Information Driven Stock Price Comovement

*Travis Box, Danjue Shang**

ABSTRACT

In this paper we examine the process by which qualitative information gets to drive stock price comovement. We start with the examination of information production, and subsequently investigate what type of information that has been produced is consumed by investors and eventually gets capitalized into stock price. We find that investors' information consumption varies across market conditions and firm characteristics. Consistent with the predictions of the information driven comovement hypothesis (Veldkamp 2006a), we find that stock price comovement is stronger when investors consume qualitative information about firms whose payoffs covary strongly with many others. Furthermore, as aggregate correlation falls, so does the demand for these high covariance signals. Our findings imply that investor information consumption choices are shaped by a market for information, and that these choices can drive excessive stock price comovement.

JEL classification: C33, C53, D83, G00, G11, G12, G14

Keywords: Comovement, Information Consumption, Textual Analysis, Correlation

* Travis Box is an Assistant Professor of Finance at University of Mississippi (tbox@bus.olemiss.edu, 340 Holman, P.O. Box 1848, University, MS 38677-1848, Office: (662) 915-2553, Fax: (662) 915-5821), and Danjue Shang is an Assistant Professor of Finance at Utah State University (danjue.shang@usu.edu, 3565 Old Main Hill, Logan, UT 84322, Office: (435)797-6379). We thank the University of Mississippi and Utah State University for their research support. We are grateful for the helpful comments of Eric Kelley, Andrew Lynch, Ryan Davis, Robert Van Ness, Geert Bekaert, Kenneth Singleton, and George Jiang. We 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 our own.

I. Introduction

The process by which investors make information consumption choices is poorly understood, but critical to the functioning of financial markets. These consumption decisions are necessary because individual investors cannot keep pace with the combined volume of press releases, regulatory filings and news reports from more than just a few firms. In this paper, we investigate which types of information are consumed and incorporated into asset prices by observing changes in the relation between firm-pair stock return correlation and the similarity of their qualitative information.

Liberti and Petersen (2017) describe how hard information, which is often recorded quantitatively, and soft information, which is often communicated as text, can both be applied to financial market decisions. When quantitative information is collected by one person and transmitted to another, both people know exactly the same thing. This characteristic of hard information makes it possible to delegate the collection of quantitative data to someone other than the investor. Soft information, however, is often more difficult to code and catalog for future use. An individual charged with collecting qualitative data may not know which parts are relevant until much later. They can recall the collected information when confronted with an investment decision, but it is only then that it becomes apparent how the qualitative data will be useful. For this reason, soft information must be collected in person by the same individual that is responsible for making the investment decision. It is this characteristic of qualitative information and investor's limited attention that we exploit to identify and study the variation in the type and quantity of information consumed by equity market investors and capitalized into market prices.

More specifically, we study how investors' information consumption drives the time-variant stock price comovement. Understanding what drives asset price comovement is important for both policy makers and investors. The field of finance has identified a variety of individual characteristics that, when shared across firms, might predict comovement in their equity returns. Many of these characteristics, such as as firm beta (Ledoit and Wolf 2003), size (Pindyck and Rotemberg 1993), book-to-market (Bekaert, Hodrick and Zhang 2009), momentum (Asness, Moskowitz and Pedersen 2013) and industry ((Campbell, et al. 2001), (Irvine and Pontiff 2009), and (Brandt, et al. 2010)), are measured quantitatively and easily disseminated to investors. The comovement predictors we study in this paper, the textual similarity in newswire content (Box 2018), are based on qualitative information that cannot be easily categorized and transmitted to other investors. Before reading the newswire text related to a particular firm, an investor does not understand how the qualitative information collected from this content will be similar to the text they read previously about other companies. Recognizing the similarity in newswire content is difficult when part of the information is collected by another individual. Therefore, analyzing which types of qualitative information get to be capitalized in financial markets and drive stock price comovement provides us with a unique opportunity to study the implied information consumption process by investors.

We postulate that information only gets capitalized into stock prices when it is consumed by investors. The correlation between the stock returns of a firm pair more strongly reflects the similarity in qualitative information between the firm pair when more investors consume the qualitative information about both firms and recognize the qualitative similarity shared by the said pair. By measuring changes in the relation between firm-pair stock return correlation and the similarity of their qualitative information across market conditions and

firm characteristics, we are able to determine which types of information are capitalized in stock prices, and in turn, infer when and which types of information investors consume. We find that the inferred information consumption expands with aggregate market capitalization and uncertainty; the consumption of firm-specific information increases with individual stock return and payoff volatility. Overall, market-wide correlations are higher when many investors consume qualitative information about firms with higher average measures of textual similarity. Furthermore, as aggregate market price correlation falls, so does the demand for these high covariance signals.

Our findings shed lights on the empirical implications of the information driven comovement hypothesis by Veldkamp (2006). In her model, the value of a signal is determined by its ability to reduce total payoff variance, where total payoff variance depends on risk and the value of the asset at risk. With regards to risk, asset-specific information becomes more valuable as security payoffs become less predictable.[1] Likewise, demand for asset-specific information increases whenever the asset comprises a larger share of the average investor's portfolio. These same predictions also apply to aggregate information consumption. In times of uncertainty, the marginal benefit of observing additional signals rises, causing market-wide information consumption to increase. Similarly, when the total value of an asset rises, investors must hold that additional asset value for the asset market to clear. Therefore, aggregate demand for information should increase when many assets are highly valued.[2]

[1] The investor attention models developed by Peng and Xiong (2006), Mondria (2010) and Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016) suggest a similar relation between payoff variance and information processing.

[2] In models where incomplete information is motivated by limited attention, as opposed to costly information, aggregate information consumption is usually determined by a fixed processing capacity. Andrei and Hasler (2015) model the relation between attention to news, return volatility, and risk premia, but they avoid providing a

Note that Veldkamp (2006) is motivated by the observation that information is fundamentally distinct from other goods because of its high fixed cost of production and near-zero cost of replication. This information production technology, coupled with free entry in the information market, can create a strategic complementarity in information acquisition that works through the market price for information when the aggregate information consumption level is low. In that scenario, information producers will focus on producing highly demanded payoff signals that are indicative of many firms and investors will coordinate on these signals due to the lower price; then excess comovement ensues. In contrary, a strategic substitutability in information acquisition arises when the aggregate information consumption level is high enough, i.e., investors still choose to consume more "unique" information, the kind about firms that are less interconnected with many other firms. Without controlling for market states that influence the demand for information, we find that equity investors consume less qualitative information about companies whose payoffs covary strongly with most other assets. This result implies that the market's aggregate level of information consumption is usually high enough to support a strategic substitutability in information acquisition. However, we also find that coordination on high covariance signals becomes more common whenever market-wide return correlations increase.

Theories of investor information choice have been unable to achieve broad acceptance because they are difficult to analyze without reliable quantitative measures describing investor information sets. Certain implications of the information driven comovement hypothesis have been tested previously by examining changes in the production of information (Brockman, Liebenberg, and Schutte (2010) and Hameed, Morck, Shen, and Yeung (2015)). However, our

theoretical foundation for fluctuating attention. Andrei and Hasler (2016) investigate a costly attention allocation decision. But, with just one risky asset their model is silent on comovement.

paper is the first to demonstrate empirically that the consumption of information is determined by firm-specific characteristics and ambient market conditions, and how such consumption forms the dynamic connection between similarity in information signals and stock price comovement.

Our results are also in line with the behavioral finance and psychology studies that explore human attention; more specifically, when investors pay their limited attention to what type of information. Sicherman et al. (2016) document that investors avoid looking at their portfolios after market decline. Da et al. (2011) find that firm size can still partially explain the variation in investor attention, proxied by search frequency in Google, even after controlling for characteristic-adjusted return, trading volume, news variables, and advertising expenses. Consistently, there is the psychological remark that human beings have the tendency to avoid information when the reality is discouraging (investment loss), [3] and are more willing to absorb information when the reality is pleasant (investment gain).

The rest of the paper is organized as follows. Section II discusses our datasets and the construction of our main variable of interest, the news similarity measure. Section III presents the empirical analysis, and discusses the empirical results in line with the information driven comovement hypothesis by Veldkamp (2006). The final section concludes the paper.

II. Sample description and newswire similarity measures

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. We calculate the NYSE price and size decile breakboints each six-month period from January 2003 to December

[3] See Sweeny, Melnyk, Miller, and Shepperd (2010) for a comprehensive review of the information avoidance literature in psychology.

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 1,982 firms at the beginning of each period with 2,723 unique firms appearing in at least one interval.

For our analysis, the data for the textual similarity of newswire content is from the Thomson Reuters NewsScope Archive. The Archive is derived from the Reuters Integrated Data Network (IDN) newswire feed and consists of the message stream which communicates text produced by *Reuters News* and select third party providers directly to client workstations.

A. Term-document matrix

Our approach to calculating the textual similarity of newswire text is identical to the process described in Box (2018). The basic object of our 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.

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.[4] The field

[4] 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

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.[5]

B. Similarity of qualitative information

The term-document matrix itself can be thought of as the raw quantitative data for our analysis. However, to compare the qualitative information about different firms, the similarity of their newswire content must be computed explicitly. First, we calculate the cosine similarity, $\widetilde{WireSim}_{ijt}$, between the firm vectors i and j in the term-document matrix for period t. Following Hoberg and Phillips (2010a) and (2010b), the elements of these term-document vectors consist only of 1's and 0's to indicate whether or not a firm document contains a particular word. Next, firms with at least some relevant text are classified into deciles based on total word counts during each 6-month period in the sample. The variable $\overline{WireSim_{ijt}}$ represents the average document similarity between firms appearing in the same word count deciles as i

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.

[5] 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.

and j during period t. Finally, the similarity of qualitative information, $WireSim_{ijt}$ is calculated by subtracting $\overline{WireSim_{ijt}}$ from $\widetilde{WireSim_{ijt}}$.

C. *Content attributions*

The Thomson Reuters NewsScope Archive also describes the attribution, or source, of each story. There are a total of 12 attributions contributing relevant text to our sample, however, only *Reuters News* consists primarily of content produced by journalists. Other attributions, such as *Business Wire* or *PR Newswire*, distribute content generated by the firms themselves in the form of press releases, legal disclosures and regulatory filings. While individuals are likely to base investment decisions on text produced by both companies and journalists, firm-generated content more accurately reflects the universe of primary sources available in the market.[6] Nevertheless, special attention is still given to text generated by *Reuters News* during certain parts of our analysis. $WireSim_{ijt}^{all}$ represents the similarity of firm-specific content drawn from all attributions appearing on the IDN, whereas $WireSim_{ijt}^{rtrs}$ and $WireSim_{ijt}^{firm}$ describe the similarity of qualitative information attributed to *Reuters* News and all other sources, respectively.

III. Empirical analysis

The subsequent analysis will attempt to answer two economic questions. First, do information producers focus their efforts on firms whose payoffs covary most strongly with other companies? If journalists and analysts process more information about firms whose qualitative information is similar to most other companies, this would imply that information producers

[6] The information contained in journalist-generated content is likely to have been collected from primary sources that are produced by the firms themselves.

provide the type of signals capable of generating comovement. Second, can information consumption choices help us understand the origins of comovement? If investors cluster their information demand on a few signals that predict the values of many companies, price comovement will be high relative to the covariance of underlying fundamentals.

