | (2.242) | (2.241) | (20.10) | (20.23) | (2.366) | (2.367) | (1.948) | (1.957) | (12.99) | (12.97) |
| $\rho_{ijt}^{2mo}$ | 0.0398$^{**}$ | 0.0395$^{**}$ | 0.0482$^{**}$ | 0.0477$^{**}$ | 0.0283$^{**}$ | 0.0285$^{**}$ | 0.165$^{**}$ | 0.163$^{**}$ | 0.202$^{**}$ | 0.200$^{**}$ | 0.0399$^{**}$ | 0.0402$^{**}$ |
| | (3.307) | (3.311) | (3.181) | (3.185) | (21.26) | (21.41) | (8.229) | (8.194) | (8.077) | (8.013) | (15.61) | (15.73) |
| $AnaDum_{ijt}$ | −0.0174 | −0.0225 | −0.000758 | −0.00881 | 0.00164 | 2.62e-05 | −0.0382 | −0.0415 | 0.00121 | −0.00276 | −0.00394 | −0.00421 |
| | (−0.611) | (−0.800) | (−0.0220) | (−0.266) | (0.201) | (0.00321) | (−1.121) | (−1.213) | (0.0290) | (−0.0660) | (−0.487) | (−0.522) |
| $AnaCorr_{ijt}$ | −0.0200 | −0.0253 | 0.00128 | −0.00697 | 5.21e-05 | −0.00167 | −0.0426 | −0.0459 | 0.00345 | −0.000295 | −0.00733 | −0.00760 |
| | (−0.660) | (−0.845) | (0.0350) | (−0.198) | (0.00611) | (−0.196) | (−1.181) | (−1.268) | (0.0767) | (−0.00659) | (−0.863) | (−0.895) |
| $AmiDum_{ijt}$ | 0.0220 | 0.0157 | 0.0546 | 0.0446 | 0.115$^{**}$ | 0.113$^{**}$ | 0.0612 | 0.0571 | 0.0122 | 0.00240 | 0.0522$^{**}$ | 0.0507$^{**}$ |
| | (0.717) | (0.525) | (1.570) | (1.303) | (12.13) | (11.97) | (1.918) | (1.823) | (0.334) | (0.0677) | (5.879) | (5.718) |
| $AmiCorr_{ijt}$ | 0.0206 | 0.0143 | 0.0567 | 0.0465 | 0.116$^{**}$ | 0.114$^{**}$ | 0.0619 | 0.0579 | 0.0142 | 0.00456 | 0.0506$^{**}$ | 0.0489$^{**}$ |
| | (0.649) | (0.460) | (1.585) | (1.326) | (11.99) | (11.81) | (1.874) | (1.786) | (0.375) | (0.124) | (5.573) | (5.388) |
| $SP500_{ijt}$ | 0.0172$^{*}$ | 0.0163 | 0.0135$^{**}$ | 0.0132$^{**}$ | 0.00420 | 0.00342 | 0.0284$^{*}$ | 0.0271$^{*}$ | 0.00816 | 0.00782 | 0.0605$^{**}$ | 0.0606$^{**}$ |
| | (2.134) | (2.037) | (3.304) | (3.242) | (1.790) | (1.459) | (2.536) | (2.441) | (1.475) | (1.427) | (13.97) | (14.01) |
| $SPVal_{ijt}$ | 0.0229$^{**}$ | 0.0225$^{**}$ | 0.0213$^{**}$ | 0.0209$^{**}$ | −0.00575$^{**}$ | −0.00562$^{**}$ | 0.0276$^{**}$ | 0.0269$^{**}$ | 0.0228$^{**}$ | 0.0220$^{**}$ | −0.00398$^{**}$ | −0.00360$^{*}$ |
| | (5.618) | (5.581) | (4.299) | (4.224) | (−8.062) | (−7.903) | (4.776) | (4.740) | (3.779) | (3.681) | (−2.819) | (−2.565) |
| $SPGrw_{ijt}$ | −0.00327 | −0.00333 | 0.000537 | 0.000636 | −0.00418$^{**}$ | −0.00460$^{**}$ | −0.00503 | −0.00512 | −0.00781 | −0.00768 | 0.000684 | 0.000638 |
| | (−0.952) | (−0.975) | (0.144) | (0.172) | (−5.847) | (−6.443) | (−1.006) | (−1.029) | (−1.444) | (−1.426) | (0.475) | (0.444) |
| $PrcDum_{ijt}$ | 0.00391 | 0.00332 | 0.0169 | 0.0173 | 0.0208$^{**}$ | 0.0198$^{**}$ | −0.0138 | −0.0141 | −0.00513 | −0.00441 | −0.000211 | −0.00136 |
| | (0.135) | (0.117) | (0.575) | (0.590) | (5.655) | (5.377) | (−0.602) | (− 0.624) | (− 0.455) | (−0.387) | (−0.0456) | (−0.293) |
| $PrcCorr_{ijt}$ | 0.00305 | 0.00240 | 0.0153 | 0.0158 | 0.0195$^{**}$ | 0.0183$^{**}$ | −0.0172 | −0.0176 | −0.00849 | −0.00749 | −0.00174 | −0.00310 |
| | (0.0957) | (0.0765) | (0.466) | (0.484) | (4.891) | (4.606) | (−0.671) | (− 0.693) | (− 0.664) | (−0.582) | (−0.344) | (−0.614) |
| $MSA_{ijt}$ | 0.000257 | 8.10e-05 | 0.0108$^{**}$ | 0.0104$^{**}$ | −0.0714$^{**}$ | −0.0708$^{**}$ | 0.00607$^{*}$ | 0.00581$^{*}$ | 0.0197$^{**}$ | 0.0191$^{**}$ | −0.0773$^{*}$ | −0.0729$^{*}$ |
| | (0.153) | (0.0483) | (4.102) | (4.033) | (−3.922) | (−3.831) | (2.376) | (2.284) | (4.909) | (4.827) | (−2.104) | (−1.997) |
| $WireDum_{ijt}$ | | 0.00316 | | 0.00460 | | 0.00442$^{**}$ | | 0.00336 | | 0.00253 | | −0.000953 |
| | | (0.892) | | (0.888) | | (4.156) | | (0.851) | | (0.336) | | (−0.437) |
| $TakeSim_{ijt}$ | | 0.214 | | 0.213$^{*}$ | | −0.333$^{**}$ | | 0.521$^{**}$ | | 0.485$^{**}$ | | −0.505$^{**}$ |
| | | (2.056) | | (2.423) | | (−4.548) | | (3.708) | | (4.060) | | (−4.043) |
| $WireSim_{ijt}$ | | 0.158$^{**}$ | | 0.171$^{**}$ | | 0.0421$^{**}$ | | 0.211$^{**}$ | | 0.273$^{**}$ | | 0.0134$^{*}$ |
| | | (7.237) | | (5.255) | | (8.223) | | (6.204) | | (5.030) | | (2.249) |
| Time fixed effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-pair and firm-specific panel effects | No | No | No | No | Yes | Yes | No | No | No | No | Yes | Yes |
| Total lags | 1 | 1 | 1 | 1 | 5 | 5 | 1 | 1 | 1 | 1 | 5 | 5 |
| Adjusted $R$-squared | 0.528 | 0.528 | 0.590 | 0.591 | | | 0.259 | 0.259 | 0.343 | 0.344 | | |
| AR(2) test | | | | | 0.260 | 0.419 | | | | | −0.605 | −0.531 |
| Observations | 19,750,851 | | 7,373,461 | | 1,367,394 | | 7,373,461 | | 7,373,461 | | 1,367,394 | |

$\rho_{ijt}^{2mo}$, $AnaDum_{ijt}$, $AnaCorr_{ijt}$, $AmiDum_{ijt}$, $AmiCorr_{ijt}$, $SP500_{ijt}$, $SPVal_{ijt}$, $SPGrw_{ijt}$, $PrcDum_{ijt}$, $PrcCorr_{ijt}$ and $MSA_{ijt}$ are also included in all specifications as additional controls.

When the focus is narrowed to text produced by journalists, the magnitude of the coefficient on $WireSim_{ijt}^{rtrs}$ is just marginally significant for the DPE specification and only significant for the OLS specification when the sample is truncated by firm size. Conversely, the coefficients for $WireSim_{ijt}^{firm}$ are much larger and, at least marginally, significant in all three specifications. Thus, there is less evidence that Reuters News produces qualitative information that can predict future cross-firm comovement for a broad range of companies. These results do not imply that the output from financial journalists lacks information relevant to predicting future comovement. It may be that this output just fails to make a marginal contribution to the companies' information flows relative to the content they produce themselves. If, as claimed by Ahern and Sosyura (2014), Reuters News merely summarizes firm-generated content instead of contributing their own analysis, a thorough examination of their output will produce little incremental information.

When newswire text is separated by topic, I find that $WireSim_{ijt}^{resno}$ is a stronger predictor of future return correlation, in terms of coefficient magnitude, than $WireSim_{ijt}^{resyes}$. Once again, the comparably weaker predictive performance of newswire content focused on corporate results does not imply that these stories contain less information. If most of the information revealed in newswire takes with topic code RES is communicated through a numerical value such as an earnings level, then the text accompanying this release might just contain less relevant qualitative information.

