Internet Appendix: Time Variation in the News-Returns Relationship

Paul Glasserman* Fulin Li† Harry Mamaysky‡

July 24, 2023

Contents

*Columbia Business School, pg20@columbia.edu.
†Texas A&M University Mays Business School, fli3@mays.tamu.edu.
‡Columbia Business School, hm2646@columbia.edu.

A1 Mapping of news data to S&P 500 firms

A1.1 Map Thomson-Reuters articles to S&P 500 firms

To select Thomson Reuters (TR) articles that mention S&P 500 firms, we map CRSP PERMNO to Reuters Instrument Code (RIC), where RIC is the stock identifier from TR. Unfortunately, RICs are not unique identifiers, and we have not been able to obtain a historical RIC mapping from the company. This section gives the full details of our mapping from PERMNOs to RICs, and we summarize the process here: (1) Obtain the augmented article body by combining the headline and body text of an article; (2) Select articles that contain standardized S&P 500 company names (from the CRSP historical names table) in the augmented article body and associate these articles with S&P 500 PERMNOs; (3) For each PERMNO, find the top three most frequently occurring RICs in the selected articles, override unreliable RICs (i.e., those which do not occur sufficiently frequently) then fetch all articles tagged with these RICs; (4) Keep an article selected from (3) if it loosely mentions any S&P 500 company name. Steps (1)-(4) allow us to create a robust mapping from TR articles to S&P 500 firms.

A1.1.1 Create variants of company names

We create two variants of each S&P 500 company name, denoted as *Variant-1* and *Variant-2*. *Variant-1* is the pattern used in the first pass search and *Variant-2* is the pattern used in the second pass search. See Section A1.1.2 and Section A1.1.3 for the discussion of first and second pass search.

For each historical firm name, we perform steps 1-11 to get *Variant-1*, and steps 1-12 to get *Variant-2*. And we only keep unique *Variant-1* and *Variant-2* for each PERMNO.

1. Remove extra spaces between words.
2. Replace abbreviations in Table A1.
3. Replace ‘’ / - with space.
4. Replace & if it occurs between words while keep it if it occurs inside a word.
5. Remove all other punctuation marks and do not replace with space.
6. Remove the space which exists between two single word characters.
7. Remove all English stopwords except "under".[1]
8. Remove words in Table A2 directly (case-insensitive).
9. Remove words in Table A3 recursively (case-insensitive), i.e. starting from the last word in the company name, if it is in Table A3 then remove it. Loop until the last word is not in Table A3.
10. Convert all names to lower case.
11. Capitalize the first character of each word in company name.
12. Remove words in Table A4 recursively (case-insensitive).

[1] "Under Armour Inc" is an S&P 500 company in our sample, so we should not remove the stopword "under" from company names.

A1.1.2 First Pass Search

We first clean Thomson Reuters news data before searching for company names in article body. We drop non-English language articles or those with urgency < 2. We keep the first article within each article chain. Two articles belong to the same article chain if they have the same PNAC and have timestamps within the same 6-hour window in a day (we divide a day into four 6-hour windows). And we augment the article body with article headline.

Then we search for *Variant-1* in the augmented article body following the steps below:

1. Tokenize *Variant-1*.
2. Process the augmented article body as follows:
    (a) Replace ‘’ / - with space.
    (b) Replace & with space if it appears between words.
    (c) Replace . with space.
    (d) Remove all other punctuation marks.
    (e) Tokenize augmented article body and only keep non-empty tokens.
    (f) Convert all tokens to lower case. If the first character of a token is capitalized, keep the first character capitalized and convert all other characters to lower case.
3. Search for tokens of *Variant-1* in the augmented article body. An article is matched with *Variant-1* if all the conditions below are satisfied:
    (a) All tokens in *Variant-1* can be found in the text.
    (b) In the text, the last matched token and the first matched token are within 5 words of each other.
    (c) The order of tokens in *Variant-1* is preserved in the text.

If an article is matched with a *Variant-1* name and the associated PERMNO, then we say that all the RICs from that article is matched to the PERMNO. We then compute the frequency of each unique (PERMNO, RIC) pair and extract the top three frequently occurring RICs for each PERMNO. For a few PERMNOs, the top three PERMNO-RIC mapping are not robust, so we override the top three RICs with more reasonable ones.