A. Information production

Our analysis begins with an examination of information production. By studying the output of analysts and journalists, we investigate whether profit-motivated information producers focus their efforts on firms whose payoffs covary most strongly with others. Using the correlation in past accounting profits to measure payoff covariance, Hameed, et al. (2015) provide evidence that equity analysts disproportionately follow firms whose historical earnings are most similar to many other companies'.[7] Box (2018) provides evidence that newswire similarity predicts how future dividends and capital gains are correlated. Thus, we propose an alternative measure of payoff covariance based on each firm's average level of newswire similarity. Specifically, we calculate firm i's average newswire similarity with all other firms j:

$$\overline{WireSim_{it}^{all}} = \frac{1}{N-1}\sum_{j\neq i} WireSim_{ijt}^{all}, \quad (1)$$

where N is the number of firms with some positive volume of text appearing on the IDN during period t.

The information driven comovement hypothesis also predicts that asset-specific information becomes more valuable as the security's payoffs become less predictable or as the security comprises a larger share of the average investor's portfolio. To measure average portfolio share, market capitalizations are calculated on the final trading day of each 6-month period, and every

[7] Fang and Peress (2009) find that journalists cluster their coverage on large firms, but they do not test whether payoff covariance is a determinant of media following.

firm i is included in a NYSE size decile, $SizeDec_{it}$, for the following period t. Payoff predictability is approximated by the firm's daily stock return standard deviation, σ_{it}.

The level of information production is measured by word count and analyst following. $WrdCnt_{it}^{rtrs}$ is the total number of words written about the firm and distributed by *Reuters News* during the 6-month span t, and $WrdCnt_{it}^{firm}$ is the total number of words contributed by all other attributions. Thus, the former applies to content produced by journalists, while the latter measures content generated by the companies themselves. With a median of 89 and an average of 560 total words, the summary statitistics reported in Table I reaffirm that the bulk of journalist coverage is focused on a very small number of companies. The number of unique analysts with an earnings prediction recorded in the I/B/E/S database during period t is represented by $AnaNum_{it}$. Compared to journalists, analysts follow a much broader universe of firms. In a typical 6-month period, 83% of the companies in our sample have an analyst earnings prediction, but only 62% have a positive quantity of text produced by *Reuters News*.

The information driven comovement hypothesis predicts that profit-motivated information producers focus their efforts on larger and more volatile firms and, given sufficiently low levels of aggregate information consumption Λ, companies whose payoffs covary most strongly with others. These predictions motivate the following model:

$$Dep_{it+1} = \beta_0 + \beta_1 SizeDec_{it} + \beta_2 \sigma_{it} + \beta_3 WireDum_{it}^{all} + \beta_4 \overline{WireSim_{it}^{all}} + \beta_5 \overline{\rho_{it}} + \sum_{k=6}^{7} \beta_k Other_{kit} + \sum_{k=8}^{K} \beta_k Control_{kit} + \alpha_{t+1} + \varepsilon_{it+1}, \quad (2)$$

where α_{t+1} is a fixed effect for each 6-month span. The variable Dep_{it+1} will be some measure of information production, $WrdCnt_{it+1}^{firm}$, $WrdCnt_{it+1}^{rtrs}$ or $AnaNum_{it+1}$, depending on the specification. Equation (2) suggests that content producers determine their coverage during period $t+1$ after observing individual firm characteristics during period t. The binary variable

$WireDum_{it}^{all}$ indicates whether the firm has some positive volume of text appearing on the IDN. This variable is necessary to differentiate when contemporaneous average newswire similarity, $\overline{WireSim_{it}^{all}}$, is 0 because the firm's qualitative information not excessively similar or dissimilar to most other firms, or because it was never mentioned on the newswire.

The information driven comovement hypothesis predicts that the coefficients β_1 and β_2 should be positive when the dependent variable is either measure of profit-motivated information production, $WrdCnt_{it+1}^{rtrs}$ or $AnaNum_{it+1}$. The coefficient β_4 will also be positive if there is a strategic complementarity in content generation. Controlling for $\overline{\rho_{it}}$, the average Pearson retrun correlation over all firms $j \neq i$, ensures that the relation between information production and average newswire similarity, $\overline{WireSim_{it}^{all}}$, does not result from historical comovement. To determine whether different types of information producers influence each other, contemporaneous observations of each of the other two production measures, $Other_{kit}$, are also included in each specification. A description for all other included controls, $Control_{kit}$, is provided in Panel B of Table A-1.

The distributions of all three information production variables are described in Figure 1. Any summation of word count or analyst following is obviously bounded below by 0, but Figure 1 demonstrates that a large portion of the pooled sample is also clustered at this bound for each variable. Moreover, even when information production is positive, realized values are still confined to a discrete set of integers. The simplest framework for analyzing counted data is the Poisson regression model (Cameron and Trivedi 2013),[8] however, an important limitation of the Poisson distribution is that the conditional variance is assumed to equal the conditional mean.

[8] Ordinary least squares estimation of Equation (2) assumes that the regression errors ε_{it+1} follow a normal distribution. This assumption is not appropriate when the left-hand side variables are limited to nonnegative integer values.

According to Table I, this assumption might be inappropriate because the unconditional variance of each information production variable is much larger than its sample mean.

A negative binomial distribution should be specified in cases where the variances derived from the data are higher than their conditional means (Gardner, Mulvey and Shaw 1995). Unlike the Poisson distribution, which is fully characterized by one parameter, the negative binomial distribution is a function of both its mean and a measure of overdispersion. Adapting Equation (2) to this framework gives:

$$
\begin{aligned}
&Dep_{it+1} \sim \text{Poisson}(\mu_{it+1}) \\
&\mu_{it+1} = \beta_0 + \beta_1 SizeDec_{it} + \beta_2 \sigma_{it} + \beta_3 WireDum_{it}^{all} + \beta_4 \overline{WireSim_{it}^{all}} \\
&\qquad\qquad + \beta_5 \overline{\rho_{it}} + \sum_{k=6}^{7} \beta_k Other_{kit} + \sum_{k=8}^{K} \beta_k Control_{kit} + \alpha_{t+1} + \nu_{it+1} \\
&e^{\nu_{it+1}} \sim Gamma\left({}^{1}/_{disp_{it+1}}, disp_{it+1}\right).
\end{aligned}
\tag{3}
$$

Equation (3) stipulates that the number of words written about, and the number of analysts following, firm i during period $t+1$ is a negative binomial random variable with mean μ_{it+1} and dispersion parameter $disp_{it+1}$.[9]

Word counts and analyst following are observed over time, so our analysis must account for the correlation between repeated measures of information production related to the same firm. Companies that are covered by analysts and the financial press, during period t are also likely to be covered during period $t+1$. The generalized estimating equations approach introduced by Liang and Zeger (1986) specifies how the average of a response variable, $\bar{\mu}$, adjusts to changes in the independent variables while allowing for correlation between repeated measurements on the

[9] When the overdispersion parameter is 0, the negative binomial distribution becomes the Poisson distribution. Equation (2) is estimated with a Poisson and a negative binomial regression on the pooled sample of observations. For all three information production variables, a likelihood ratio test strongly rejects strongly rejects the null hypothesis that the overdispersion parameter is 0.

same individual over time. Parameters from this method of estimation have a population average interpretation. For every unit increase in an independent variable across the population, generalized estimating equations reveal how much the average response $\bar{\mu}$ would change (Ballinger 2004).[10]

The results from estimating Equations (3) are reported in the first three columns of Table II with standard errors clustered by firm. If firm-generated content is often related to required disclosures, then $WrdCnt_{it}^{firm}$ measures the output volume of a primary source that is not determined by a market for information. Table II confirms that future firm-generated text volume is not positively associated with stock return volatility or average newswire similarity. However, companies that move into a higher size decile during period t subsequently increase their self-generated word count by 34%.[11] It is not possible to determine from Table II whether larger firms produce more content because of higher investor information demand or more arduous disclosure requirements. An increase in contemporaneous analyst following also predicts future firm-generated volume, but the economic impact is small. There is no similar relation between contemporaneous journalist output and future firm-generated text volume.

[10] The generalized estimating equations model specifies only the conditional mean μ_{it+1} and treats the correlation structure as a nuisance parameter (Gardiner, Luo, and Roman (2009) and Hardin and Hilbe (2013)). The algebraic form of the correlation structure is specified by the researcher through a working correlation matrix whose parameters are estimated by the method of moments. When the mean response is correctly specified, consistent parameter estimates will be derived even if the algebraic form of the correlation structure is misspecified. However, some loss of efficiency could result if the specified working correlation matrix is far from the true correlation. we estimate Equation (3) assuming an autoregressive correlation structure for each measure of information production. Pan (2001) proposed a model-selection method for generalized estimating equations known as the quasi-likelihood information criterion. The specification of a negative binomial distribution with an autoregressive correlation structure is supported by this criterion.

[11] For a one-unit change in the predictor variable, the difference in the logs of expected counts of the dependent variable is expected to change by the respective regression coefficient. For the coefficient on $SizeDec_{it}$, $e^{0.293} = 1.34$.

Consistent with the predictions of the information driven comovement hypothesis, Table II shows that analysts coordinate on firms whose average newswire similarity, $\overline{WireSim_{it}^{all}}$, is high. However, we find no evidence that journalist-produced text volume is positively influenced by total payoff covariance. Thus, there is a strategic complementarity in information produced by analysts, but a strategic substitutability in information distributed by *Reuters News*.

Table II also demonstrates that contemporaneous average price comovement $\overline{\rho_{it}}$ has only a modest impact on analyst following and does not contribute positively to future text volume. Thus, average newswire similarity, $\overline{WireSim_{it}^{all}}$, is a better predictor of analyst information production than historical comovement, $\overline{\rho_{it}}$. We also find that future analyst following and journalist coverage increase with firm size, but only journalists are influenced positively by contemporaneous volatility. While journalists and analyst should both be motivated to focus their efforts on generating the most profitable content, their methods for creating value seem to diverge. Overall, we find that analysts concentrate on firms whose fundamentals are good predictors of other companies', whereas journalists focus on recent volatility.

The positive and significant coefficient on $WrdCnt_{it}^{firm}$ in the second column provides evidence that future journalist coverage is positively influenced by contemporaneous firm-generated text volume. Journalists are portrayed as information producers in the Veldkamp (2006) model, however, the positive association with contemporaneous firm-generated output implies that *Reuters News* may function more like an echo for primary sources.