## 6. Other sources of qualitative information

Israelsen (2015) and Muslu et al. (2014) both examine the effect of common analyst coverage on stock return comovement. For my study, the measure of analyst coverage provided by Israelsen (2015) will be used to determine the proportion of a

**Table 6**
Frequency of topic codes across attributions.
This table provides the frequency, type and definition for the most common topic codes across all takes appearing in my sample. The frequency of each topic code is also reported separately for *Reuters News* and for all other attributions.

| | All content | | Reuters news | | Other attributions | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Code | *N* | % | *N* | % | *N* | % | Type | Name | Definition |
| LEN | 2,624,133 | 100.0 | 1,539,884 | 100.0 | 1,084,249 | 100.0 | Language | English | Stories in English. |
| US | 2,178,709 | 83.0 | 1,474,515 | 95.8 | 704,194 | 64.9 | Geography | United States | United States of America |
| BACT | 887,250 | 33.8 | 512,920 | 33.3 | 374,330 | 34.5 | Event Type | Corporate Events | All business events relating to companies and other issuers of securities. |
| CMPNY | 878,185 | 33.5 | 582,307 | 37.8 | 295,878 | 27.3 | Broad News Topic | Company News | Company news (added automatically when a story contains any company RIC). |
| RES | 751,094 | 28.6 | 556,283 | 36.1 | 194,811 | 18.0 | Event Type | Performance / Results / Earnings | All corporate financial results; tabular and textual reports; dividends; accounts, annual reports; forecasts and estimates of future earnings; corporate insolvencies and bankruptcies. |
| NEWR | 728,457 | 27.8 | 18 | 0.0 | 728,439 | 67.2 | Genre | News Announcements | Announcements made as news releases to media organizations, including corporate announcements and regulatory disclosures. |
| BUS | 586,821 | 22.4 | 282,232 | 18.3 | 304,589 | 28.1 | Business Sector | Business, Public Services | Services to business and consumers including office supplies; advertising / marketing; data vendors, software development and data processing; security; transporters, custom agents, package & mail delivery; port-harbour transport & warehousing; airport, port, tunnel, highway management; agencies; water distribution; waste management, cleaning, water filtration. |
| FIN | 484,107 | 18.4 | 248,832 | 16.2 | 235,275 | 21.7 | Business Sector | Financials | Companies engaged in the operation of retail and commercial banks, insurance companies, real estate operations, investment trusts and other financial service providers. |
| AMERS | 456,544 | 17.4 | 293,164 | 19.0 | 163,380 | 15.1 | Geography | Americas | |
| INDS | 407,005 | 15.5 | 226,873 | 14.7% | 180,132 | 16.6 | Business Sector | Industrials | Manufacturers of industrial equipment and commercial supplies, as well as providers of related services, such as diversified trading, distribution operations and transportation services. |
| FINS | 405,776 | 15.5 | 264,396 | 17.2 | 141,380 | 13.0 | Business Sector | Financials | Operators of commercial and investment banks, investment trusts and financial markets, as well as providers of investment, insurance and real estate services. |
| CYCS | 334,969 | 12.8 | 203,374 | 13.2 | 131,595 | 12.1 | Business Sector | Cyclical Consumer Goods & Services | Manufacturers of automobiles, household goods, textiles and other products, as well as homebuilders and retailers, and providers of consumer services, such as hotel, entertainment and media services. |
| BNK | 333,216 | 12.7 | 164,938 | 10.7 | 168,278 | 15.5 | Business Sector | Banking Services | Companies engaged in retail and commercial banking, providers of consumer financial services, investment services, mortgage REITs, insurance brokers and other loan and financing operations. |
| TECH | 329,487 | 12.6 | 176,487 | 11.5 | 153,000 | 14.1 | Business Sector | Technology | Manufacturers of semiconductors, communications equipment, computer hardware and technology related office equipment, as well as providers of consulting and IT services. |
| DRU | 320,738 | 12.2 | 171,945 | 11.2 | 148,793 | 13.7 | Business Sector | Biotechnology / Pharmaceuticals | Companies engaged in manufacturing and marketing generic and specialty drugs as well as research and development activities for new drugs, medical products and procedures. |
| RESF | 248,087 | 9.5% | 239,256 | 15.5 | 8,831 | 0.8 | Event Type | Results Forecasts / Warnings | Forecasts or "guidance" given by a company about its future results, including profit warnings. |
| RCH | 243,642 | 9.3 | 231,352 | 15.0 | 12,290 | 1.1 | Event Type | Broker Research / Recommendations | The issuing of an investment opinion by a broker / analyst about whether a given stock is a 'buy', 'sell' or a 'hold', or giving a target share price. |

firm-pair's information-related comovement that is attributable to commonality in their analyst following. This variable is defined as:

$$EPSSim_{ijt} = N^{an}_{ijt} / \sqrt{N^{an}_{it} N^{an}_{jt}} \tag{8}$$

where $N^{an}_{ij}$ is the number of analysts from the I/B/E/S database following both firms $i$ and $j$ in a period $t$, and $N^{an}_{it}$ and $N^{an}_{jt}$ are the number of analysts following firms $i$ and $j$ respectively. Variables for instiutional and mutual fund ownership, $S34Sim_{ijt}$ and $S12Sim_{ijt}$, are constructed in an analgolous way, and the variable $S12Sim_{ijt}$ should be highly correlated with the common ownership measure $FCAP_{ijt}$ proposed in Anton and Polk (2014).

Across the broad sample, Table 9 reports that average commonality in ownership is higher for institutions, 37%, than for mutual funds, 21%, though a sizable quantity of firm-pairs are owned by the same organizations in either case. However, with an average of only 0.5%, commonality in analyst following, $EPSSim_{ijt}$, is rare because a large number of firms have no analyst following at all. Therefore, the narrow scope of this variable may limit the practical applicability of this measure.

From the online Hoberg-Phillips Industry Classification Library, the variable $HobPhiScr_{ijt}$ is the yearly firm-by-firm pairwise similarity score calculated by parsing the product descriptions of company 10Ks, then forming word vectors for each firm to compute continuous measures of product similarity. Their variable is very similar to unadjusted document similarity $\widetilde{WireSim}_{ijt}$ discussed in Section 1.3. Unfortunately, Hoberg and Phillips (2015c) only make their measure publicly available for firms having pairwise sim-

**Table 7**
Summary statistics for qualitative similarity across different sources and subjects.
A description for all variable calculations is provided in the surrounding text and in Panel C of Table A-1.

| | Broad sample | | | | | | Larger Firms | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | P1 | P5 | P50 | P95 | P99 | Mean | P1 | P5 | P50 | P95 | P99 |
| $WireDum_{ijt}^{rtrs}$ | 0.491 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.748 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| $WireDum_{ijt}^{firm}$ | 0.828 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.895 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| $WireDum_{ijt}^{resno}$ | 0.838 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.928 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| $WireDum_{ijt}^{resyes}$ | 0.656 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.759 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| $TakeSim_{ijt}^{rtrs}$ | 0.000060 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00017 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $TakeSim_{ijt}^{firm}$ | 0.000044 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00016 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $TakeSim_{ijt}^{resno}$ | 0.000056 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00019 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $TakeSim_{ijt}^{resyes}$ | 0.000029 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.000079 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $WireSim_{ijt}^{rtrs}$ | 0.0020 | −0.092 | −0.058 | 0.000 | 0.064 | 0.113 | 0.0075 | −0.084 | −0.052 | 0.000 | 0.077 | 0.151 |
| $WireSim_{ijt}^{firm}$ | 0.0024 | −0.088 | −0.057 | 0.000 | 0.066 | 0.113 | 0.0077 | −0.076 | −0.050 | 0.002 | 0.072 | 0.129 |
| $WireSim_{ijt}^{resno}$ | 0.0009 | −0.137 | −0.093 | 0.000 | 0.093 | 0.155 | 0.0067 | −0.132 | −0.086 | 0.000 | 0.101 | 0.164 |
| $WireSim_{ijt}^{resyes}$ | 0.0024 | −0.159 | −0.074 | 0.000 | 0.089 | 0.194 | 0.0070 | −0.104 | −0.059 | 0.000 | 0.089 | 0.165 |

ilarities that are above a certain threshold. The binary variable $HobPhiDum_{ijt}$ is set to 1 if both firms $i$ and $j$ are above this minimum level. According to Table 9, the scope of the product similarity measures is also limited, with $HobPhiDum_{ijt}$ averaging only 2% across the broad sample. However, unlike commonality in analyst following, The Hoberg and Phillips (2010a, 2010b) methodology could easily be adapted to produce a product similarity score that spans nearly all possible firm-pairs.

For tonal measures of newswire content, the Thomson Reuters News Analytics dataset contains a proprietary variable designed to measure the sentiment of newswire text. Every firm mentioned in a particular take is given a positive, negative and neutral sentiment probability by Thomson Reuters, and the three values must sum to 1. Each period $t$, the word weighted average positive and negative sentiment across all takes is calculated for every firm with some positive volume of text. The dummy variable $SentPos_{ijt}$ ($SentNeg_{ijt}$) is set to 1 if the average positive (negative) sentiment value of both firms $i$ and $j$ is above the median level for period $t$. According to Table 9, the tone of newswire text written about a pair is either jointly positive or jointly negative a litle more than 37% of the time.

To control for cross-firm return predicatability between companies frequently mentioned in the same newswire items, the variable $TakeSim_{ijt}^{Scher}$ is computed just as the $TakeSim_{ijt}$ variable described in Eq. (4). However, take counts, $N_{ijt}^{take}$, $N_{it}^{take}$ and $N_{jt}^{take}$, are now based only on the subset of newswire items that would be eligible for the Scherbina and Schlusche (2015) sample. Therefore, I only consider takes that mention exactly two firms, and I remove all takes where the sentiment classes of co-mentioned firms differ by an absolute value of two.[9] Next, I discard news items that contain variations of the words “rival” or “competitior” in the headline. While $TakeSim_{ijt}$ is greater than zero for 44,389 observations in my sample, $TakeSim_{ijt}^{Scher}$ is only positive during 8956 firm-pair periods.

The relation between these other sources of qualitative information and future daily return correlation is examined in Panel A of Table 10. Once again, specifications with correlations based on ten-day cumulative returns are untabulated to conserve space, however, the inferences are qualitatively similar to those reported. The results demonstrate that both of the Hoberg and Phillips (2010a, 2010b) product similarity measures are strong predictors of future comovement. Furthermore, the significance of the broad qualitative similarity measure $WireSim_{ijt}$ is unaffected by the inclusion of these other text-based measures. This implies that newswire text contains at least some information about future return correlations that is orthogonal to what is found in company product descriptions, or any of the other sources of qualitative information included in these specifications.

With regards to these other potential sources of qualitative information, however, the results are less consistent. The coefficients on the variables measuring shared mutual fund ownership $S12Sim_{ijt}$, common analyst following $EPSSim_{ijt}$ and newswire co-mentions $TakeSim_{ijt}^{Scher}$ are only positive and signifigant in the OLS specifications, whereas the institutional ownership variable $S34Sim_{ijt}$ only has a positive coefficient in the DPE specifications. Obviously, all four variables are correlated with the firm and firm-pair panel effects. Prior studies reporting a positive and significant relation between ownership or coverage variables and future price comovement might not account for persistent unobservable heterogeneity between firm-pairs. In most cases, analysts and reporters will follow, and institutions and mutual funds will hold, firms with similar characteristics. If a variable like $TakeSim_{ijt}^{Scher}$is truly a source of comovement, and not just a proxy for shared firm characterisitics, we should observe a rise in return correlation, above and beyond the persistent heterogeneity observed during prior periods, after the number of newswire stories mentioning a firm-pair increases.