A1.1.3 Second Pass Search

For each (PERMNO, RIC) pair, we search for the corresponding *Variant-2* in the augmented article body and only keep the matched articles after performing the following steps:

1. Tokenize *Variant-2*.
2. Process the augmented article body as follows:
   (a) Replace ‘’ / - with space.
   (b) Replace & with space if it appears between words.
   (c) Replace . with space.
   (d) Remove all other punctuation.
   (e) Tokenize augmented article body and only keep non-empty tokens.
   (f) Convert all tokens to lower case. If the first character of a token is capitalized, keep the first character capitalized and convert all other characters to lower case.
3. Search for tokens of *Variant-2* in the augmented article body. An article is matched with *Variant-2* if all the conditions below are satisfied:
   (a) The article is tagged with a top three frequently occurring RIC in its subject.
   (b) All tokens in *Variant-2* can be found in the text.
   (c) In the text, the last matched token and the first matched token are within 5 words of each other.
   (d) The order of tokens in *Variant-2* is preserved in the text.

A1.2 Mapping other news sources to PERMNOs

Since we have created a mapping from TR articles to PERMNOs, we need to map the alternative news archives (DJ and WSJ from RavenPack and the Financial Times) to PERMNOs as well. At that point, we can connect TR news with the alternative news sources via the PERMNO mapping.

We map RavenPack ENTITY IDs to CRSP PERMNOs through the following two steps. First, we match them based on the first six digits of their CUSIPs, which serve as unique identifiers for the companies in both data sources. Out of the 1,042 PERMNOs in the S&P 500 historical list, we identify a unique RavenPack ENTITY ID for each of the 648 PERMNOs. Second, in cases where a match is not found, we perform steps 1-11, as outlined in Section A1.1.1, for the company names in both RavenPack and CRSP datasets. By utilizing the tokenized names from both sources, we further map 90 RavenPack ENTITY IDs to CRSP PERMNOs on a one-to-one basis through exact matches. As a result of the aforementioned two steps, we obtain a total of 738 PERMNO-RavenPack ID pairs.

We map Financial Times (FT) articles to CRSP PERMNOs by linking CUSIPs in the CRSP dataset with the Financial Instrument Global Identifiers (FIGIs) associated with each FT article from the FT dataset. To match the two identifiers, we utilize the OpenFIGI API, which provides a list of related FIGIs for each CUSIP of companies in the S&P 500 historical list. By matching the FIGIs, we obtain a mapping between 1,005 PERMNOs and 251,345 FT articles, resulting in a total of 484,565 PERMNO-FT article pairs. To ensure the accuracy and reliability of the mapping based on CUSIPs, we collect, for each FIGI in the mapping, the company's ticker from the OpenFIGI API and its name from the FT dataset. We then filter and retain only the PERMNO-FT article pairs where the company's ticker and tokenized name match the ticker and name in the CRSP dataset on the published dates of the articles. This step results in a refined set of 337,918 PERMNO-FT article pairs.

We only keep day t articles about firms that are in S&P 500 on day t. Table A6 shows how many firm-day observations we have from each news source, as well as how much overlap there is in firm-day coverage.

A2 Data construction

A2.1 Constructing text measures

We construct two article-level text measures, sentiment and entropy, from our news data. Sentiment involves counting positive and negative words in articles, and entropy involves counting n-grams in the training corpus and the new text.

A2.1.1 Sentiment

For an article j, we first clean the augmented body text following steps 1-3 and 5 below. Then we do a case-insensitive search for positive and negative words in the augmented body using the Loughran and McDonald (2011) sentiment dictionary, and count the number of positive words n_j^{pos} and the number of negative words n_j^{neg} in article j. We also count the total number of words n_j in article j after we apply steps 1-4 to the augmented article body.

1. Convert the augmented body text to lower case.
2. Replace non-alphabet characters with space.
3. Tokenize the text.
4. Drop English stopwords.
5. Mark negation using the Das and Chen (2007) method. This appends the string `_NEG` to all words following a negating word until an end-of-statement punctuation mark. Such modified words are then ignored when calculating sentiment.

The sentiment of article j is defined as

$$Sent^j = \frac{n_j^{pos} - n_j^{neg}}{n_j}$$

A2.1.2 Entropy

We extract n-grams from each article following the steps below:

1. Convert the augmented body text to lower case.
2. Replace date strings, entity names, numerical strings and punctuation marks between sentences, as shown in Table A5 panel A-D.
3. Break the augmented body text into sentences by $***$.
4. Within each sentence, replace punctuation marks in Table A5 panel E.
5. Tokenize each sentence and stem the tokens.
6. Obtain the sequence of n-grams in the article, $n = 3, 4$.