This result is consistent with the findings of Ahern and Sosyura (2014), who show that firms originate and disseminate information through the financial media.[12] Their conclusions are based

[12] The Pew Research Center (2011) analyzed several major storylines reported on television, radio, newspaper or online outlets and found that only 14% originated with reporters.

on an even narrower classification of journalist-produced content. Publications like *The Wall Street Journal*, *The New York Times* and *The Washington Post* are described as media sources in their study, whereas *Reuters News*, *Dow Jones News Service* and *Business Wire* are lumped together as "firm-originated news." While the *Business Wire* stories included in our sample are clearly firm-generated, those from *Reuters News* have journalist bylines. Still, Ahern and Sosyura (2014) justify their classification by arguing that newswire stories often provide little analysis. Nevertheless, if content from *Reuters News* is at least somewhat "firm-originated," the market for information will play a smaller role in determining their coverage decisions.

B. Average comovement

In addition to analyzing the determinants of information production, we are also interested in whether the availability of firm-specific information reduces comovement. Veldkamp's (2006) model predicts that comovement will be excessively high between two assets when investors must make correlated inferences about their values instead of directly observing payoff signal of even one asset. Thus, conditional on total payoff covariance, higher volumes of firm-specific information consumption should be inversely related to that firm's average level of comovement.

Our analysis of average comovement is summarized by the following regression model:

$$\begin{aligned}\overline{\rho_{it+1}} = \beta_0 + \beta_1 SizeDec_{it} + \beta_2 \sigma_{it} + \beta_3 WireDum_{it}^{all} + \beta_4 \overline{WireSim_{it}^{all}} \\ + \beta_5 \frac{WrdCnt_{it}^{firm}}{1{,}000} + \beta_6 \frac{WrdCnt_{it}^{rtrs}}{1{,}000} + \beta_7 AnaNum_{it} + \beta_8 \overline{\rho_{it}} \\ + \sum_{k=9}^{K} \beta_k Control_{kit} + \alpha_{t+1} + Ind_i + \varepsilon_{it+1}.\end{aligned} \tag{4}$$

To account for varying levels of average correlation between industries, every firm in the sample is assigned to one of the 49 industry portfolios as defined on Kenneth French's website. Ind_i is a fixed effect describing industry affiliation. The coefficients β_6 and β_7 will be negative if the

availability of firm-specific information produced by journalists and analysts reduces stock price comovement.

The results from estimating Equations (4) are reported in the fourth column of Table II with standard errors clustered by firm and time using the Cameron, Gelbach and Miller (2011) multi-way clustering procedure. The relation between contemporaneous analyst following and future comovement is consistent with the predictions of the information driven comovement hypothesis. Specifically, the coefficient on $AnaNum_{it}$ is negative and significant, implying that a firm's average level of comovement with all other firms in the market, $\overline{\rho_{it+1}}$, is inversely related to the amount of information produced by analysts. Thus, future comovement is highest when analyst following is low and investors are most likely to be making correlated inferences about a particular firm's value. The availability of relevant firm- and journalist-produced content, however, does not reduce a particular company's average level of stock price comovement with all other firms. Therefore, the production of information by either firms or journalists may not mirror investor information demand.

C. *Information driven price comovement*

While the results in Table II imply that analyst output may be determined by a market for information, there is less evidence that the volume of firm-generated, or perhaps even journalist-generated, newswire content is similarly influenced by investor demand. Thus, individual investors must choose which pieces of newswire text to consume because it is not economical to process all of the content appearing on the IDN. The proceeding analysis investigates whether investors cluster their information demand on the types of signals that cause stock price comovement to be high relative to the covariance of underlying fundamentals. First, we analyze how aggregate information consumption changes with market conditions. Second, we examine

how investors choose which types of information to consume. Finally, we study whether the type of information consumed differs across market states.

C.1. Market conditions and information consumption

The information driven comovement hypothesis by Veldkamp (2006) suggests that, on average, the relation between stock price comovement and similarity in information signals should become stronger as total information consumption increases. From the information driven comovement hypothesis, we identify three maket conditions that should influence aggregate demand for qualitative information. First, when the value of an asset rises, investors must hold that additional asset value in order for the asset market to clear. Therefore, there will be more aggregate demand for qualitative information about high-value assets whenever most assets are highly valued. This implies that the relation between newswire similarity and future Pearson return correlation becomes stronger as aggregate market levels rise. Changing asset values will be measured by the total return of the CRSP Value Weighted Index, R_t^{Mkt}, during period t. Panel A of Figure 2 portrays the level and return of the CRSP Market Weighted Index over the entire sample period. The market loses and regains half of its value during this span, providing ample opportunity to examine how information consumption responds to market-wide fluctuations.

Next, we use the daily return standard deviation σ_t^{Mkt} of the CRSP Market Weighted Index during period t to gauge the importance of asset-relevant information in times of uncertainty. As equity payoffs become less predictable, the marginal benefit of observing additional signals rises, causing market-wide information consumption, Λ, to increase. Thus, $WireSim_{ijt}$ will be a better predictor of stock return correlation when market-wide uncertainty σ_t^{Mkt} is high. In untabulated results, an alternative measure of payoff predictability, the Chicago Board Options Exchange Market Volatility Index (VIX), is substituted in our analysis and the inferences are unchanged.

Panel B of Figure 2 shows that both measures of uncertainty are highly correlated throughout the sample period.

Finally, price comovement will be highest when investors make correlated inferences about the values of many assets. As demand for asset-specific information increases, however, the pairwise return correlation between firms should track more closely to the covariance of their payoff signals. Thus, the relation between newswire similarity and future price comovement should vary inversely with aggregate return correlation. The variable $\bar{\rho}_t$, defined as the sample average of all pairwise return correlations, ρ_{ijt}, in a given period t, is used to capture the aggregate level of equity price comovement. According to Panel C of Figure 2, $\bar{\rho}_t$ has also varied considerably across the sample period, rising as high as 61.8% in the third quarter of 2011.

Most of our subsequent analysis centers on the following basic regression model:

$$\begin{aligned}\rho_{ijt+1} = {} & \beta_0 + \beta_1 S34Sim_{ijt} + \beta_2 S12Sim_{ijt} + \beta_3 EPSSim_{ijt} + \beta_4 \max_{k\in i,j} Size_{kt} \\ & + \beta_5 \max_{k\in i,j} \sigma_{kt} + \beta_6 WireDum_{ijt} + \beta_7 TakeSim_{ijt} + \beta_8 WireSim_{ijt} \\ & + \beta_9 \left(WireSim_{ijt} \times R_t^{Mkt}\right) + \beta_{10} \left(WireSim_{ijt} \times \sigma_t^{Mkt}\right) \\ & + \beta_{11} \left(WireSim_{ijt} \times \bar{\rho}_t\right) + \sum_{k=12}^{K} \beta_k Control_{kijt} + \alpha_{t+1} + \gamma_{i\wedge j} \\ & + \delta_{i\vee j} + + \varepsilon_{ijt+1},\end{aligned} \tag{5}$$

where α_{t+1} is a time series fixed effect, $\gamma_{i\wedge j}$ 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. The first three variables in Equation (5) control for qualitative information generated by certain types of information producers. A measure introduced by Israelsen (2015) accounts for information-related comovement that is attributable to commonality analyst following. This variable is defined as:

$$EPSSim_{ijt} = {N_{ijt}^{an}} \Big/ \sqrt{N_{it}^{an} N_{jt}^{an}}, \tag{6}$$

where N_{ij}^{an} is the number of analysts from the I/B/E/S database following both firms i and j in a period t, and N_{it}^{an} and N_{jt}^{an} are the number of analysts following firms i and j respectively. Measures of commonality in instiutional and mutual fund ownership, $S34Sim_{ijt}$ and $S12Sim_{ijt}$, are constructed in an analgolous way.

The next two variables in Equation (5) account for cross-sectional differences in average correlations based on individual firm characteristics. First, the market capitalizations of individual firms are calculated on the final trading day of period $t-1$, and the variable $\max_{k \in i,j} Size_{kt}$ represents the maximum market value of the firm-pair. Next, the daily return standard deviation is calculated for each firm over all of the trading days in period t, and $\max_{k \in i,j} \sigma_{kt}$ is the maximum volatility between firm i and j.

$WireSim_{ijt}$ is either the textual similarity of all content appearing on the Reuters IDN, $WireSim_{ijt}^{all}$, or the textual similarity of newswire content contributed by only the firms themselves, $WireSim_{ijt}^{firm}$. $WireDum_{ijt}$ is a binary variable indicating that both firms had some positive volume of text during period t. $TakeSim_{ijt}$ is included to account for situations where newswire similarity is high because two firms are frequently mentioned in the same newswire item. The variable $TakeSim_{ijt}$ is defined analogously to $EPSSim_{ijt}$ in Equation (6), except that the numerator represents the number of newswire items that mention both firms i and j, and the denominator includes the number of items mentioning each individual firm. All of the systematic and alternative controls, $Control_{kijt}$, introduced in Box (2018) are included in every specification. A description of these variables is also provided in Panel C of Table A-1.

Equation (5) measures the change in future return correlation that results from a change in contemporaneous newswire similarity. It is possible that contemporaneous changes in newswire similarity are themselves responses to changes in return correlation occurring earlier in the same period. Therefore, the specification should account for the current period's, and possibly even earlier periods', observations of pairwise return correlation. Furthermore, all estimated return correlations have a value bounded between -1 and 1, but the error term ε_{ijt+1} is assumed to be distributed over a range of $-\infty$ to ∞. To improve the accuracy of the coefficient standard errors, the Fisher transformation is applied to the correlation estimates:

$$z_{ijt} = \frac{1}{2}\ln\frac{1+\rho_{ijt}}{1-\rho_{ijt}}. \tag{7}$$

Together, these concerns motivate the following model with transformed and lagged dependent variables:

$$\begin{aligned} z_{ijt+1} = \sum_{s=0}^{S} \phi_s z_{ijt-s} &+ \beta_1 S34Sim_{ijt} + \beta_2 S12Sim_{ijt} + \beta_3 EPSSim_{ijt} \\ &+ \beta_4 \max_{k\in i,j} Size_{kt} + \beta_5 \max_{k\in i,j} \sigma_{kt} + \beta_6 WireDum_{ijt} + \beta_7 TakeSim_{ijt} \\ &+ \beta_8 WireSim_{ijt} + \beta_9\left(WireSim_{ijt} \times R_t^{Mkt}\right) \\ &+ \beta_{10}\left(WireSim_{ijt} \times \sigma_t^{Mkt}\right) + \beta_{11}\left(WireSim_{ijt} \times \overline{\rho_t}\right) \\ &+ \sum_{k=12}^{K} \beta_k Control_{kijt} + \alpha_{t+1} + \gamma_{i\wedge j} + \delta_{i\vee j} + +\varepsilon_{ijt+1}. \end{aligned} \tag{8}$$

For unbiased and consistent estimation of Equation (8), we proceed with the dynamic panel estimator (henceforth DPE) proposed by Arellano and Bover (1995) and Blundell and Bond (1998).[13] Pearson correlations, ρ_{ijt}, and their Fisher transformations, z_{ijt}, are calculated from

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

daily 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 Equation (8) contains 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 computational demands of the Arellano and Bover (1995) and Blundell and Bond (1998) estimation procedure are immense due to the dimensions of the instrument matrix required for efficient parameter estimation. Thus, 150,000 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 1,977,933 firm-pair-period observations.[14]

Variation in aggregate information consumption across market states is examined in Table III. Four lags of the systematic variables, z_{ijt}, $BetaDum_{ijt}$, $BetaCorr_{ijt}$, $SizeDum_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktDum_{ijt}$, $Bk/MktCorr_{ijt}$, $MomDum_{ijt}$, $MomCorr_{ijt}$, $IndDum_{ijt}$ and $IndCorr_{ijt}$, are included to remove any evidence of serial correlation in the first-differenced residuals and validate the moment conditions of the dynamic panel estimator.[15] Just as in Box

[14] 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 2,723 firms that stay in the sample at least 6 periods.