Finally, the sentiment variables, $SentPos_{ijt}$ and $SentNeg_{ijt}$, are not positively related to future comovement in any of the specificaitons. Furthermore, the coefficients on both variables are negative and significant in the DPE specifications. This implies that firm-pairs whose newswire text falls in either tonal extreme have lower future return correlation. To ensure that the tone of newswire text is not influencing my main result, I also interact $SentPos_{ijt}$ and $SentNeg_{ijt}$ with my measure of qualitative similarity. According to Table 10, the average tone of newswire text does not affect the relation between $WireSim_{ijt}$ and future return correlation.

Next, I will examine another important difference between my empirical approach and what was proposed in some of these related projects. At a minimum, qualitative similarity should be related to comovement in returns after subtracting off the risk-free rate. Other authors analyzing return correlation have instead focused on some measure of excess comovement, but the field of finance lacks a widely accepted definition of what constitutes “excess.” Building on an approach utilized in Ledoit and Wolf (2003) and Bekaert et al. (2005, 2009), Anton and Polk (2014) analyze covariance in the context of risk based models; where excess comovement is approximated by the correlation between traditional factor model residuals. Israelsen (2015) also analyzes a variety of specifications where the dependent variable is the corre-

[9] Following Scherbina and Schlusche (2015), I also delete takes with topic codes INSI, STX, HOT, INDX, AAA, LIST1, USC, MEVN, RCH, FUND and DBT.

**Table 8**
Qualitative similarity of text across different sources and subjects.
The dependent variable in all specifications is the Fisher transformation $z_{ijt+1}$ of the Pearson correlation $\rho_{ijt+1}$ calculated from the daily returns of firms $i$ and $j$ in excess of the risk-free rate for each six-month period $t+1$. A description for all other included variable calculations is provided in the surrounding text and in Panel C of Table A-1. Eq. (6) is estimated with ordinary least squares and Eq. (7) is estimated with a dynamic panel estimation (DPE) methodology. Ordinary least squares standard errors are clustered by firm-pair, both individual firms and time using the Cameron et al. (2011) multi-way clustering procedure. DPE results are generated using the approach described in Arellano and Bover (1995) and Blundell and Bond (1998) with bias-corrected robust variance-covariance estimates of the model parameters. Coefficients marked * and ** are significant at the 5% and 1% level, respectively, and $t$-statistics are reported in parenthesis. All of the independent variables are used as predetermined instruments in the DPE specifications. "Systematic lags" refers to the total number of lags included in each specification for the variables $z_{ijt}$, $BetaDum_{ijt}$, $BetaCorr_{ijt}$, $SizeDum_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktDum_{ijt}$, $Bk/MktCorr_{ijt}$, $MomDum_{ijt}$, $MomCorr_{ijt}$, $IndDum_{ijt}$ and $IndCorr_{ijt}$. "Alternative Controls" refers to the inclusion of $\rho^{1mo}_{ijt}$, $\rho^{2mo}_{ijt}$, $AnaDum_{ijt}$, $AnaCorr_{ijt}$, $AmiDum_{ijt}$, $AmiCorr_{ijt}$, $SP500_{ijt}$, $SPVal_{ijt}$, $SPGrw_{ijt}$, $PrcDum_{ijt}$, $PrcCorr_{ijt}$ and $MSA_{ijt}$ as untabulated controls.

| | Ordinary least squares | | | | | | | | Dynamic panel estimator | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Broad sample | | | | Larger firms | | | | Broad sample | | | |
| $z_{ijt}$ | 0.355$^{**}$ | 0.354$^{**}$ | 0.354$^{**}$ | 0.356$^{**}$ | 0.392$^{**}$ | 0.392$^{**}$ | 0.391$^{**}$ | 0.392$^{**}$ | 0.192$^{**}$ | 0.194$^{**}$ | 0.195$^{**}$ | 0.193$^{**}$ |
| | (23.12) | (23.14) | (23.16) | (23.12) | (25.29) | (25.19) | (25.14) | (25.30) | (117.0) | (118.7) | (119.2) | (117.6) |
| $WireDum^{rtrs}_{ijt}$ | −0.00450 | | | | −0.00575 | | | | −0.00199$^{**}$ | | | |
| | (−1.470) | | | | (−1.936) | | | | (−4.580) | | | |
| $TakeSim^{rtrs}_{ijt}$ | 0.151$^{*}$ | | | | 0.188$^{*}$ | | | | −0.128$^{**}$ | | | |
| | (2.643) | | | | (2.628) | | | | (−3.095) | | | |
| $WireSim^{rtrs}_{ijt}$ | 0.0108 | | | | 0.0803$^{**}$ | | | | 0.00519 | | | |
| | (1.163) | | | | (4.531) | | | | (1.804) | | | |
| $WireDum^{firm}_{ijt}$ | | 0.00278 | | | | 0.00287 | | | | 0.00374$^{**}$ | | |
| | | (1.093) | | | | (0.848) | | | | (4.221) | | |
| $TakeSim^{firm}_{ijt}$ | | 0.139$^{*}$ | | | | 0.0901 | | | | −0.0764 | | |
| | | (2.715) | | | | (1.602) | | | | (−1.170) | | |
| $WireSim^{firm}_{ijt}$ | | 0.129$^{**}$ | | | | 0.115$^{**}$ | | | | 0.00946 | | |
| | | (5.672) | | | | (3.640) | | | | (1.911) | | |
| $WireDum^{resno}_{ijt}$ | | | 0.000243 | | | | 0.00258 | | | | 0.000943 | |
| | | | (0.0776) | | | | (0.605) | | | | (1.389) | |
| $TakeSim^{resno}_{ijt}$ | | | 0.240$^{**}$ | | | | 0.263$^{**}$ | | | | −0.159$^{*}$ | |
| | | | (3.235) | | | | (4.615) | | | | (−2.487) | |
| $WireSim^{resno}_{ijt}$ | | | 0.157$^{**}$ | | | | 0.140$^{**}$ | | | | 0.0191$^{**}$ | |
| | | | (7.302) | | | | (3.983) | | | | (3.949) | |
| $WireDum^{resyes}_{ijt}$ | | | | 0.000868 | | | | 0.000143 | | | | 0.00332$^{**}$ |
| | | | | (0.404) | | | | (0.0579) | | | | (5.893) |
| $TakeSim^{resyes}_{ijt}$ | | | | 0.0540 | | | | 0.0192 | | | | −0.157$^{**}$ |
| | | | | (1.009) | | | | (0.373) | | | | (−3.007) |
| $WireSim^{resyes}_{ijt}$ | | | | 0.00939 | | | | 0.0640$^{**}$ | | | | 0.0106$^{**}$ |
| | | | | (0.926) | | | | (6.466) | | | | (3.041) |
| Time fixed effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-pair and Firm-specific panel effects | No | No | No | No | No | No | No | No | Yes | Yes | Yes | Yes |
| Alternative controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Systematic lags | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 5 | 5 | 5 |
| Adjusted $R$-squared | 0.528 | 0.528 | 0.528 | 0.528 | 0.591 | 0.591 | 0.591 | 0.590 | | | | |
| AR(2) test | | | | | | | | | 0.403 | 0.625 | 0.442 | 0.536 |
| Observations | 19,750,851 | | | | 7,373,461 | | | | 1,367,394 | | | |

lation between the daily residuals from an estimated asset pricing equation.

For Panel B of Table 10, the Carhart (1997) four-factor model is estimated over two years of returns ending on the last day of period $t+1$. The dependent variable in these specifications is the Fisher transformation $z'_{ijt+1}$ of the Pearson correlation $\rho'_{ijt+1}$ between the residuals of firms $i$ and $j$ during period $t+1$. The returns used to construct all of the portfolio correlation variables ($BetaCorr_{ijt}$, $SizeCorr_{ijt}$, etc.) are also calculated from factor model residuals. The first thing that stands out in Panel B is the dramatic decline in the goodness of fit measure. With 35 independent variables and time series fixed effects, the adjusted $R$-squared still falls to 0.081 in the broad sample OLS specifications. Thus, it seems that the residuals estimated from the four-factor model are mostly devoid of any meaningful structure.

Most importantly, the sign and significance of the qualitative similarity measure $WireSim_{ijt}$ does not change between panels. However, the product similarity measure $HobPhiScr_{ijt}$ is no longer significant in some of the DPE specifications. The coefficients on $S12Sim_{ijt}$ and $EPSSim_{ijt}$ are now positive and signifigant even when the estimation allows for firm-pair panel effects. Thus, the impact of shared mutual fund ownership and common analyst following does change based on factor model specification.

## 7. Investor inattention and economically linked firms

Regardless of whether interfirm linkages are based on firm-specific customer-supplier relationships (Cohen and Frazzini, 2008), conglomeration across different lines of business (Cohen and Lou, 2012) or formal strategic alliances (Cao et al., 2016), the newswire content associated with these connected firms should have higher relative qualitative similarity. Therefore, it is possible that qualitative similarity is only able to predict return correlation between a pair of firms because their newswire text describes the relationships that generate these economic linkages. To ensure that qualitative similarity is not a proxy for these relationships, I perform my analysis on subsets of firms that are less likely to have formal economic linkages.

My identification of connected firms is based on two dimensions. First, I remove all observations for the 29,892 firm-pairs that are mentioned in the same newswire take ($TakeSim_{ijt} > 0$) at least once during my sample period. Next, I remove all 120,908 firm-pairs with Hoberg and Phillips (2010a, 2010b) product similarity measures above the minimum threshold ($HobPhiDum_{ijt} = 1$)

**Table 9**
Summary statistics for other sources of qualitative information.
A description for all variable calculations is provided in the surrounding text and in Panel D of Table A-1.

| | Broad sample | | | | | | Larger firms | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | P1 | P5 | P50 | P95 | P99 | Mean | P1 | P5 | P50 | P95 | P99 |
| Panel A: Correlations from daily returns | | | | | | | | | | | | |
| $S34Sim_{ijt}$ | 0.365 | 0.00 | 0.00 | 0.43 | 0.57 | 0.62 | 0.411 | 0.00 | 0.00 | 0.48 | 0.60 | 0.65 |
| $S12Sim_{ijt}$ | 0.207 | 0.00 | 0.00 | 0.17 | 0.51 | 0.60 | 0.234 | 0.00 | 0.00 | 0.22 | 0.50 | 0.56 |
| $EPSSim_{ijt}$ | 0.0048 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.0089 | 0.00 | 0.00 | 0.00 | 0.04 | 0.28 |
| $HobPhiDum_{ijt}$ | 0.022 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.022 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| $HobPhiScr_{ijt}$ | 0.001 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.001 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 |
| $SentPos_{ijt}$ | 0.378 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.430 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $SentNeg_{ijt}$ | 0.372 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.395 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| $TakeSim_{ijt}^{Scher}$ | 0.00010 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00034 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Panel B: Correlations from daily Carhart residuals | | | | | | | | | | | | |
| $\rho'_{ijt}$ | 0.0049 | −0.24 | −0.16 | 0.00 | 0.18 | 0.28 | 0.0099 | −0.25 | −0.17 | 0.01 | 0.20 | 0.34 |
| $z'_{ijt}$ | 0.0051 | −0.25 | −0.17 | 0.00 | 0.18 | 0.29 | 0.0105 | −0.26 | −0.17 | 0.01 | 0.20 | 0.36 |

at least one time in my sample. These filters should remove firm-pairs where the companies sell at least one similar product, announce a strategic partnership or have a public supply chain relationship at any point during the eleven year sample period. There is no guarantee, however, that this sort of sample truncation will account for less direct industry-specific customer-supplier relationships (Menzly and Ozbas, 2010).

Scherbina and Schlusche (2015) discuss "second-tier" linkages whereby two firms are not linked directly but are linked through their respective direct linkages to some other firm. Along these lines, I also eliminate firm-pairs that are indirectly linked by their appearance in the same newswire take or by their product market similarity. Even though only 29,892 firm-pairs are mentioned in the same take, my sample includes 605,591 firm-pairs that are indirectly linked at least one time. Indirect product market linkages are even more common. Out of the 3,146,459 firm-pairs in my sample, 2,423,970 have second-tier product market connections. Despite the severity of these filters, the truncated sample still retains observations related to all 2723 firms in my initial sample.

Table 11 reports estimates of the same basic specifications analyzed in Table 3, except that the sample firms must belong to one of three subsets: firm-pairs without newswire linkages, firm-pairs without product market linkages, and firm-pairs without newswire or product market linkages. The first subset is large enough that firm-pairs must still be randomly selected before estimating the OLS model. Even though the last two subsets are small enough to preclude sampling, I was unable to find a lag structure that would satisfy the DPE assumptions after so many observations are removed. Therefore, I only report the results of OLS specifications for these two subsets. For all specifications in Table 11, the coefficient on qualitative similarity $WireSim_{ijt}$ remains positive and significant, even though coefficients on firm size, book-to-market ratio, momentum and industry change signs or become insignificant. Thus, qualitative similarity is still able to predict future comovement between firms that lack direct or indirect economic linkages.

## 8. Qualitative similarity and portfolio risk

Regardless of truncation scheme, estimation methodology or model specification, the relation between contemporaneous qualitative similarity and future stock return correlation has been positive and statistically significant within my sample period. To evaluate the economic significance of this finding, I test whether forecasts of rolling correlations based on qualitative similarity can reduce the out-of-sample volatility of a minimum-variance portfolio.

For most of my analysis, individual firm-pairs were randomly selected to produce a manageable subset of firm-pair-period observations. Therefore, data that is specific to certain combinations of firms is sometimes omitted. To produce minimum variance portfolio weights, however, I require predicted correlations for all possible combinations of firms within a particular subset. To create such a sample, I randomly select 500 firms that have eligible return observations on the last trading day of 2004. Should any of these firms later become ineligible through bankruptcy, acquisition or a decline in market capitalization, another company is randomly chosen as a replacement. The result is a randomly generated, but persistent, collection of firms from all industries and size deciles. For each six-month period from January 2005 to June 2014, I generate out-of-sample correlation forecasts based on all historical information available about the firms in this subset.

Aside from the sampling approach, two other elements of my empirical methodology must be altered to produce reliable out-of-sample correlation forecasts. First, the DPE is no longer practical because the appropriate lag structure changes for each rolling sample window. In the early years of the sample, when the estimation window consists of only a few six-month periods, I am unable to find a lag structure that can satisfy the DPE assumptions. Second, I am unable to include time fixed effects in my specifications because the out-of-sample time-specific shocks cannot be known in advance. Therefore, I proceed by forecasting daily and ten-day cumulative return correlations with specifications that are similar to Eq. (6), without the time series fixed effects $\alpha_{t+1}$. I also apply the cyclical coordinate descent algorithm for elastic net regression, developed by Friedman et al. (2010), to find sparse parameter estimates and improve the out-of-sample fit relative to OLS.

Table 12 reports regression coefficients averaged across each of the nineteen rolling sample windows. For daily and ten-day cumulative return correlations, I estimate two specifications that include $WireDum_{ijt}$, $TakeSim_{ijt}$ and $WireSim_{ijt}$ as explanatory variables, and two that do not. Newey–West (1987) $t$-statistics are reported below each average in parenthesis. Alpha describes the weight placed on the lasso norm, or one minus the weight placed on the ridge norm, and lambda represents the penalty placed on larger coefficients. As in Table 3, the coefficient averages for qualitative similarity, and most of the other independent variables, are positive and significant. However, both of the variables controlling for the firm-pair's book-to-market ratios lose their significance. Possibly due to the omission of time series fixed effects, the average adjusted R-squared is also much lower than what was reported in Table 3.

To generate portfolio weights for each six-month period, a covariance matrix is created based on the correlation forecasts produced by the regression estimates summarized in Table 12. The predicted covariance for period $t+1$ between the returns of firms

**Table 10**

Other sources of qualitative information. The dependent variable in all specifications is the Fisher transformation $z_{ijt+1}$ of the Pearson correlation $\rho_{ijt+1}$ calculated from the returns of firms $i$ and $j$ in excess of the risk-free rate for each six-month period $t+1$. Correlations are calculated from daily returns in Panel A, whereas Panel B reports estimates that are based on Carhart (1997) residuals. The binary variable $WireDum_{ijt}$ is set to 1 whenever both firms have some positive number of total words transmitted across the Reuters Integrated Data Network. $TakeSim_{ijt}$, defined in Eq. (4), accounts for how often firm-pairs are mentioned in the same newswire take. $WireSim_{ijt}$ is a measure of qualitative similarity defined in Eq. (2). A description for all other included variable calculations is provided in the surrounding text and in Panel D of Table A-1. Eq. (6) is estimated with ordinary least squares and Eq. (7) is estimated with a dynamic panel estimation (DPE) methodology. Ordinary least squares standard errors are clustered by firm-pair, both individual firms and time using the Cameron et al., (2011) multi-way clustering procedure. DPE results are generated using the approach described in Arellano and Bover (1995) and Blundell and Bond (1998) with bias-corrected robust variance-covariance estimates of the model parameters. Coefficients marked $^{*}$ and $^{**}$ are significant at the 5% and 1% level, respectively, and $t$-statistics are reported in parenthesis. All of the independent variables are used as predetermined instruments in the DPE specifications. "Systematic lags" refers to the total number of lags included in each specification for the variables $z_{ijt}$, $BetaDum_{ijt}$, $BetaCorr_{ijt}$, $SizeDum_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktDum_{ijt}$, $Bk/MktCorr_{ijt}$, $MomDum_{ijt}$, $MomCorr_{ijt}$, $IndDum_{ijt}$ and $IndCorr_{ijt}$. "Alternative Controls" refers to the inclusion of $\rho^{1mo}_{ijt}$, $\rho^{2mo}_{ijt}$, $AnaDum_{ijt}$, $AnaCorr_{ijt}$, $AmiDum_{ijt}$, $AmiCorr_{ijt}$, $SP500_{ijt}$, $SPVal_{ijt}$, $SPGrw_{ijt}$, $PrcDum_{ijt}$, $PrcCorr_{ijt}$ and $MSA_{ijt}$ as untabulated controls.