[15] For all of the dynamic panel specifications reported in this paper, the model was first estimated with one contemporaneous observation of each systematic variable. As recommended by Arellano and Bond (1991),

(2018), newswire similarity, whether it be calculated from all attributions, $WireSim_{ijt}^{all}$, or only firm-generated content, $WireSim_{ijt}^{firm}$, is a positive and significant predictor of future stock price comovement. Furthermore, the relation between contemporaneous newswire similarity and future return correlation becomes stronger as market values rise and aggregate payoff uncertainty σ_t^{Mkt} increases. Thus, the degree to which the signals contained in primary sources of information are incorporated into asset prices is consistent with the predictions of the information driven comovement hypothesis. Specifically, Table III implies that investors are more willing to bear the cost of information discovery as the variance of their total payoff increases. These results also demonstrate that the relation between newswire similarity and future comovement weakens when aggregate return correlation, $\bar{\rho}_t$, increases. Thus, the consumption of firm-specific qualitative information, Λ, is lower during periods when market-wide comovement is high.

It is possible that realized variations in the relation between $WireSim_{ijt}$ and ρ_{ijt+1} are caused by intertemporal changes in newswire text instead of fluctuations in information consumption. For example, periods of higher average stock return correlation might simply coincide with a prevalence of firm-speicifc information that is unusually similar across companies. Figure 3 depicts the raw, undifferenced document similarity variables, $\widetilde{WireSim}_{ijt}^{all}$ and $\widetilde{WireSim}_{ijt}^{firm}$, averaged across all firm-pairs with a positive quantity of text for each six-month period. Both time series averages are also pictured for two subsamples of firms truncated by NYSE size deciles. Regardless of attribution or truncation scheme, there does not appear to be

additional lags were added until the moment conditions were satisfied. The untabulated lags do not affect the economic inferences in any way.

much variation in average document similarity across time periods.[16] The lack of systematic variation in document similarity observed in Figure 3, lessens the possibility that our results in Table III stem from market-wide changes in textual similarity.

C.2. Firm characteristics and information consumption

The market-level analysis demonstrates that the consumption of qualitative information adjusts to fluctuations in aggregate returns, volatility and correlation. While these dynamics are consistent with the information driven comovement hypothesis, Table III does not consider why investors choose to consume specific pieces of information. The subsequent analysis examines whether firm-specific information consumption increases as security i's payoffs become less predictable, the stock comprises a larger share of the average investor's portfolio, or signals about the firm contain more information that is relevant to the valuation of other companies.

If firm i is larger and more volatile than firm j, the information driven comovement hypothesis by Veldkamp (2006) suggests that investors should consume more information about firm i because its signals can reduce more total payoff variance. Thus, as firm i's size and standard deviation increase, the fraction of investors that demand information about that company should also rise. Furthermore, the price comovement between firms i and j should move closer to the covariance of their payoff signals, as stock-specific information consumption grows. Likewise, the relation between newswire similarity and future return correlation should be strongest when $\max_{k\in i,j} Size_{kt}$ and $\max_{k\in i,j} \sigma_{kt}$ are large.

[16] The time series averages for $WireSim_{ijt}^{all}$, $WireSim_{ijt}^{firm}$ and $WireSim_{ijt}^{rtrs}$ are mechanically centered at 0 zero during each 6-month period. Therefore, it would be impossible to describe aggregate variation in newswire similarity using these measures.

According to the information driven comovement hypothesis, a signal must contain information about the value of many assets and must be observed by many investors in order for it to produce comovement. To gauge whether signals about a particular firm contain information about the value of many other companies, we rely on the same proxy for total payoff covariance, $\overline{WireSim_{it}^{all}}$, introduced in Section III.A. If there is a strategic complementarity in information consumption, investors will process more information about firms with higher aggregate signal correlation. Similar to our strategy for examining how firm-specific information consumption grows to changes in individual firm size and volatility, the variable $\max_{k \in i,j} \overline{WireSim_{kt}^{all}}$ represents the maximum average payoff covariance between firms i and j. If the relation between newswire similarity and future return correlation is stronger when $\max_{k \in i,j} \overline{WireSim_{kt}^{all}}$ is large, then investors coordinate on the types of signals that generate excessive comovement.

Table IV shows how the consumption of qualitative information relates to the value, risk and average payoff covariance of individual firms. Untabulated in every specification are three lags of each systematic variable and the 11 of the alternative controls, $AnaDum_{ijt}$, $AnaCorr_{ijt}$, $AmiDum_{ijt}$, $AmiCorr_{ijt}$, $SP500_{ijt}$, $PrcDum_{ijt}$, $PrcCorr_{ijt}$, $InstDum_{ijt}$, $InstCorr_{ijt}$, ρ_{ijt}^{1mo} and ρ_{ijt}^{2mo}, included in Table III. Inferences from the untabulated variables are the same as in previous tables. For every interacted variable, the multiplier and multiplicand are also included individually as regressors. Once again, the significance of newswire similarity is not diminished by the inclusion of interacted variables.

As expected, the coefficients on $WireSim_{ijt}^{all} \times \max_{k \in i,j} Size_{kt}$ and $WireSim_{ijt}^{firm} \times \max_{k \in i,j} Size_{kt}$ are positive and significant, implying that information consumption λ_i increases with firm size. Thus, when it is not economical to process all of the content appearing in the IDN feed, investors

focus their resources on the subset qualitative information that can be used to evaluate the greatest asset value. According to Table IV, the coefficient on $WireSim_{ijt}^{all} \times \max_{k \in i,j} \sigma_{kt}$ is also positive and significant, implying that investors consume more qualitative information when the content relates to firms with high daily return standard deviations. However, the consumption of only firm-generated content is not significantly related to firm volatility.

The results in Table IV are not consistent with a strategic complementarity in information acquisition. The coefficients on $WireSim_{ijt}^{all} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}}$ and $WireSim_{ijt}^{firm} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}}$ are negative significant in all specifications. This implies that investors consume less qualitative information about firms whose payoff signals covary strongly with most other companies. When investors can eliminate the more uncertainty by observing low covariance signals and inferring the values of the high covariance firms, there is a strategic substitutability in information acquisition. Overall, this finding suggests that, on average, investors do not cluster their information demand on the types of signals that can cause stock price comovement to be high relative to the covariance of underlying fundamentals.

C.3. Firm characteristics, market conditions and information consumption

Direct empirical tests of Veldkamp's (2006) information driven comovement hypothesis are complicated by aggregate changes in information consumption. In her model, investors only coordinate on high covariance signals when aggregate information consumption is sufficiently low. As information consumption begins to rise, however, signal demand can spill over into other assets, and a strategic substitutability in information acquisition begins to appear. Thus, whether or not investors coordinate on high covariance signals depends on the aggregate level of information consumption.

Without controlling for market conditions that could influence the overall demand for information, Table IV shows that investors consume less qualitative information about firms whose payoffs have higher average covariances. However, Table III reveals that the aggregate level of information consumption varies with market-wide average comovement, cumulative returns and volatility. Table V examines whether or not these same market conditions influence how investors choose which types of information to consume. If aggregate information consumption recedes when market returns R_t^{Mkt} are negative, aggregate return volatilities σ_t^{Mkt} are low and average return correlation $\bar{\rho}_t$ is high, then these same conditions should encourage investors to coordinate on a limited number of high covariance signals.

Once again, the multiplier and multiplicand are included individually as regressors for every interacted variable. Therefore, all of the interaction terms appearing in Table III and Table IV are included in Table V's specifications. Inferences from all other untabulated variables are the same as in previous tables. Once again, the significance of the newswire similarity measures, $WireSim_{ijt}^{all}$ and $WireSim_{ijt}^{firm}$, are not diminished by including additional interacted variables.

When firm documents are constructed from text combined across all attributions, the negative and significant coefficients on $WireSim_{ijt}^{all} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}} \times R_t^{Mkt}$ and $WireSim_{ijt}^{firm} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}} \times R_t^{Mkt}$ imply that investor coordination on high covariance signals becomes more common when market values are falling. While insignificant, the negative coefficients on $WireSim_{ijt}^{all} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}} \times \sigma_t^{Mkt}$ and $WireSim_{ijt}^{firm} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}} \times \sigma_t^{Mkt}$ are consistent with a preference for high covariance signals when aggregate volatility is low. According to the information driven comovement hypothesis, lower

market returns and standard deviations make it less economical to read and evaluate primary sources of information.

Table III demonstrated that firm-specific information consumption is low whenever market-wide comovement is high. In Table V, the positive and significant coefficients on $WireSim_{ijt}^{all} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}} \times \bar{\rho}_t$ and $WireSim_{ijt}^{firm} \times \max_{k \in i,j} \overline{WireSim_{kt}^{all}} \times \bar{\rho}_t$ imply that coordination on high covariance signals becomes more common as market-wide return correlations $\bar{\rho}_t$ increase. Thus, episodes of high average stock price comovement coincide with an increased consumption of information related to firms with higher average newswire similarities. Consistent with the information driven comovement hypothesis, we find that market-wide correlations are higher when many investors consume qualitative information about firms whose payoffs covary most strongly with many other companies. Likewise, as aggregate correlation falls, so does the demand for these high covariance signals.

Overall, the results in Table IV and Table V imply that comovement rises when many investors observe a limited number of high covariance signals, but also that demand for low covariance signals is higher on average. Thus, complementarity leads to comovement, but substitutability typically prevails.

IV. Closing remarks

The process by which investors choose the type and quantity of information to consume is poorly understood, but critical to the functioning of financial markets. This paper provides a new empirical strategy to identify investors' information choices by inferring the type of information that is consumed and incorporated into asset prices. Consistent with a theoretical model presented by Veldkamp (2006), we find that stock price comovement is high relative to the

covariance of underlying fundamentals when investors cluster their information demand on just a few firms whose payoffs covary strongly with many other companies. However, as the breadth of information consumption increases, we also find that stock return correlations move closer to their fundamental covariances. Overall, Our findings imply that investor information consumption choices are influenced by a market for information.

References

Ahern, Kenneth R., and Denis Sosyura. 2014. "Who Writes the News? Corporate Press Releases during Merger Negotiations." *The Journal of Finance Vol. LXIX, No. 1* 241-291.

Andrei, Daniel, and Michael Hasler. 2016. "Dynamic Attention Behavior Under Return Predictability." *Working paper.*

Andrei, Daniel, and Michael Hasler. 2015. "Investor Attention and Stock Market Volatility." *The Review of Financial Studies / v 28 n 1* 33-72.