| | Ordinary least squares | | | | | | | | | | Dynamic panel estimator | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Broad sample | | | | | Larger firms | | | | | Broad sample | | | | |
| Panel A: Correlations from daily returns | | | | | | | | | | | | | | | |
| $z_{ijt}$ | 0.404** | 0.405** | 0.410** | 0.398** | 0.345** | 0.440** | 0.450** | 0.456** | 0.435** | 0.378** | 0.226** | 0.231** | 0.231** | 0.224** | 0.189** |
| | (22.43) | (22.36) | (22.75) | (22.40) | (22.86) | (23.58) | (23.84) | (24.27) | (23.21) | (24.09) | (145.4) | (148.1) | (147.9) | (146.8) | (117.2) |
| $S34Sim_{ijt}$ | −0.00908 | | | −0.00902 | −0.0161 | −0.0275** | | | −0.0273** | −0.0115 | 0.117** | | | 0.0956** | 0.0299** |
| | (−0.790) | | | (−0.802) | (−1.389) | (−2.893) | | | (−2.908) | (−1.075) | (20.74) | | | (17.73) | (7.775) |
| $S12Sim_{ijt}$ | 0.0491** | | | 0.0471** | 0.0472* | 0.0836** | | | 0.0813** | 0.0453** | −0.0353** | | | −0.0257** | 0.0114* |
| | (3.011) | | | (2.935) | (2.418) | (6.305) | | | (6.185) | (2.863) | (−6.710) | | | (−4.965) | (2.483) |
| $EPSSim_{ijt}$ | 0.179** | | | 0.126** | 0.121** | 0.208** | | | 0.163** | 0.164** | −0.0533** | | | −0.0581** | −0.0635** |
| | (13.95) | | | (9.201) | (9.331) | (15.85) | | | (11.98) | (12.17) | (−3.088) | | | (−3.519) | (−3.834) |
| $HobPhiDum_{ijt}$ | | 0.0326** | | 0.0244** | 0.0247** | | 0.0464** | | 0.0331** | 0.0331** | | 0.0136** | | 0.0137** | 0.0143** |
| | | (7.705) | | (5.761) | (5.806) | | (9.162) | | (6.248) | (6.064) | | (3.069) | | (3.117) | (3.252) |
| $HobPhiScr_{ijt}$ | | 0.566** | | 0.523** | 0.538** | | 0.469** | | 0.291** | 0.290** | | 0.320** | | 0.398** | 0.404** |
| | | (6.879) | | (6.373) | (6.652) | | (5.811) | | (3.552) | (3.630) | | (4.117) | | (5.173) | (5.215) |
| $SentPos_{ijt}$ | | | −0.00242 | −0.00280 | −0.00353* | | | 0.000874 | −0.000799 | −0.00106 | | | −0.00299** | −0.00276** | −0.00315** |
| | | | (−1.411) | (−1.857) | (−2.518) | | | (0.360) | (−0.334) | (−0.461) | | | (−6.295) | (−5.799) | (−6.664) |
| $SentNeg_{ijt}$ | | | −0.00439* | −0.00563** | −0.00618** | | | −0.00412* | −0.00609** | −0.00601** | | | −0.00354** | −0.00326** | −0.00328** |
| | | | (−2.478) | (−3.826) | (−4.280) | | | (−2.138) | (−3.400) | (−3.707) | | | (−8.060) | (−7.377) | (−7.478) |
| $TakeSim^{Scher}_{ijt}$ | | | 0.110** | 0.0310** | 0.0269* | | | 0.116** | 0.0203 | 0.0186 | | | −0.0398 | −0.0210 | −0.0179 |
| | | | (7.970) | (2.986) | (2.697) | | | (8.620) | (1.623) | (1.536) | | | (−1.926) | (−0.955) | (−0.811) |
| $WireSim_{ijt} \times SentPos_{ijt}$ | | | | 0.0158 | 0.00770 | | | | 0.0345 | 0.0295 | | | | 0.0135 | 0.0112 |
| | | | | (1.031) | (0.504) | | | | (1.369) | (1.202) | | | | (1.337) | (1.111) |
| $WireSim_{ijt} \times SentNeg_{ijt}$ | | | | 0.0119 | 0.0113 | | | | −0.00542 | −0.00578 | | | | 0.00787 | 0.00794 |
| | | | | (0.641) | (0.607) | | | | (−0.157) | (−0.172) | | | | (0.807) | (0.816) |
| $WireDum_{ijt}$ | | | | 0.00651 | 0.00562 | | | | 0.00633 | 0.00486 | | | | 0.00578** | 0.00640** |
| | | | | (1.839) | (1.597) | | | | (1.181) | (0.938) | | | | (5.328) | (6.005) |
| $TakeSim_{ijt}$ | | | | −0.0792 | −0.0863 | | | | −0.156* | −0.139* | | | | −0.276** | −0.305** |
| | | | | (−1.189) | (−1.300) | | | | (−2.478) | (−2.322) | | | | (−3.586) | (−3.949) |
| $WireSim_{ijt}$ | | | | 0.137** | 0.131** | | | | 0.134** | 0.128** | | | | 0.0304** | 0.0349** |
| | | | | (5.745) | (5.723) | | | | (3.510) | (3.478) | | | | (3.258) | (3.754) |
| Time fixed effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-pair and Firm-specific panel effects | No | No | No | No | No | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes |
| Alternative controls | No | No | No | No | Yes | No | No | No | No | Yes | No | No | No | No | Yes |
| Systematic lags | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 5 | 5 | 5 | 5 |
| Adjusted $R$-squared | 0.525 | 0.525 | 0.523 | 0.527 | 0.532 | 0.589 | 0.587 | 0.585 | 0.591 | 0.594 | | | | | |
| AR(2) test | | | | | | | | | | | −0.341 | −0.767 | −0.613 | −0.354 | 0.220 |
| Observations | 19,750,851 | | | | | 7,373,461 | | | | | 1,367,394 | | | | |
| Panel B: Correlations from daily Carhart residuals | | | | | | | | | | | | | | | |
| $z'_{ijt}$ | 0.140** | 0.144** | 0.152** | 0.135** | 0.131** | 0.185** | 0.198** | 0.211** | 0.179** | 0.172** | 0.0365** | 0.0366** | 0.0364** | 0.0367** | 0.0381** |
| | (12.51) | (13.11) | (13.30) | (12.35) | (11.34) | (12.62) | (13.61) | (14.08) | (12.42) | (12.01) | (34.84) | (34.95) | (34.70) | (35.08) | (34.44) |
| $S34Sim_{ijt}$ | −0.0162** | | | −0.0144** | −0.0168** | −0.0147** | | | −0.0126** | −0.0270** | −0.0104** | | | −0.00917* | −0.0116** |
| | (−6.397) | | | (−6.059) | (−6.740) | (−7.296) | | | (−6.126) | (−11.18) | (−2.737) | | | (−2.523) | (−3.856) |

(*continued on next page*)

**Table 10** (*continued*)

| | Ordinary least squares | | | | | | | | | | Dynamic panel estimator | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Broad sample | | | | | Larger firms | | | | | Broad sample | | | | |
| $S12Sim_{ijt}$ | 0.0326** | | | 0.0303** | 0.0309** | 0.0246** | | | 0.0215** | 0.0473** | 0.0397** | | | 0.0384** | 0.0420** |
| | (7.741) | | | (7.457) | (7.193) | (9.013) | | | (7.869) | (11.28) | (12.17) | | | (11.94) | (14.98) |
| $EPSSim_{ijt}$ | 0.288** | | | 0.241** | 0.239** | 0.317** | | | 0.265** | 0.264** | 0.0680** | | | 0.0806** | 0.0699** |
| | (25.25) | | | (21.60) | (21.58) | (25.67) | | | (21.80) | (21.75) | (5.651) | | | (6.718) | (5.910) |
| $HobPhiDum_{ijt}$ | | 0.0466** | | 0.0347** | 0.0345** | | 0.0600** | | 0.0414** | 0.0407** | | 0.0109** | | 0.0126** | 0.0129** |
| | | (14.12) | | (12.25) | (12.25) | | (9.657) | | (8.794) | (8.552) | | (3.735) | | (4.369) | (4.464) |
| $HobPhiScr_{ijt}$ | | 0.393** | | 0.326** | 0.325** | | 0.570** | | 0.308** | 0.305** | | 0.0917 | | 0.104 | 0.119* |
| | | (5.789) | | (5.296) | (5.319) | | (5.778) | | (4.312) | (4.246) | | (1.533) | | (1.781) | (2.050) |
| $SentPos_{ijt}$ | | | −0.000367 | 0.000519 | 0.000572* | | | −0.000459 | −0.000250 | 8.86e-06 | | | −0.000136 | 1.07e-05 | 2.39e-05 |
| | | | (−1.246) | (2.010) | (2.189) | | | (−0.629) | (−0.370) | (0.0131) | | | (−0.418) | (0.0329) | (0.0733) |
| $SentNeg_{ijt}$ | | | 7.09e-05 | 0.000485* | 0.000689** | | | −0.000303 | −0.000590 | 0.000101 | | | 9.86e-05 | 0.000254 | 0.000195 |
| | | | (0.278) | (2.201) | (3.052) | | | (−0.728) | (−1.395) | (0.233) | | | (0.326) | (0.833) | (0.638) |
| $TakeSim^{Scher}_{ijt}$ | | | 0.177** | 0.0491** | 0.0489** | | | 0.169** | 0.0252* | 0.0254* | | | −0.00602 | −0.00806 | −0.00796 |
| | | | (11.47) | (4.033) | (4.038) | | | (12.90) | (2.249) | (2.271) | | | (−0.422) | (− 0.553) | (− 0.552) |
| $WireSim_{ijt} \times SentPos_{ijt}$ | | | | 0.000323 | −0.000411 | | | | −0.0168 | −0.0171 | | | | 0.00289 | 0.00270 |
| | | | | (0.0619) | (−0.0809) | | | | (−1.365) | (−1.387) | | | | (0.415) | (0.388) |
| $WireSim_{ijt} \times SentNeg_{ijt}$ | | | | 0.00736 | 0.00723 | | | | 0.000734 | 0.00222 | | | | 0.00434 | 0.00306 |
| | | | | (1.292) | (1.286) | | | | (0.0728) | (0.228) | | | | (0.639) | (0.451) |
| $WireDum_{ijt}$ | | | | −0.00121* | −0.00122* | | | | −0.000799 | −0.000358 | | | | −0.00124 | −0.00139* |
| | | | | (−2.439) | (−2.467) | | | | (−0.918) | (−0.440) | | | | (− 1.809) | (− 2.042) |
| $TakeSim_{ijt}$ | | | | 0.232** | 0.238** | | | | 0.0620 | 0.0785 | | | | 0.0528 | 0.0281 |
| | | | | (3.262) | (3.358) | | | | (1.356) | (1.756) | | | | (0.996) | (0.526) |
| $WireSim_{ijt}$ | | | | 0.0525** | 0.0494** | | | | 0.101** | 0.0950** | | | | 0.00707* | 0.00708* |
| | | | | (9.683) | (9.159) | | | | (6.706) | (6.750) | | | | (1.977) | (2.008) |
| Time fixed effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-pair and firm-specific panel effects | No | No | No | No | No | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes |
| Alternative controls | No | No | No | No | Yes | No | No | No | No | Yes | No | No | No | No | Yes |
| Systematic lags | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Adjusted $R$-squared | 0.076 | 0.073 | 0.067 | 0.080 | 0.081 | 0.164 | 0.156 | 0.147 | 0.168 | 0.170 | | | | | |
| AR(2) test | | | | | | | | | | | −0.411 | −0.328 | −0.432 | −0.314 | −0.00829 |
| Observations | 19,750,851 | | | | | 7,373,461 | | | | | 1,698,561 | | | | |

**Table 11**

Investor inattention and economically linked firms. The dependent variable in all specifications is the Fisher transformation $z_{ijt+1}$ of the Pearson correlation $\rho_{ijt+1}$ calculated from the returns of firms $i$ and $j$ in excess of the risk-free rate for each six-month period $t+1$. The sample is truncated to remove observations where firms are linked directly or linked indirectly, through their respective direct linkages to some other firm. Firms that are mentioned in the same newswire take at least once during the sample period have "news story linkages," and firms with $HobPhiDum_{ijt} = 1$ at least once have "product market linkages." The binary variable $WireDum_{ijt}$ is set to 1 whenever both firms have some positive number of total words transmitted across the Reuters Integrated Data Network. $TakeSim_{ijt}$, defined in Eq. (4), accounts for how often firm-pairs are mentioned in the same newswire take. Document similarity $\widetilde{WireSim}_{ijt}$ is based on text transmitted across the Reuters Integrated Data Network. $WireSim_{ijt}$ is a measure of qualitative similarity defined in Eq. (2). A description for all other included variable calculations is provided in the surrounding text and in Table A-1. Eq. (6) is estimated with ordinary least squares and Eq. (7) is estimated with a dynamic panel estimation (DPE) methodology. Ordinary least squares standard errors are clustered by firm-pair, both individual firms and time using the Cameron et al. (2011) multi-way clustering procedure. DPE results are generated using the approach described in Arellano and Bover (1995) and Blundell and Bond (1998) with bias-corrected robust variance-covariance estimates of the model parameters. Coefficients marked * and ** are significant at the 5% and 1% level, respectively, and $t$-statistics are reported in parenthesis. All of the independent variables are used as predetermined instruments in the DPE specifications. "Systematic lags" refers to the total number of lags included in each specification for the variables $z_{ijt}$, $BetaDum_{ijt}$, $BetaCorr_{ijt}$, $SizeDum_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktDum_{ijt}$, $Bk/MktCorr_{ijt}$, $MomDum_{ijt}$, $MomCorr_{ijt}$, $IndDum_{ijt}$ and $IndCorr_{ijt}$.

| | Firms without news story linkages | | | | | | | | Firms without product market linkages | | | | Firms without news story or product market linkages | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Correlations from daily returns | | | | Correlations from 10-day returns | | | | Correlations from daily returns | | Correlations from 10-day returns | | Correlations from daily returns | | Correlations from 10-day returns | |
| | Ordinary least squares | | Dynamic panel estimator | | Ordinary least squares | | Dynamic panel estimator | | Ordinary least squares | | Ordinary least squares | | Ordinary least squares | | Ordinary least squares | |
| $z_{ijt}$ | 0.399** | 0.397** | 0.225** | 0.225** | 0.110** | 0.109** | 0.0465** | 0.0455** | 0.301** | 0.300** | 0.0912** | 0.0909** | 0.305** | 0.303** | 0.0876** | 0.0873** |
| | (21.42) | (21.65) | (145.0) | (145.7) | (10.84) | (10.87) | (32.95) | (32.40) | (7.587) | (7.622) | (7.985) | (8.006) | (7.617) | (7.654) | (7.751) | (7.765) |
| $BetaDum_{ijt}$ | 0.0731** | 0.0733** | 0.0350** | 0.0342** | 0.0900** | 0.0899** | 0.0530** | 0.0532** | 0.0621* | 0.0612* | 0.0651** | 0.0647** | 0.0650* | 0.0641* | 0.0645** | 0.0642** |
| | (3.850) | (3.841) | (15.79) | (15.45) | (5.120) | (5.085) | (19.49) | (19.56) | (2.515) | (2.511) | (3.092) | (3.066) | (2.708) | (2.701) | (3.170) | (3.148) |
| $BetaCorr_{ijt}$ | 0.0799** | 0.0801** | 0.0370** | 0.0357** | 0.0957** | 0.0956** | 0.0554** | 0.0553** | 0.0668* | 0.0658* | 0.0694** | 0.0690** | 0.0709* | 0.0698* | 0.0696** | 0.0692** |
| | (3.872) | (3.867) | (15.17) | (14.69) | (5.183) | (5.151) | (19.32) | (19.29) | (2.352) | (2.346) | (2.945) | (2.918) | (2.567) | (2.560) | (3.056) | (3.033) |
| $SizeDum_{ijt}$ | 0.172** | 0.164** | 0.0919** | 0.0908** | 0.177** | 0.169** | 0.0905** | 0.0907** | 0.153** | 0.147* | 0.232 | 0.228 | 0.177** | 0.172** | 0.230 | 0.227 |
| | (4.777) | (4.545) | (10.33) | (10.22) | (4.274) | (4.140) | (12.28) | (12.30) | (2.910) | (2.746) | (1.977) | (1.939) | (3.133) | (3.027) | (2.070) | (2.040) |
| $SizeCorr_{ijt}$ | 0.172** | 0.165** | 0.0893** | 0.0879** | 0.179** | 0.171** | 0.0812** | 0.0814** | 0.152* | 0.146* | 0.238 | 0.234 | 0.178** | 0.173** | 0.237 | 0.233 |
| | (4.724) | (4.489) | (9.690) | (9.540) | (4.305) | (4.171) | (10.51) | (10.54) | (2.756) | (2.601) | (1.914) | (1.879) | (3.023) | (2.922) | (2.016) | (1.988) |
| $Bk/MktDum_{ijt}$ | 0.0895** | 0.0873** | 0.0975** | 0.0972** | 0.0529* | 0.0513* | 0.0396** | 0.0392** | −0.0972 | −0.0956 | 0.0543 | 0.0551 | −0.0928 | −0.0915 | 0.0514 | 0.0520 |
| | (4.829) | (4.647) | (23.08) | (23.06) | (2.700) | (2.640) | (8.216) | (8.143) | (−0.698) | (−0.689) | (1.253) | (1.270) | (−0.661) | (−0.654) | (1.221) | (1.233) |
| $Bk/MktCorr_{ijt}$ | 0.0974** | 0.0951** | 0.107** | 0.106** | 0.0580* | 0.0564* | 0.0421** | 0.0416** | −0.107 | −0.105 | 0.0634 | 0.0644 | −0.103 | −0.101 | 0.0597 | 0.0604 |
| | (4.838) | (4.661) | (23.01) | (22.96) | (2.707) | (2.653) | (7.784) | (7.693) | (−0.695) | (−0.686) | (1.286) | (1.304) | (−0.660) | (−0.654) | (1.249) | (1.262) |
| $MomDum_{ijt}$ | 0.0717** | 0.0713** | 0.0510** | 0.0495** | 0.0707** | 0.0700** | 0.0341** | 0.0333** | −0.0267 | −0.0261 | 0.0271 | 0.0272 | −0.0251 | −0.0247 | 0.0237 | 0.0238 |
| | (3.948) | (3.964) | (25.38) | (24.66) | (3.893) | (3.892) | (10.91) | (10.65) | (−0.432) | (−0.426) | (0.953) | (0.961) | (−0.408) | (−0.404) | (0.886) | (0.892) |
| $MomCorr_{ijt}$ | 0.0790** | 0.0786** | 0.0557** | 0.0540** | 0.0785** | 0.0779** | 0.0396** | 0.0386** | −0.0327 | −0.0320 | 0.0312 | 0.0314 | −0.0300 | −0.0296 | 0.0282 | 0.0283 |
| | (3.785) | (3.801) | (25.16) | (24.38) | (3.766) | (3.768) | (11.49) | (11.20) | (−0.460) | (−0.453) | (0.914) | (0.924) | (−0.424) | (−0.420) | (0.876) | (0.883) |
| $IndDum_{ijt}$ | 0.0934** | 0.0892** | 0.0671** | 0.0709** | 0.124** | 0.118** | 0.198** | 0.197** | 0.0203 | 0.0184 | 0.0451** | 0.0433** | 0.0173 | 0.0154 | 0.0400* | 0.0383* |
| | (5.705) | (5.385) | (10.68) | (11.26) | (8.661) | (8.076) | (14.67) | (14.66) | (0.748) | (0.671) | (3.161) | (3.025) | (0.659) | (0.582) | (2.828) | (2.704) |
| $IndCorr_{ijt}$ | 0.0552* | 0.0532* | −0.0698** | −0.0699** | 0.0643** | 0.0626** | −0.0146** | −0.0143** | 0.0130 | 0.0118 | 0.0432* | 0.0425* | 0.0135 | 0.0121 | 0.0422* | 0.0415* |
| | (2.635) | (2.551) | (−32.20) | (−32.39) | (3.829) | (3.742) | (−6.395) | (−6.255) | (0.324) | (0.294) | (2.521) | (2.485) | (0.346) | (0.311) | (2.488) | (2.450) |
| $WireDum_{ijt}$ | | 0.00488 | | 0.00763** | | 0.00303 | | 0.00197 | | 0.0108 | | 0.00613 | | 0.0117 | | 0.00563 |
| | | (1.336) | | (7.764) | | (0.701) | | (0.980) | | (1.409) | | (0.892) | | (1.581) | | (0.841) |
| $TakeSim_{ijt}$ | | | | | | | | | | 0.163 | | 0.273 | | | | |
| | | | | | | | | | | (0.634) | | (0.901) | | | | |
| $WireSim_{ijt}$ | | 0.157** | | 0.0445** | | 0.228** | | 0.0145** | | 0.154** | | 0.167** | | 0.145** | | 0.156** |
| | | (6.536) | | (8.809) | | (6.300) | | (2.893) | | (4.095) | | (4.975) | | (4.242) | | (5.124) |
| Time fixed effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-pair and Firm-specific panel effects | No | No | Yes | Yes | No | No | Yes | Yes | No | No | No | No | No | No | No | No |
| Systematic lags | 1 | 1 | 4 | 4 | 1 | 1 | 4 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Adj. $R$-squared | 0.515 | 0.516 | | | 0.233 | 0.234 | | | 0.299 | 0.300 | 0.159 | 0.159 | 0.299 | 0.300 | 0.152 | 0.153 |
| AR(2) $T$est | | | 1.646 | 1.816 | | | −0.770 | −0.747 | | | | | | | | |
| Observations | 18,652,056 | | 1,336,491 | | 18,652,056 | | 1,336,491 | | 9,732,039 | | 9,732,039 | | 8,207,752 | | 8,207,752 | |

**Table 12**
Rolling correlation forecast regressions. The dependent variable in all specifications is the Fisher transformation $z_{ijt+1}$ of the Pearson correlation $\rho_{ijt+1}$ calculated from the returns of firms $i$ and $j$ in excess of the risk-free rate for each six-month period $t+1$. The binary variable $WireDum_{ijt}$ is set to 1 whenever both firms have some positive number of total words transmitted across the Reuters Integrated Data Network. $TakeSim_{ijt}$, defined in Eq. (4), accounts for how often firm-pairs are mentioned in the same newswire take. $WireSim_{ijt}$ is a measure of qualitative similarity defined in Eq. (2). A description for all other included variable calculations is provided in the surrounding text and in Table A-1. The cyclical coordinate descent algorithm for elastic net regression, developed by Friedman et al. (2010), is used to find sparse parameter estimates across nineteen rolling sample windows. Alpha describes the weight placed on the lasso norm, or one minus the weight placed on the ridge norm, and lambda represents the penalty placed on larger coefficients. Newey–West (1987) $t$-statistics are reported below each average in parenthesis. Coefficients marked $^{*}$,$^{**}$ and $^{***}$ are significant at the 10%, 5% and 1% level, respectively, and $t$-statistics are reported in parenthesis.