Arellano, Manuel, and Olympia Bover. 1995. "Another look at the instrumental variable estimation of error-components models." *Journal of Econometrics 68 (1)* 29-51.

Asness, Clifford S., Tobias J. Moskowitz, and Lasse Heje Pedersen. 2013. "Value and Momentum Everywhere." *The Journal of Finance Vol. LXVIII, No. 3* 929–985.

Ballinger, Gary A. 2004. "Using Generalized Estimating Equations for Longitudinal Data Analysis." *Organizational Research Methods vol. 7 no. 2* 127-150.

Bekaert, Geert, Robert J. Hodrick, and Xiaoyan Zhang. 2009. "International Stock Return Comovements." *The Journal of Finance Vol. LXIV, No. 6* 2591-2627.

Bilisoly, Roger. 2008. *Practical Text Mining with Perl.* Hoboken, New Jersey: John Wiley and Sons. Inc.

Blundell, Richard, and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics 87 (1)* 115-143.

Box, Travis. 2018. "Qualitative Similarity and Stock Price Comovement." *Journal of Banking and Finance 91* 49-69.

Brandt, Michael W., Alon Brav, John R. Graham, and Alok Kumar. 2010. "The Idiosyncratic Volatility Puzzle: Time Trend of Specualtive Episode." *The Review of Financial Studies / v 23 n 2* 865-899.

Brockman, Paul, Ivonne Liebenberg, and Maria Schutte. 2010. "Comovement, information production, and the business cycle." *Journal of Financial Economics 97* 107-129.

Cameron, A. Colin, and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data, Second Edition.* New York: Cambridge University Press.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multi-way Clustering." *Journal of Business and Economic Statistics 29(2)* 238-249.

Campbell, John Y., Martin Lettau, Burton G. Malkiel, and Yexiao Xu. 2001. "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk." *The Journal of Finance Vol. LVI, No. 1* 1-43.

Fang, Lily, and Joel Peress. 2009. "Media Coverage and the Cross-section of Stock Returns." *The Journal of Finance Vol. LXIV, No. 5* 2023-2052.

Gardiner, Joseph C, Zhehui Luo, and Lee Anne Roman. 2009. "Fixed effects, random effects and GEE: what are the differences?" *Statistics in Medicine Volume 28, Issue 2* 181–360.

Gardner, William, Edward P Mulvey, and Esther C. Shaw. 1995. "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." *Psychological Bulletin, Vol 118(3)* 392-404.

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *The Amercian Economic Review, Vol. 70, No. 3* 393-408.

Hameed, Allaudeen, Randall Morck, Jianfeng Shen, and Bernard Yeung. 2015. "Information, Analysts, and Stock Return Comovement." *The Review of Financial Studies / v 28 n 11* 3153-3187.

Hardin, James W., and Joseph M. Hilbe. 2013. *Generalized Estimating Equations, Second Edition.* Boca Raton, FL: CRC Press.

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

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

Hoberg, Gerard, and Gordon Phillips. 2010a. "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis." *The Review of Financial Studies / v 23 n 10* 3773-3811.

Irvine, Paul J., and Jeffrey Pontiff. 2009. "Idiosyncratic Return Volatility, Cash Flows, and Product Markets." *The Review of Financial Studies / v 22 n 3* 1149-1177.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp. 2016. "A Rational Theory of Mutual Funds' Attention Allocation." *Econometrica 84 (2)* 571-626.

Ledoit, Olivier, and Michael Wolf. 2003. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection." *Journal of Empirical Finance 10* 603-621.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika Vol. 73* 13-22.

Liberti, José María, and Mitchell A. Petersen. 2017. "Information: Hard and Soft." *The Review of Corporate Finance Studies (forthcoming).*

Loughran, Tim, and Bill McDonald. 2015. "Information Decay and Financial Disclosures." *Working paper.*

Table I

Summary statistics for production regressions

This table presents summary statistics for the variables appearing in Equation (4). $WrdCnt_{it+1}^{rtrs}$ is the total number of words written about firm i and distributed by *Reuters News* during each 6-month period t, and $WrdCnt_{it}^{firm}$ is the total number of words contributed by all other attributions. $AnaNum_{it}$ is the number of unique analysts with an earnings prediction recorded in the I/B/E/S database during period t. $\overline{\rho_{it}}$ is calculated by averaging ρ_{ijt}, the Pearson correlation in the daily stock returns of firms i and j, over all firms j. Similarly, $\overline{WireSim_{it}^{all}}$ is firm i's average newswire similarity $WireSim_{ijt}^{all}$ over all firms j. σ_{it} is firm i's daily stock return standard deviation during period t.

	Mean	Std Dev	P1	P10	P25	P50	P75	P90	P99
σ_{it}	2.74	1.69	0.83	1.26	1.68	2.33	3.28	4.60	9.08

Mondria, Jordi. 2010. "Portfolio choice, attention allocation, and price comovement." *Journal of Economic Theory 145* 1837-1864.

Pan, Wei. 2001. "Akaike's Information Criterion in Generalized Estimating Equations." *Biometrics Vol. 57, No. 1* 120-125.

Peng, Lin, and Wei Xiong. 2006. "Investor attention, overconfidence and category learning." *Journal of Financial Economics 80* 563-602.

Pew Research Center. 2011. *How News Happens.* January 11.

Pindyck, Robert S., and Julio Rotemberg. 1993. "The Comovement of Stock Prices." *The Quarterly Journal of Economics, Vol. 108, No. 4* 1073-1104.

Veldkamp, Laura. 2011. "Information Choice with Sunstitutability in Actions." In *Information Choice in Macroeconomics and Finance*, by Laura Veldkamp, 83-138. Princeton, New Jersey: Princeton University Press.

Veldkamp, Laura. 2006. "Information Markets and the Comovement of Asset Prices." *Review of Economic Studies 73* 823-845.

$WireDum_{it}^{all}$	0.95	0.22	0	1	1	1	1	1	1
$\overline{WireSim_{it}^{all}}$	0.003	0.023	-0.066	-0.025	-0.008	0.004	0.017	0.028	0.060
$WrdCnt_{it}^{firm}$	4,123.78	6,142.75	0	154	1,020	2,531	5,108	9,164	25,724
$WrdCnt_{it}^{rtrs}$	559.53	2,114.31	0	0	0	89	403	1,145	7,899
$AnaNum_{it}$	9.28	8.49	0	0	3	7	14	21	35
$\overline{\rho_{it}}$	0.30	0.13	0.06	0.16	0.22	0.28	0.38	0.49	0.68

Table II
Information production and firm characteristics

This table reports the estimation of Equations (3) and (4). $WrdCnt_{it+1}^{rtrs}$ is the total number of words written about firm i and distributed by *Reuters News* during each 6-month period t, and $WrdCnt_{it}^{firm}$ is the total number of words contributed by all other attributions. $AnaNum_{it}$ is the number of unique analysts with an earnings prediction recorded in the I/B/E/S database during period t. $\overline{\rho_{it}}$ is calculated by averaging ρ_{ijt}, the Pearson correlation in the daily stock returns of firms i and j, over all firms j. Similarly, $\overline{WireSim_{it}^{all}}$ is firm i's average newswire similarity $WireSim_{it}^{all}$ over all firms j. $WireDum_{it}^{all}$ is a binary variable set to 1 whenever firm i has any positive number of words appearing on the Reuters Integrated Data Network during period t. $SizeDec_{it}$ is firm i's NYSE decile based on market value from the last trading day of period $t-1$, and σ_{it} is firm i's daily stock return standard deviation during period t. A description for all other included variable calculations is provided in Panel B of Table A-1. A generalized estimating equations approach, specified with a negative binomial distribution and an autoregressive correlation structure, is used when the dependent variable measures future information production, either $WrdCnt_{it+1}^{firm}$, $WrdCnt_{it+1}^{rtrs}$ or $AnaNum_{it+1}$. Ordinary least squares is used when the dependent variable measures future average comovement, $\overline{\rho_{it+1}}$. The t-statistics (reported in parenthesis) in the information production specifications are calculated from standard errors clustered by firm, and t-statistics in the comovement specification are derived from standard errors clustered by firm and time using the Cameron, Gelbach and Miller (2011) multi-way clustering procedure. * and ** represent significance at the 5% and 1% level, respectively.

Table II—Continued

	Generalized Estimating Equations—Negative Binomial Distribution				Ordinary Least Squares
	$WrdCnt_{it+1}^{firm}$	$WrdCnt_{it+1}^{rtrs}$	$AnaNum_{it+1}$	$ServCount_{it+1}$	$\overline{\rho_{it+1}}$
$BetaDec_{it}$	0.0128**	0.0293**	0.00730**	-0.0124**	0.0182**
	(5.001)	(4.930)	(6.517)	(-4.332)	(4.085)
$Bk/MktDec_{it}$	-0.00574	0.0332**	-0.00302*	0.0117**	0.0113**
	(-1.811)	(4.436)	(-2.200)	(2.688)	(4.271)
$MomDec_{it}$	0.00364**	0.0155**	-0.00156**	0.00282	-0.000386
	(3.130)	(4.300)	(-3.529)	(1.830)	(-0.130)
$AmiDec_{it}$	-0.0562**	-0.201**	-0.0593**	-0.0685**	0.0104*
	(-9.244)	(-11.34)	(-20.89)	(-9.209)	(2.037)
$PrcDec_{it}$	-0.0256**	-0.0540**	-0.00426**	-0.0150*	0.00716**
	(-6.896)	(-5.877)	(-2.639)	(-2.564)	(2.902)
$InstDec_{it}$	0.00513	-0.00703	0.0214**	-0.0115*	-0.000316
	(1.521)	(-0.892)	(12.64)	(-2.551)	(-0.199)
$SP500_{it}$	0.292**	0.343**	0.0978**	0.214**	0.0464*
	(8.637)	(6.422)	(4.971)	(5.113)	(2.382)
$SizeDec_{it}$	0.0532**	0.193**	0.0361**	0.0564**	0.0160
	(8.320)	(12.22)	(14.31)	(6.651)	(1.868)
σ_{it}	-0.0469**	0.120**	-0.0202**	0.0628**	-0.0183
	(-7.758)	(5.915)	(-7.979)	(7.527)	(-0.553)
$WireDum_{it}^{all}$	0.556**	0.605**	0.0380**	0.0141	0.0135
	(26.33)	(7.565)	(4.385)	(0.443)	(0.965)
$\overline{WireSim_{it}^{all}}$	-0.573**	-2.445**	0.197**	-0.470*	0.590**
	(-4.180)	(-4.025)	(3.597)	(-2.269)	(4.008)
$WrdCnt_{it}^{firm}/1{,}000$		0.0267**	0.000856	0.00733**	0.000421
		(10.04)	(1.642)	(3.082)	(0.746)
$WrdCnt_{it}^{rtrs}/1{,}000$	0.00159		-0.000690	0.0173**	-5.67e-05
	(0.360)		(-0.663)	(2.665)	(-0.0338)
$AnaNum_{it}$	0.0101**	0.00282		0.0190**	-0.00225**
	(6.883)	(0.693)		(8.799)	(-3.128)
$ServCount_{it}$	0.00202	0.0433**	0.00290		-0.00159
	(0.630)	(6.896)	(1.640)		(-0.486)
$\overline{\rho_{it}}$	-0.00330	-0.0608**	0.00911**	0.00281	0.401**
	(-0.867)	(-5.548)	(6.334)	(0.498)	(18.77)
Working Correlation Matrix	AR(1)	AR(1)	AR(1)		
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	No	No	No	No	Yes
R-squared					0.777
Dispersion	1.982	4.094	0.836	1.581	
Observations	40,155	40,155	40,155	40,155	40,155

Table III
Market conditions and information consumption

The dependent variable in all specifications is the Fisher transformation z_{ijt+1} of the Pearson correlation ρ_{ijt+1} calculated from the daily returns of firms i and j in excess of the risk free rate for each 6-month period $t+1$. The market condition variables R_t^{Mkt}, σ_t^{Mkt}, and $\bar{\rho}_t$, defined in Figure 2, are standardized with a mean of 0 and a standard deviation of unity. A description for all other included variable calculations is provided in Table A-1. 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 dynamic panel estimation. "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 $IndCorr_{ijt}$.