| | Correlations calculated from daily returns | | | | Correlations calculated from 10-day returns | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (1) | (2) | (3) | (4) |
| $z_{ijt}$ | 0.4585$^{***}$ | 0.4502$^{***}$ | 0.3975$^{***}$ | 0.3932$^{***}$ | 0.1623$^{***}$ | 0.1586$^{***}$ | 0.101$^{***}$ | 0.0997$^{***}$ |
| | (25.62) | (25.97) | (27.6) | (27.91) | (9.771) | (9.856) | (12.39) | (12.65) |
| $BetaDum_{ijt}$ | | | 0.1026$^{***}$ | 0.09812$^{***}$ | | | 0.1099$^{***}$ | 0.1068$^{***}$ |
| | | | (5.225) | (5.24) | | | (7.537) | (7.66) |
| $BetaCorr_{ijt}$ | | | 0.1142$^{***}$ | 0.1093$^{***}$ | | | 0.1209$^{***}$ | 0.1175$^{***}$ |
| | | | (5.053) | (5.062) | | | (7.296) | (7.422) |
| $SizeDum_{ijt}$ | | | 0.07082$^{***}$ | 0.06635$^{***}$ | | | 0.1913$^{***}$ | 0.187$^{***}$ |
| | | | (19.48) | (18.47) | | | (12.06) | (11.76) |
| $SizeCorr_{ijt}$ | | | 0.06624$^{***}$ | 0.0619$^{***}$ | | | 0.1901$^{***}$ | 0.1862$^{***}$ |
| | | | (15.6) | (14.75) | | | (11.18) | (10.87) |
| $Bk/MktDum_{ijt}$ | | | −0.008355 | −0.007532 | | | 0.003567 | 0.008392 |
| | | | (−1.144) | (−1.094) | | | (0.4821) | (1.253) |
| $Bk/MktCorr_{ijt}$ | | | −0.01138 | −0.01035 | | | 0.00175 | 0.007598 |
| | | | (−1.323) | (−1.273) | | | (0.2115) | (1.03) |
| $MomDum_{ijt}$ | | | 0.0224$^{***}$ | 0.02244$^{***}$ | | | 0.1336$^{***}$ | 0.1322$^{***}$ |
| | | | (5.617) | (5.607) | | | (4.59) | (4.627) |
| $MomCorr_{ijt}$ | | | 0.02175$^{***}$ | 0.02185$^{***}$ | | | 0.1537$^{***}$ | 0.1522$^{***}$ |
| | | | (4.683) | (4.673) | | | (4.558) | (4.6) |
| $IndDum_{ijt}$ | | | 0.125$^{***}$ | 0.1187$^{***}$ | | | 0.1822$^{***}$ | 0.1735$^{***}$ |
| | | | (74.25) | (53.32) | | | (95.52) | (69.58) |
| $IndCorr_{ijt}$ | | | 0.07436$^{***}$ | 0.07151$^{***}$ | | | 0.1075$^{***}$ | 0.1046$^{***}$ |
| | | | (33.92) | (34.49) | | | (41.57) | (42.68) |
| $WireDum_{ijt}$ | | 0.02478$^{***}$ | | 0.02449$^{***}$ | | 0.02928$^{***}$ | | 0.02829$^{***}$ |
| | | (4.939) | | (5.761) | | (3.455) | | (4.32) |
| $TakeSim_{ijt}$ | | 0.9531$^{***}$ | | 0.6539$^{***}$ | | 2.058$^{***}$ | | 1.522$^{***}$ |
| | | (15.49) | | (14.25) | | (14.9) | | (14.73) |
| $WireSim_{ijt}$ | | 0.2854$^{***}$ | | 0.211$^{***}$ | | 0.4501$^{***}$ | | 0.2951$^{***}$ |
| | | (23.19) | | (11.23) | | (35.83) | | (12.2) |
| Avg. alpha | 1 | 1 | 0.94 | 0.97 | 1 | 1 | 0.99 | 1 |
| Avg. lambda | 0.0001 | 0.0001 | 0.000105 | 0.000105 | 0.000055 | 0.000055 | 0.000055 | 0.000055 |
| Avg. cross-validation MSE | 0.02476 | 0.02457 | 0.02417 | 0.02404 | 0.09338 | 0.09296 | 0.09067 | 0.09043 |
| Avg. $R$-squared | 0.2041 | 0.2102 | 0.2243 | 0.2284 | 0.02724 | 0.03143 | 0.05496 | 0.0573 |

$i$ and $j$ is defined as:

$$\widehat{Cov}_{ijt+1} = \hat{\rho}_{ijt+1}\sigma_{it}\sigma_{jt} \tag{9}$$

where $\hat{\rho}_{ijt+1}$ is the correlation forecast and $\sigma_{it}$ and $\sigma_{jt}$ are the actual period $t$ return volatilities for each firm. Minimum-variance portfolio weights are found using the primal-dual interior point method for nonlinear optimization. To generate weights that would appear reasonable in an investment setting, I limit each holding in the portfolio to 1% of total value. In cases where short-sales are allowed, weights are also bounded below by − 1%. Given that the equal-weighted allocation to each of the 500 firms is 0.2%, these constraints still allow for considerable variability in the individual allocations.

Table 13 describes the out-of-sample performance of several portfolios from January of 2005 to June of 2014. While holdings are adjusted to the new minimum-variance weights at the beginning of each six-month period, the portfolios are not rebalanced within each period. Thus, performance results reported in Table 13 could have been achieved with minimal trading costs. For comparative purposes, performance statistics are also provided for other common portfolios. First, I use the actual sample covariance matrix in each forecast period $t+1$ to generate weights for the realized minimum-variance portfolio. The standard deviation for this portfolio represents what could be achieved with perfect foresight using the same constraints and optimization routine described above. Next, Bekaert et al. (2009) show that risk-based factor models capture the covariance structure of the international stock market better than a variety of alternatives. Therefore, I generate portfolio weights for period $t+1$ based on a covariance matrix of Carhart (1997) model predicted values calculated during the preceding six-month period $t$. To limit the impact of noise in the factor-model estimation, parameters are calculated with two full years of returns ending on the last trading day of period $t$. Finally, I also report the market- and equal-weighted portfolio performance, rebalanced once every six months, for my collection of 500 firms during the out-of-sample period.

For both panels of Table 13, I match the return frequency of the performance statistics with the return frequency of the correlation forecasts, and portfolios are sorted by standard deviation within each category. Several things stand out from the reported performance statistics. First, while the realized minimum-variance portfolios have much lower volatilities than any of those generated with predicted correlations, all of the forecasted portfolios have dramatically lower standard deviations than the market- and equal-weighted portfolios. Even when risk is measured by exposure to the market factor, or the range of out-of-sample returns, the forecasted portfolios still dominate the two passive strategies. Furthermore, the passive strategies offer no clear benefit in the form of higher returns. Thus, the market- and equal-weighted strategies appear to be suboptimal for most investors during this sample period.

**Table 13**
Out-of-sample portfolio performance. This table describes the out-of-sample performance of several portfolios from January of 2005 to June of 2014. Holdings are adjusted to the new minimum-variance weights at the beginning of each six-month period, and the portfolios are not rebalanced within each period. To generate portfolio weights for each six-month period, a covariance matrix is created based on the correlation forecasts (Spec (1), Spec (2), etc.) produced by the regression estimates summarized in Table 12. The “Realized Min Variance” portfolio is the minimum variance portfolio based on the actual sample covariance matrix calculated during the forecast period $t+1$. The “Carhart” portfolio is based on a covariance matrix of Carhart (1997) model predicted values estimated during the preceding six-month period $t$. “Market Weight” and “Equal Weight” represent the market- and equal-weighted portfolio performance of the sample firms during period $t+1$. The return frequency of the performance statistics is matched with the return frequency of the correlation forecasts, and portfolios are sorted by standard deviation within each category.

| | | Summary statistics | | | | | | Factor model coefficients | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Correlation forecast | Short sales allowed | Mean | Std dev | Skew | Kurt | Max | Min | Mkt | SMB | HML | UMD |
| Panel A: Correlations predicted from and statistics reported in daily returns | | | | | | | | | | | |
| Realized min variance | Yes | 0.033% | 0.706% | 0.691 | 39.40 | 8.28% | −6.85% | 0.219 | 0.016 | −0.008 | −0.280 |
| Spec (2) | Yes | 0.044% | 0.951% | 0.332 | 14.66 | 8.19% | −6.83% | 0.538 | −0.172 | −0.078 | −0.144 |
| Spec (1) | Yes | 0.041% | 0.955% | 0.315 | 15.45 | 8.24% | −6.81% | 0.538 | −0.169 | −0.071 | −0.159 |
| Spec (4) | Yes | 0.039% | 0.999% | 0.308 | 13.59 | 8.29% | −6.86% | 0.591 | −0.036 | −0.025 | −0.138 |
| Spec (3) | Yes | 0.039% | 1.001% | 0.298 | 13.57 | 8.39% | −6.76% | 0.593 | −0.034 | −0.025 | −0.143 |
| Carhart | Yes | 0.044% | 0.999% | 0.178 | 10.81 | 8.23% | −6.82% | 0.387 | −0.035 | −0.081 | −0.171 |
| Realized min variance | No | 0.045% | 0.990% | 0.087 | 12.04 | 8.25% | −6.68% | 0.619 | 0.175 | 0.075 | −0.124 |
| Spec (2) | No | 0.037% | 1.133% | 0.069 | 10.92 | 9.40% | −7.03% | 0.772 | 0.112 | 0.015 | −0.099 |
| Spec (1) | No | 0.037% | 1.136% | 0.081 | 11.08 | 9.62% | −7.07% | 0.773 | 0.113 | 0.014 | −0.101 |
| Spec (4) | No | 0.037% | 1.139% | 0.050 | 10.96 | 9.40% | −7.17% | 0.779 | 0.109 | 0.017 | −0.095 |
| Spec (3) | No | 0.038% | 1.146% | 0.066 | 11.17 | 9.68% | −7.22% | 0.781 | 0.113 | 0.017 | −0.101 |
| Carhart | No | 0.038% | 1.142% | −0.049 | 10.80 | 9.69% | −7.20% | 0.778 | 0.126 | 0.027 | −0.080 |
| Market weight | | 0.030% | 1.252% | −0.120 | 11.04 | 10.84% | −9.37% | 0.966 | −0.121 | 0.008 | 0.027 |
| Equal weight | | 0.044% | 1.512% | −0.169 | 5.74 | 9.09% | −10.35% | 1.006 | 0.553 | 0.175 | −0.071 |
| Panel B: Correlations predicted from and statistics reported in 10-day cumulative returns | | | | | | | | | | | |
| Realized min variance | Yes | 0.235% | 1.635% | 0.539 | 22.57 | 11.68% | −11.08% | 0.184 | 0.045 | 0.127 | −0.192 |
| Spec (2) | Yes | 0.268% | 2.461% | −0.650 | 11.10 | 11.70% | −16.36% | 0.577 | −0.167 | −0.006 | −0.091 |
| Spec (1) | Yes | 0.281% | 2.534% | −0.814 | 10.25 | 11.50% | −16.98% | 0.620 | −0.227 | −0.008 | −0.054 |
| Spec (4) | Yes | 0.260% | 2.693% | −1.032 | 12.41 | 12.46% | −19.34% | 0.665 | 0.003 | −0.017 | −0.087 |
| Spec (3) | Yes | 0.276% | 2.704% | −0.961 | 11.48 | 12.51% | −18.76% | 0.656 | 0.014 | −0.008 | −0.092 |
| Carhart | Yes | 0.278% | 2.463% | −0.854 | 11.62 | 11.72% | −16.84% | 0.492 | −0.002 | 0.074 | −0.123 |
| Realized min variance | No | 0.332% | 2.255% | −0.778 | 10.86 | 11.59% | −13.83% | 0.520 | 0.129 | 0.155 | −0.085 |
| Spec (4) | No | 0.266% | 2.848% | −1.152 | 12.39 | 14.40% | −20.06% | 0.749 | 0.131 | 0.085 | −0.073 |
| Spec (3) | No | 0.261% | 2.852% | −1.163 | 12.40 | 14.49% | −20.06% | 0.751 | 0.133 | 0.083 | −0.074 |
| Spec (2) | No | 0.260% | 2.852% | −1.171 | 12.80 | 14.51% | −20.39% | 0.752 | 0.120 | 0.075 | −0.077 |
| Spec (1) | No | 0.263% | 2.853% | −1.152 | 12.50 | 14.30% | −20.19% | 0.749 | 0.125 | 0.080 | −0.077 |
| Carhart | No | 0.278% | 2.879% | −1.369 | 12.41 | 13.35% | −20.31% | 0.751 | 0.190 | 0.140 | −0.039 |
| Market weight | | 0.206% | 3.065% | −1.269 | 14.23 | 18.28% | −22.11% | 0.937 | −0.124 | 0.014 | 0.030 |
| Equal weight | | 0.299% | 3.733% | −0.887 | 9.04 | 20.52% | −22.18% | 0.967 | 0.574 | 0.199 | −0.074 |

Next, adding independent variables to the correlation regressions does not typically improve out-of-sample portfolio performance. Standard deviations for Specifications 1 and 2 are lower than 3 and 4 in all cases except when correlations are calculated in 10-day cumulative returns and short sales are forbidden. Moreover, the Carhart portfolio has the highest standard deviation in all four cases. Finally, the best performing portfolio in three of the four cases is Specification 2, which only includes a contemporaneous correlation and the three newswire measures as explanatory variables. Furthermore, the performance disparity between Specifications 2 and 1, and 4 and 3, imply that these newswire variables improve out-of-sample correlation forecasts in all cases. Therefore, incorporating the qualitative similarity of firm-specific newswire text into covariance predictions can reduce the out-of-sample volatility of a minimum-variance portfolio.

## 9. Closing remarks

In this article, I introduce a novel approach for quantifying a firm’s flow of information and use the cross-firm similarity of this measure to predict future price comovement. Commonality in information flows is gauged by the textual similarity of firm-specific content appearing on the Reuters Integrated Data Network from 2003 to 2013. This measure of qualitative similarity predicts an economically meaningful portion of future return correlation after controlling for numerous alternative explanations of comovement that have been suggested in prior literature. Previous research shows that newswire text is informative about future stock returns. My paper is the first to show that this type of qualitative information also predicts price comovement.

The time series of a firm’s stock returns are the single-dimensional output of a pricing function containing a broad range of inputs. Because this function evolves over time, the influence of inputs relevant to future prices may not be present in distant historical return series. The newswire text written about a firm describes these relevant inputs. Both in research and in practice, estimating the market correlation structure has sensibly relied on a lengthy historical times series. The depth of this qualitative information can amend the shortcomings of using distant historical prices for predicting comovement. Quantifying these inputs provides an opportunity to predict only the comovement that is implied by the contemporaneous pricing function. Thus, the approach introduced in my paper can produce estimates of future return correlation that do not require, or significantly benefit from, an abundant individual price history.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jbankfin.2018.04.010.

## References

Ahern, K.R., Sosyura, D., 2014. Who writes the news? Corporate press releases during merger negotiations. J. Financ. LXIX (1), 241–291.

Amihud, Y., Mendelson, H., Lauterbachc, B., 1997. Market microstructure and securities values: evidence from the Tel Aviv Stock exchange. J. Financ. Econ. 45, 365–390.

Anton, M., Polk, C., 2014. Connected stocks. J. Financ. LXIX (3), 1099–1127.

Arellano, M., Bond, S., 1991. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. Rev. Econ. Stud. 58 (2), 277–297.

Arellano, M., Bover, O., 1995. Another look at the instrumental variable estimation of error-components models. J. Econom. 68 (1), 29–51.

Asness, C.S., Moskowitz, T.J., Pedersen, L.H., 2013. Value and momentum everywhere. J. Financ. LXVIII (3), 929–985.

Barber, B.M., Loeffler, D., 1993. The "Dartboard" column: second-hand information and price pressure. J. Financ. Quant. Anal. 28 (2), 273–284.

Barberis, N., Schleifer, A., 2003. Style investing. J. Financ. Econ. 75, 161–200.

Barberis, N., Shleifer, A., Wurgler, J., 2005. Comovement. J. Financ. Econ. 75, 283–317.

Bekaert, G., Campbell, R.H., Angela, N., 2005. Market integration and contagion. J. Bus. 78, 39–79.

Bekaert, G., Robert, J.H., Xiaoyan, Z., 2009. International stock return comovements. J. Financ. LXIV (6), 2591–2627.

Bilisoly, R., 2008. Practical Text Mining with Perl. John Wiley & Sons. Inc, Hoboken, New Jersey.

Blundell, R., Bond, S., 1998. Initial conditions and moment restrictions in dynamic panel data models. J. Econom. 87 (1), 115–143.

Box, T., Shang, D., 2018. Information Driven Stock Price Comovement. Working paper.

Box, T., Davis, R., Hill, M., Lawrey, C., 2018. Operating performance and aggressive trade credit policies. J. Bank. Financ. 89, 192–208.

Boyer, B.H., 2011. Style-related comovement: fundamentals or labels. J. Financ. LXVI (1), 307–332.

Brandt, M.W., Brav, A., Graham, J.R., Kumar, A., 2010. The idiosyncratic volatility puzzle: time trend of specualtive episode. Rev. Financ. Stud. 23 (2), 865–899.

Cameron, A.C., Gelbach, J.B., Miller, D.L., 2011. Robust inference with multi-way clustering. J. Bus. Econ. Stat. 29 (2), 238–249.

Campbell, J.Y., Lettau, M., Malkiel, B.G., Xu, Y., 2001. Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. J. Financ. LVI (1), 1–43.

Cao, J., Chordia, T., Lin, C., 2016. Alliances and return predictability. J. Financ. Quant. Anal. 51 (5), 1689–1717.

Carhart, M.M., 1997. On persistence in mutual fund performance. J. Financ. LII (1), 57–82.

Chen, H., Chen, S., Li, F., 2012. Empirical Investigation of an Equity Pairs Trading Strategy. Working paper.

Cohen, L., Frazzini, A., 2008. Economic links and predictable returns. J. Financ. LXII (4), 1977–2011.

Cohen, L., Lou, D., 2012. Complicated firms. J. Financ. Econ. 104 (2), 383–400.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? Rev. Financ. Stud. 22 (5), 1915–1953.

Fang, L., Peress, J., 2009. Media coverage and the cross-section of stock returns. J. Financ. LXIV (5), 2023–2052.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 23 (1), 1–22.

Green, T.C., Hwang, B.-H., 2009. Price-based return comovement. J. Financ. Econ. 93, 37–50.

Hoberg, G., Phillips, G., 2010b. Dynamic Text-Based Industries and Endogenous Product Differentiation. National Bureau of Economic Research, Inc *NBER Working Papers 15991*.

Hoberg, G., and G. Phillips. 2015c. *Hoberg-Phillips Industry Classification Library*. Accessed 11 6, 2015. http://cwis.usc.edu/projects/industrydata/industryclass.htm.

Hoberg, G., Phillips, G., 2010a. Product market synergies and competition in mergers and acquisitions: a text-based analysis. Rev. Financ. Stud. 23 (10), 3773–3811.

Irvine, P.J., Pontiff, J., 2009. Idiosyncratic return volatility, cash flows, and product markets. Rev. Financ. Stud. 22 (3), 1149–1177.

Israelsen, R.D., 2015. Does common analyst coverage explain excess comovement? J. Financ. Quant. Anal 51 (4), 1193–1229.

Kleinbaum, A.M., Stuart, T.E., Tushman, MichaelL., 2013. Discretion within constraint: homophily and structure in a formal organization. Organ. Sci. 24 (5), 1316–1336.

Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empir. Financ. 10, 603–621.

Menzly, L., Ozbas, O., 2010. Market segmentation and cross-predictability of returns. J. Financ. LXV (4), 1555–1580.

Muslu, V., Rebello, M., Xu, Y., 2014. Sell-side analyst research and stock comovement. J. Account. Res. 52, 911–954.

Peress, J., 2014. The media and the diffusion of information in financial markets: evidence from newspaper strikes. J. Financ. LXIX (5), 2007–2043.

Pindyck, R.S., Rotemberg, J., 1993. The comovement of stock prices. Q. J. Econ. 108 (4), 1073–1104.

Pirinsky, C., Wang, Q., 2006. Does corporate headquarters location matter for stock returns. J. Financ. LXI (4), 1991–2015.

Scherbina, A., Schlusche, B., 2015. Economic Linkages Inferred from News Stories and the Predictability of Stock Returns. Working paper.

Tetlock, P.C., 2011. All the News that's fit to reprint: do investors react to stale information. Rev. Financ. Stud. 24, 1481–1512.

Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. J. Financ. LXII (3), 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. J. Financ. LXIII (3), 1437–1467.

Windmeijer, F., 2005. A finite sample correction for the variance of linear efficient two-step GMM estimators. J. Econom. 126, 25–51.

Wintoki, M.B., Linck, J.S., Netter, J.M., 2012. Endogeneity and the dynamics of internal corporate governance. J. Financ. Econ. 105, 581–606.