Table III—Continued (Control Variables)

	(1)	(2)	(3)	(4)
z_{ijt}	0.194**	0.196**	0.194**	0.194**
	(127.8)	(128.9)	(127.5)	(127.7)
$BetaDum_{ijt}$	0.0246**	0.0232**	0.0252**	0.0248**
	(11.60)	(10.92)	(11.88)	(11.65)
$BetaCorr_{ijt}$	0.0243**	0.0227**	0.0250**	0.0245**
	(10.41)	(9.721)	(10.70)	(10.48)
$SizeDum_{ijt}$	0.0743**	0.0674**	0.0739**	0.0684**
	(7.720)	(7.000)	(7.686)	(7.100)
$SizeCorr_{ijt}$	0.0711**	0.0643**	0.0707**	0.0650**
	(7.182)	(6.483)	(7.143)	(6.563)
$Bk/MktDum_{ijt}$	0.105**	0.107**	0.106**	0.107**
	(26.53)	(26.82)	(26.67)	(26.72)
$Bk/MktCorr_{ijt}$	0.114**	0.116**	0.115**	0.116**
	(26.33)	(26.60)	(26.48)	(26.51)
$MomDum_{ijt}$	0.0409**	0.0401**	0.0414**	0.0411**
	(21.63)	(21.17)	(21.92)	(21.69)
$MomCorr_{ijt}$	0.0403**	0.0394**	0.0411**	0.0406**
	(19.44)	(18.93)	(19.79)	(19.52)
$IndDum_{ijt}$	0.0826**	0.0891**	0.0834**	0.0904**
	(13.89)	(14.96)	(14.00)	(15.16)
$IndCorr_{ijt}$	-0.0737**	-0.0765**	-0.0740**	-0.0766**
	(-38.04)	(-39.70)	(-38.19)	(-39.77)
ρ_{ijt}^{1mo}	0.0173**	0.0175**	0.0172**	0.0173**
	(20.19)	(20.33)	(19.97)	(20.19)
ρ_{ijt}^{2mo}	0.0263**	0.0264**	0.0261**	0.0261**
	(20.46)	(20.57)	(20.36)	(20.33)
$AnaDum_{ijt}$	-0.0443**	-0.0444**	-0.0453**	-0.0456**
	(-5.536)	(-5.530)	(-5.649)	(-5.675)
$AnaCorr_{ijt}$	-0.0468**	-0.0469**	-0.0478**	-0.0482**
	(-5.613)	(-5.610)	(-5.728)	(-5.759)
$InstDum_{ijt}$	0.0173**	0.0186**	0.0185**	0.0186**
	(3.581)	(3.823)	(3.819)	(3.828)
$InstCorr_{ijt}$	0.0176**	0.0189**	0.0188**	0.0189**
	(3.424)	(3.665)	(3.660)	(3.662)
$AmiDum_{ijt}$	0.0993**	0.0972**	0.0993**	0.0975**
	(10.89)	(10.63)	(10.89)	(10.67)
$AmiCorr_{ijt}$	0.100**	0.0979**	0.100**	0.0981**
	(10.71)	(10.45)	(10.71)	(10.48)
$PrcDum_{ijt}$	0.0147**	0.0143**	0.0149**	0.0145**
	(4.166)	(4.056)	(4.222)	(4.098)
$PrcCorr_{ijt}$	0.0134**	0.0130**	0.0136**	0.0131**
	(3.510)	(3.394)	(3.560)	(3.425)
$SP500_{ijt}$	-0.00677**	-0.00203	-0.00592*	-0.00231
	(-2.859)	(-0.860)	(-2.507)	(-0.984)

Each specification continues on following page

Table III—Continued (Information Variables)

	(1)	(2)	(3)	(4)
$S34Sim_{ijt}$	0.00694*	0.00579	0.00938**	0.00901*
	(1.988)	(1.653)	(2.681)	(2.567)
$S12Sim_{ijt}$	0.0398**	0.0444**	0.0379**	0.0410**

	(9.256)	(10.32)	(8.831)	(9.521)
$EPSSim_{ijt}$	-0.0614**	-0.0402*	-0.0600**	-0.0503**
	(-3.796)	(-2.501)	(-3.668)	(-3.082)
$\max_{k \in i,j} Size_{kt}$	0.00854**	0.00795**	0.00838**	0.00794**
	(23.12)	(21.93)	(22.75)	(21.82)
$\max_{k \in i,j} \sigma_{kt}$	0.00111**	0.000937**	0.00112**	0.00102**
	(3.169)	(2.676)	(3.201)	(2.897)
$WireDum_{ijt}^{all}$	0.00716**	0.00852**		
	(7.179)	(8.605)		
$TakeSim_{ijt}^{all}$	-0.173**	-0.190**		
	(-2.754)	(-2.995)		
$WireSim_{ijt}^{all}$	0.0426**	0.0779**		
	(8.629)	(18.97)		
$WireSim_{ijt}^{all} \times R_t^{Mkt}$		0.0288**		
		(6.315)		
$WireSim_{ijt}^{all} \times \sigma_t^{Mkt}$		0.0431**		
		(6.127)		
$WireSim_{ijt}^{all} \times \bar{\rho}_t$		-0.0310**		
		(-5.778)		
$WireDum_{ijt}^{firm}$			0.000878	0.00180*
			(1.044)	(2.153)
$TakeSim_{ijt}^{firm}$			-0.0105	-0.0392
			(-0.160)	(-0.598)
$WireSim_{ijt}^{firm}$			0.0168**	0.0468**
			(3.502)	(11.53)
$WireSim_{ijt}^{firm} \times R_t^{Mkt}$				0.0210**
				(4.852)
$WireSim_{ijt}^{firm} \times \sigma_t^{Mkt}$				0.0386**
				(5.753)
$WireSim_{ijt}^{firm} \times \bar{\rho}_t$				-0.0418**
				(-8.128)
Time Fixed Effects	Yes	Yes	Yes	Yes
Firm-pair Panel Effects	Yes	Yes	Yes	Yes
Systematic Lags	4	4	4	4
AR(2) Test	-0.369	0.321	-0.278	0.531
Observations	1,452,317	1,452,317	1,452,317	1,452,317

Table IV
Firm characteristics and information consumption

The dependent variable in all specifications is the Fisher transformation z_{ijt+1} of the Pearson correlation ρ_{ijt+1} calculated from the daily returns of firms i and j in excess of the risk-free rate for each 6-month period $t+1$. Firm i's and j's average newswire similarity is calculated for each period t, and $\max_{k\in i,j} \overline{WireSim_{kt}^{all}}$ is the standardized maximum average newswire similarity between both firms. Similarly, $\max_{k\in i,j} Size_{kt}$ and $\max_{k\in i,j} \sigma_{kt}$ are the standardized maximum market value and daily return standard deviation between the firms. A description for all other included variable calculations is provided in Table A-1. 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 dynamic panel estimation. "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 $AnaDum_{ijt}$, $AnaCorr_{ijt}$, $AmiDum_{ijt}$, $AmiCorr_{ijt}$, $SP500_{ijt}$, $PrcDum_{ijt}$, $PrcCorr_{ijt}$, $InstDum_{ijt}$, $InstCorr_{ijt}$, ρ_{ijt}^{1mo} and ρ_{ijt}^{2mo} as untabulated controls.

Table IV—Continued

	(1)		(2)
$S34Sim_{ijt}$	0.0658** (11.14)	$S34Sim_{ijt}$	0.0719** (12.06)
$S12Sim_{ijt}$	0.0400** (7.324)	$S12Sim_{ijt}$	0.0334** (6.123)
$EPSSim_{ijt}$	0.0808** (3.280)	$EPSSim_{ijt}$	0.0751** (2.937)
$\max_{k\in i,j} Size_{kt}$	0.0155** (22.37)	$\max_{k\in i,j} Size_{kt}$	0.0161** (23.44)
$\max_{k\in i,j} \sigma_{kt}$	0.00170** (4.448)	$\max_{k\in i,j} \sigma_{kt}$	0.00152** (3.991)
$\max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	0.000268 (1.356)	$\max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	0.000679** (3.822)
$WireDum_{ijt}^{all}$	0.00551** (5.732)	$WireDum_{ijt}^{firm}$	0.00141 (1.733)
$TakeSim_{ijt}^{all}$	-0.131* (-2.097)	$TakeSim_{ijt}^{firm}$	0.00526 (0.0801)
$WireSim_{ijt}^{all}$	0.0838** (14.87)	$WireSim_{ijt}^{firm}$	0.0451** (9.071)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} Size_{kt}$	0.0554** (7.505)	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} Size_{kt}$	0.0137* (2.212)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} \sigma_{kt}$	0.0194** (4.399)	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} \sigma_{kt}$	0.00349 (0.822)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	-0.0152** (-5.778)	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	-0.0173** (-6.060)
Time Fixed Effects	Yes	Time Fixed Effects	Yes
Firm-pair Panel Effects	Yes	Firm-pair Panel Effects	Yes
Alternative Controls	Yes	Alternative Controls	Yes
Systematic Lags	3	Systematic Lags	3
AR(2) Test	1.192	AR(2) Test	1.710
Observations	1,534,833	Observations	1,534,833

Table V

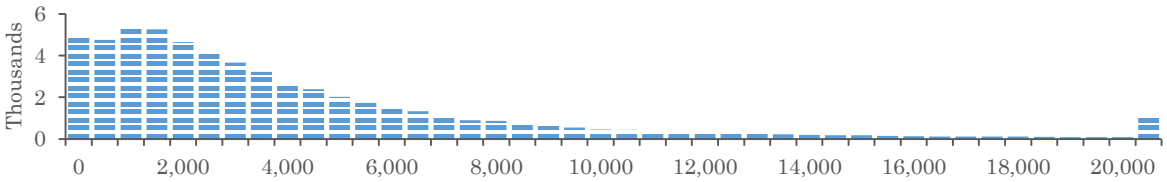
Firm characteristics, market conditions and information consumption

The dependent variable in all specifications is the Fisher transformation z_{ijt+1} of the Pearson correlation ρ_{ijt+1} calculated from the daily returns of firms i and j in excess of the risk free rate for each 6-month period $t+1$. Firm i's and j's average newswire similarity is calculated for each period t, and $\max_{k \in i,j} \overline{WireSim_{kt}^{all}}$ is the standardized maximum average newswire similarity between both firms. A description for all other included variable calculations is provided in Table A-1. 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 dynamic panel estimation. "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 $AnaDum_{ijt}$, $AnaCorr_{ijt}$, $AmiDum_{ijt}$, $AmiCorr_{ijt}$, $SP500_{ijt}$, $PrcDum_{ijt}$, $PrcCorr_{ijt}$, $InstDum_{ijt}$, $InstCorr_{ijt}$, ρ_{ijt}^{1mo} and ρ_{ijt}^{2mo} as untabulated controls.

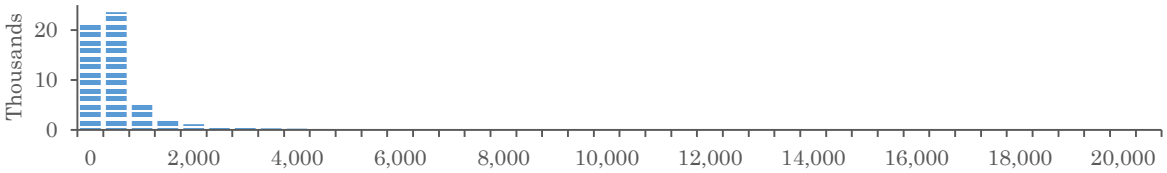
Table V—Continued

	(1)		(2)
$S34Sim_{ijt}$	0.0177**	$S34Sim_{ijt}$	0.0205**
	(5.398)		(6.222)
$S12Sim_{ijt}$	0.0322**	$S12Sim_{ijt}$	0.0298**
	(8.171)		(7.559)
$EPSSim_{ijt}$	-0.0458**	$EPSSim_{ijt}$	-0.0470**
	(-3.013)		(-3.044)
$\max_{k\in i,j} Size_{kt}$	0.0101**	$\max_{k\in i,j} Size_{kt}$	0.00867**
	(26.75)		(25.27)
$\max_{k\in i,j} \sigma_{kt}$	0.00266**	$\max_{k\in i,j} \sigma_{kt}$	0.00267**
	(7.975)		(8.012)
$\max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	0.00404**	$\max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	0.00433**
	(24.86)		(29.77)
$\max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times R_t^{Mkt}$	0.000965**	$\max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times R_t^{Mkt}$	0.00102**
	(4.826)		(5.593)
$\max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \sigma_t^{Mkt}$	0.00134**	$\max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \sigma_t^{Mkt}$	0.00142**
	(4.274)		(4.938)
$\max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \overline{\rho_t}$	-0.000420	$\max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \overline{\rho_t}$	-0.000245
	(-1.714)		(-1.083)
$WireDum_{ijt}^{all}$	0.0117**	$WireDum_{ijt}^{firm}$	0.00859**
	(12.81)		(11.07)
$TakeSim_{ijt}^{all}$	-0.128*	$TakeSim_{ijt}^{firm}$	0.000118
	(-2.071)		(0.00180)
$WireSim_{ijt}^{all}$	0.0564**	$WireSim_{ijt}^{firm}$	0.0324**
	(10.87)		(7.102)
$WireSim_{ijt}^{all} \times R_t^{Mkt}$	0.0139**	$WireSim_{ijt}^{firm} \times R_t^{Mkt}$	0.0150**
	(2.639)		(3.245)
$WireSim_{ijt}^{all} \times \sigma_t^{Mkt}$	0.00746	$WireSim_{ijt}^{firm} \times \sigma_t^{Mkt}$	0.0272**
	(0.751)		(3.008)
$WireSim_{ijt}^{all} \times \overline{\rho_t}$	-0.0211**	$WireSim_{ijt}^{firm} \times \overline{\rho_t}$	-0.0439**
	(-3.335)		(-7.652)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} Size_{kt}$	0.0454**	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} Size_{kt}$	-0.00127
	(6.507)		(-0.215)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} \sigma_{kt}$	0.0212**	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} \sigma_{kt}$	0.00140
	(2.962)		(0.204)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	-0.00608**	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}}$	-0.00457
	(-2.598)		(-1.735)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times R_t^{Mkt}$	-0.00902**	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times R_t^{Mkt}$	-0.0147**
	(-2.686)		(-4.021)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \sigma_t^{Mkt}$	-0.000403	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \sigma_t^{Mkt}$	-0.00162
	(-0.0719)		(-0.267)
$WireSim_{ijt}^{all} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \overline{\rho_t}$	0.0113**	$WireSim_{ijt}^{firm} \times \max_{k\in i,j} \overline{WireSim_{kt}^{all}} \times \overline{\rho_t}$	0.0163**
	(2.625)		(3.585)
Time Fixed Effects	Yes	Time Fixed Effects	Yes
Firm-pair Panel Effects	Yes	Firm-pair Panel Effects	Yes
Alternative Controls	Yes	Alternative Controls	Yes
Systematic Lags	3	Systematic Lags	3
AR(2) Test	-1.929	AR(2) Test	-1.547
Observations	1,534,833	Observations	1,534,833

Panel A: Distribution of $WrdCnt_{it}^{firm}$



Panel B: Distribution of $WrdCnt_{it}^{rtrs}$



Panel C: Distribution of $AnaNum_{it}$

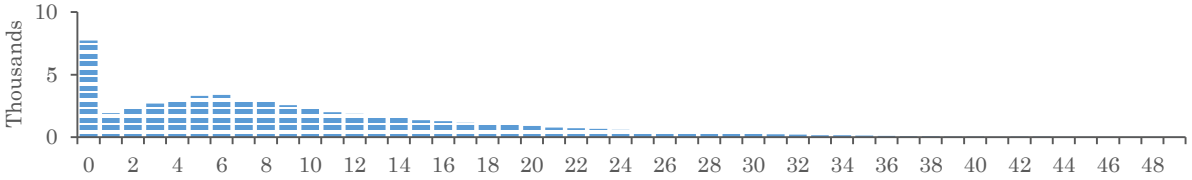
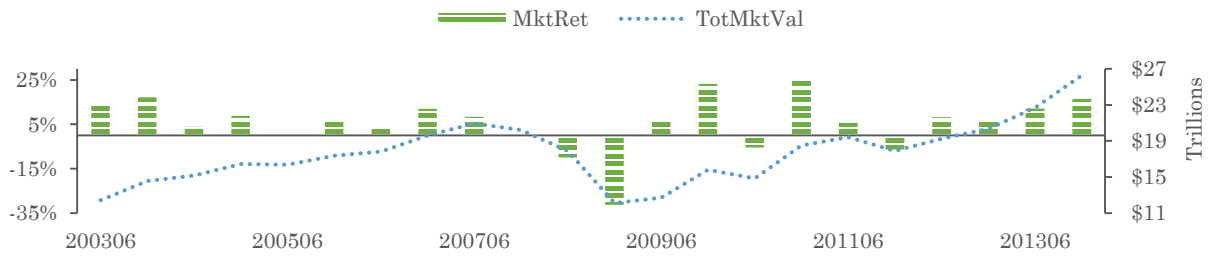


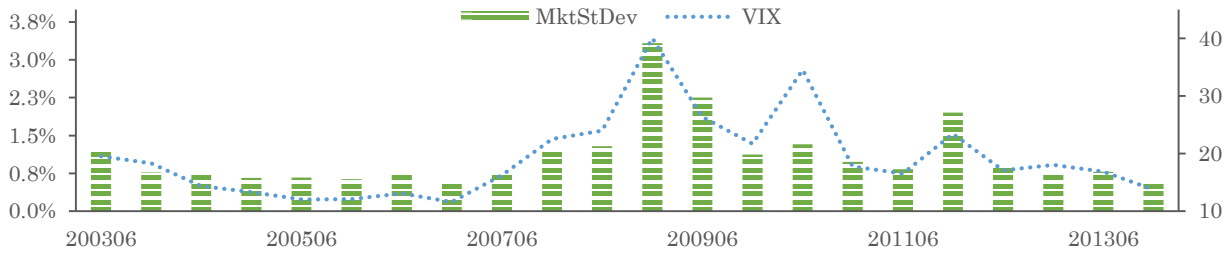
Figure 1. Production variable histograms

Panel A illustrates the pooled distribution of $WrdCnt_{it}^{firm}$, or the total number of words written about firm i and distributed by all attributions other than *Reuters News*. Panel B describes $WrdCnt_{it}^{rtrs}$, or the total number of words written about firm i and distributed by *Reuters News*. Panel C represents the distribution of $AnaNum_{it}$, or the number of unique analysts with an earnings prediction recorded in the I/B/E/S database during period t.

Panel A: Market value and 6-month cumulative return



Panel B: 6-month market daily return standard deviation and VIX level



Panel C: Average pairwise return correlation

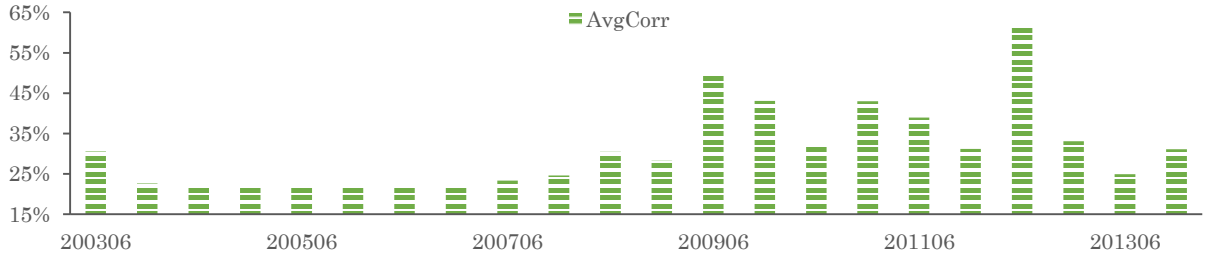
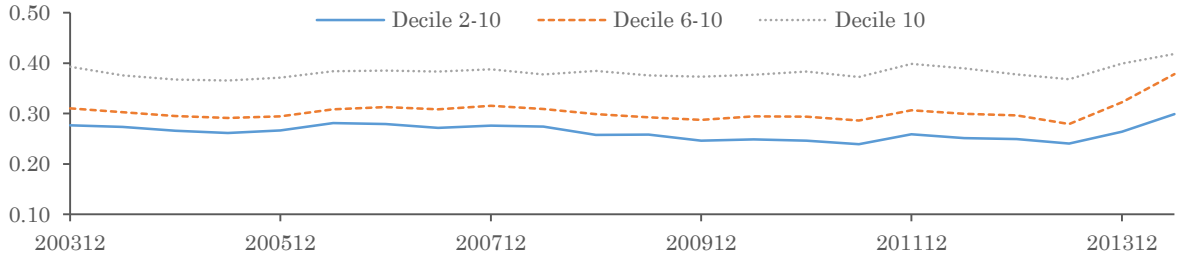


Figure 2. Market-wide financial variables 2003-2013

Panel A illustrates the closing aggregate market level $TotMktVal_t$ (right axis) from the last trading day of period t and the cumulative return R_t^{Mkt} (left axis) of the CRSP Market Weighted Index over period t. Panel B depicts the daily return standard deviation σ_t^{Mkt} (left axis) of the CRSP Market Weighted Index during period t and the Chicago Board of Options Exchange Market Volatility Index VIX_t (right axis) closing value on the last trading day of period t. Panel C represents the $\bar{\rho}_t$, or the sample average of all pairwise return correlations ρ_{ijt} during period t.

Panel A: Time series average document similarity, $\widetilde{WireSim}_{ijt}^{all}$, calculated across all attributions



Panel B: Time series average document similarity, $\widetilde{WireSim}_{ijt}^{firm}$, calculated only from firm-generated content

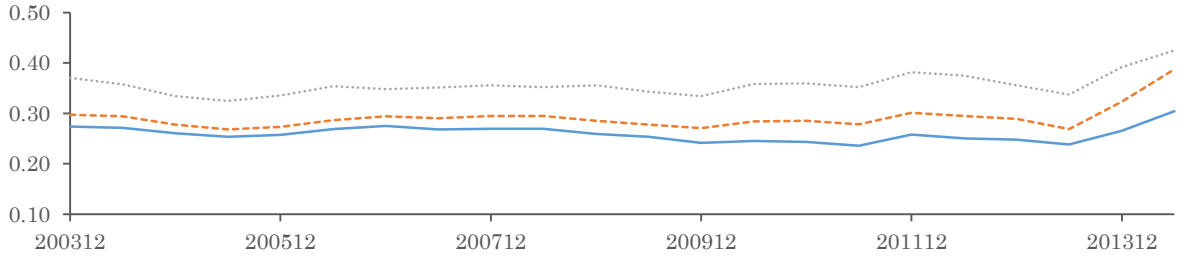


Figure 3. Average document similarity variable over time

For each 6-month period t, average document similarity is calculated across all firm-pairs with some positive quantity of text. Panel A depicts the average document similarity $\widetilde{WireSim}_{ijt}^{all}$ of text appearing on the Reuters Integrated Data Network, while Panel B depicts the average of document similarity $\widetilde{WireSim}_{ijt}^{firm}$ from text generated by the firms themselves in the form of press releases and legal disclosures and Panel C depicts the average of document similarity $\widetilde{WireSim}_{ijt}^{rtrs}$ from text produced by *Reuters News*.

A. Supplementary descriptors

Table A-1

Regression variable definitions

Variable	Definition
	Table A-1 Panel A: First appearing in Table I
$WrdCnt_{it}^{firm}$	Total number of words written about firm i and distributed by all attributions other than *Reuters News* during period t.
$WrdCnt_{it}^{rtrs}$	Total number of words written about firm i and distributed by *Reuters News* during period t.
$AnaNum_{it}$	The number of unique analysts with an earnings prediction recorded in the I/B/E/S database during period t.
$WireDum_{it}^{all}$	Binary variable has a value of 1 whenever firm i has some positive number of total words appearing on the Reuters Integrated Data Network during period t.
$\overline{WireSim_{it}^{all}}$	Firm i's average newswire similarity with all other firms j, $\overline{WireSim_{it}^{all}} = \frac{1}{N-1}\sum_{j\neq i} WireSim_{it}^{all}$, where N is the number of firms with some positive volume of text appearing on the IDN during period t.
$\overline{\rho_{it}}$	Pearson correlation ρ_{ijt} between the daily stock returns of firms i and j averaged over all firms $j \neq i$.
σ_{it}	Firm i's daily stock return standard deviation σ_{it}.
	Table A-1 Panel B: First appearing in Table II
$BetaDec_{it}$	Firm i's NYSE decile based on daily market model beta calculated over two years ending on the last day of period t.
$Bk/MktDec_{it}$	Firm i's NYSE decile based on book-to-market from the most recent quarterly report before the beginning period t.
$MomDec_{it}$	Firm i's NYSE decile based on total return over the previous $t-12$ to $t-2$ months.
$AmiDec_{it}$	Firm i's NYSE decile based on daily Amihud ratio calculated over two years ending on the last day of period t.
$PrcDec_{it}$	Firm i's NYSE decile based on closing price on the last trading day of period $t-1$.
$InstDec_{it}$	Firm i's NYSE decile based on level of institutional holdings during period t.
$SP500_{it}$	Binary variable set to 1 if firm i is a member of the S&P 500 Index on the last trading day of period t.
$SizeDec_{it}$	Firm i's NYSE decile based on market value from the last trading day of period $t-1$.
	Table A-1 Panel C: First appearing in Box (2018)
ρ_{ijt}	Pearson daily return correlation between firms i and j during period t.
z_{ijt}	Fisher transformation of Pearson return correlation. Equal to $\frac{1}{2}\ln\frac{1+\rho_{ijt}}{1-\rho_{ijt}}$.
$BetaDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on daily market model beta calculated over two years ending on the last day of period t.
$BetaCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on daily market model beta calculated over two years ending on the last day of period t. $BetaCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$SizeDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on market value from the last trading day of period $t-1$.

$SizeCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on market value from the last trading day of period $t-1$. $SizeCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$Bk/MktDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on book-to-market from the most recent quarterly report before the beginning period t.
$Bk/MktCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on book-to-market from the most recent quarterly report before the beginning period t. $Bk/MktCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$MomDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on total return over the previous $t-12$ to $t-2$ months.
$MomCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on total return over the previous $t-12$ to $t-2$ months. $MomCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$IndDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same 49-industry portfolio, as defined on Kenneth French's website.
$IndCorr_{ijt}$	Each firm in the sample is assigned to one the 49 industry portfolios, as defined on Kenneth French's website. $IndCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
ρ_{ijt}^{1mo}	Pearson daily return correlation between firms i and j during the last month of period t.
ρ_{ijt}^{2mo}	Pearson daily return correlation between firms i and j during the last two months of period t.
$AnaDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on the number of unique analyst releasing an earnings forecast during period t.
$AnaCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on the number of unique analyst releasing an earnings forecast during period t. $AnaCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$InstDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on level of institutional holdings during period t.
$InstCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on level of institutional holdings during period t. $InstCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$AmiDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on daily Amihud ratio calculated over two years ending on the last day of period t.
$AmiCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on daily Amihud ratio calculated over two years ending on the last day of period t. $AnaCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$PrcDum_{ijt}$	Binary variable set to 1 if both firms i and j are members of the same NYSE decile portfolio based on closing price on the last trading day of period $t-1$.
$PrcCorr_{ijt}$	Each firm in the sample is assigned to NYSE decile portfolios based on closing price on the last trading day of period $t-1$. $PrcCorr_{ijt}$ is the daily return correlation between the portfolios containing firms i and j during period t.
$SP500_{ijt}$	Binary variable set to 1 if both firms i and j are members of the S&P 500 Index on the last trading day of period t.
$S34Sim_{ijt}$	Equal to $N_{ijt}^{inst}/\sqrt{N_{it}^{inst}N_{jt}^{inst}}$ where N_{ij}^{inst} is the number of institutions holding both firms i and j in a period t, and N_{it}^{inst} and N_{jt}^{inst} are the number of institutions holding firms i and j respectively.
$S12Sim_{ijt}$	Equal to $N_{ijt}^{mut}/\sqrt{N_{it}^{mut}N_{jt}^{mut}}$ where N_{ij}^{mut} is the number of mutual funds holding both firms i and j in a period t, and N_{it}^{mut} and N_{jt}^{mut} are the number of mutual funds holding firms i and j respectively.

$EPSSim_{ijt}$	Equal to $N_{ijt}^{an} / \sqrt{N_{it}^{an} N_{jt}^{an}}$ where N_{ij}^{an} is the number of analysts following both firms i and j in a period t, and N_{it}^{an} and N_{jt}^{an} are the number of analysts following firms i and j respectively.
$\max_{k \in i,j} Size_{kt}$	Standardized maximum market value between firms i and j on the last trading day of period $t-1$.
$\max_{k \in i,j} \sigma_{kt}$	For each period t, the daily return standard deviation is calculated for each firm i and j. $\max_{k \in i,j} \sigma_{kt}$ is the standardized maximum standard deviation between both firms.
$WireDum_{ijt}^{all}$	Binary variable has a value of 1 whenever both firms have some positive number of total words appearing on the Reuters Integrated Data Network.
$TakeSim_{ijt}^{all}$	Equal to $N_{ijt}^{take} / \sqrt{N_{it}^{take} N_{jt}^{take}}$ where N_{ij}^{take} is the number of takes that mention both firms i and j in a period t on the Reuters Integrated Data Network, and N_{it}^{take} and N_{jt}^{take} are the number of takes mentioning firms i and j, respectively.
$\widetilde{WireSim}_{ijt}^{all}$	Document similarity variable 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.
$\overline{WireSim_{ijt}^{all}}$	For each period in the sample, firms with some relevant text are classified into deciles based on total word counts. The variable $\overline{WireSim_{ijt}^{all}}$ represents the average document similarity between firms appearing in the same word count deciles as i and j during period t. The variable is constructed from all attributions appearing on the Reuters Integrated Data Network.
$WireSim_{ijt}^{all}$	Newswire similarity variable is calculated by subtracting $\overline{WireSim_{ijt}^{all}}$ from $\widetilde{WireSim}_{ijt}^{all}$.
$WireDum_{ijt}^{firm}$	Binary variable has a value of 1 whenever both firms have some positive number of total words originating from sources other than *Reuters News*.
$TakeSim_{ijt}^{firm}$	Equal to $N_{ijt}^{take} / \sqrt{N_{it}^{take} N_{jt}^{take}}$ where N_{ij}^{take} is the number of takes that mention both firms i and j in a period t originating from sources other than *Reuters News*, and N_{it}^{take} and N_{jt}^{take} are the number of takes mentioning firms i and j, respectively.
$\widetilde{WireSim}_{ijt}^{firm}$	Document similarity variable is the cosine similarity between the firm vectors i and j in the term-document matrix for period t constructed from sources other than *Reuters News*.
$\overline{WireSim_{ijt}^{firm}}$	For each period in the sample, firms with some relevant text are classified into deciles based on total word counts. The variable $\overline{WireSim_{ijt}^{firm}}$ represents the average document similarity between firms appearing in the same word count deciles as i and j during period t. The variable is constructed from sources other than *Reuters News*.
$WireSim_{ijt}^{firm}$	Newswire similarity variable is calculated by subtracting $\overline{WireSim_{ijt}^{firm}}$ from $\widetilde{WireSim}_{ijt}^{firm}$.
R_t^{Mkt}	Standardized cumulative return of the CRSP Market Weighted Index over period t.
σ_t^{Mkt}	Standardized daily return standard deviation of the CRSP Market Weighted Index during period t.
$\bar{\rho}_t$	Standardized sample average of all pairwise return correlations ρ_{ijt} in a given period t.