We then count the frequency of each 3-gram and each 4-gram in articles of a given month. We define the training corpus for month t as articles in months $t-27, t-26, \cdots, t-$

4, and calculate the frequency of each 3-grams (4-gram) in the training corpus for month t. The entropy of article j in month t is defined as

$$\begin{aligned} Entropy_j &= -\sum_{i \in \text{4-grams}_j} \hat{p}_{i,j} \log \hat{q}_{i,j} \\ \hat{p}_{i,j} &= \frac{n_{i,j}}{\sum_{i \in \text{4-grams}_j} n_{i,j}} \\ \hat{q}_{i,j} &= \frac{\hat{c}_{t-27,t-4}\left(w_{1,i} w_{2,i} w_{3,i} w_{4,i}\right) + 1}{\hat{c}_{t-27,t-4}\left(w_{1,i} w_{2,i} w_{3,i}\right) + 10} \end{aligned}$$

where 4-grams_j is the set of distinct 4-grams in article j, $n_{i,j}$ is the count of 4-gram i in document j, $\hat{c}_{t-27,t-4}\left(w_{1,i} w_{2,i} w_{3,i} w_{4,i}\right)$ is the count of 4-gram i in the training corpus, $\hat{c}_{t-27,t-4}\left(w_{1,k} w_{2,k} w_{3,k}\right)$ is the count of the 3-gram associated with 4-gram i in the training corpus.

A2.2 Measuring passive and active ownership in stocks

We obtain mutual fund characteristics and holdings data from CRSP Survivor-Bias-Free US Mutual Fund database and Thomson Reuters (TR) Mutual Fund Holdings database. We identify index/passive mutual funds by searching for certain strings in CRSP fund names and supplement this information with the index fund indicator from CRSP.

We focus on US domestic equity mutual funds[2] from CRSP and classify them into passive, active or unclassified categories. For each CRSP fund, we do the following.

1. Fill in missing fund names using the most recently available one.
2. Replace the following characters in fund name with space: `˜! @ # $ % ˆ∗() _ + − = [ ] \{}|; : “ ” , . / <>?
3. Classify the fund based on the following criteria:
   (a) If the fund has a CRSP index fund indicator (index_fund_flag) in {B, D, E}, then it is a passive fund.
   (b) Otherwise,
      i. If the fund name includes a word/phase in {index, idx, indx, ind, russell, s␣&␣p, s␣and␣p, s&p, sandp, sp, dow, dj, msci, bloomberg, kbw, nasdaq,

[2] We focus on US domestic equity mutual funds because they have the most complete and reliable holdings data.

nyse, stoxx, ftse, wilshire, morningstar, 100, 400, 500, 600, 900, 1000, 1500, 2000, 5000}[3], then the fund is passive.

ii. Otherwise,

   A. If the fund has missing name and missing CRSP index fund indicator, then it is unclassified.

   B. In all other cases, the fund is active.

We then match CRSP funds to TR funds using the link tables from MFLINKS. MFLINKS maps CRSP funds and TR funds to a common Wharton Financial Institution Center Number (WFICN), which uniquely identifies a fund.[4] Finally, we map TR fund holdings to CRSP stocks by historical CUSIP, and construct the mutual fund holdings dataset at fund-stock level.

A2.3 Trimming mutual fund ownership variables

Figure A1 depicts the cross-sectional correlations between the passive and active ownership series. The top panel shows the correlations using all available data. The three correlations spike in early 2011. For example, Corr(Passive/Market, Active/Market) increases from 0.2573455 to 0.5809107 in the first quarter of 2011. This pattern is caused by outliers in terms of Passive/Market and Active/Market values. From Q1 2011 onward, we have stocks with very few mutual fund holders and their Passive/Market and Active/Market are close to zero, which drives up the correlations between the passive and active series. In the bottom panel of Figure A1, we exclude the bottom 2.5% observations of each series and recompute their correlations. We no longer see the spikes in the correlations. We discuss this further in Section A6.5.

[3] ␣ denotes a space character.

[4] CRSP mutual fund data is at the share-class level, so there could be multiple CRSP funds associated with the same WFICN and they all have the same holdings. We only keep one CRSP fund for each WFICN.

A3 Model development

A3.1 One-period model

Assume a one period model, where an agent (indexed by i) solves the following mean-variance portfolio problem with benchmarking penalty

$$\max_{w} w^{\top}\left(\mu_i - P\right) - \frac{\gamma_i}{2} w^{\top} \Sigma w - \frac{1}{2}(w - x)^{\top} \Lambda_i \left(w - x\right) \tag{A1}$$

for $w, x, \mu_i, P \in \mathbb{R}^N$. P is the vector of security prices, w is the agent's portfolio holdings, μ_i is agent i's expectations about end-of-period security values, and Σ is the covariance matrix of end-of-period security values conditional on the investor's information set. The vector x captures the benchmark target, and $\Lambda_i \in \mathbb{R}^{N \times N}$ is a symmetric matrix which represents the deviation penalty. $\gamma_i \geq 0$ is the investor's risk aversion.

The first-order condition for the problem is

$$\mu_i - P - \gamma_i \Sigma w - \Lambda_i \left(w - x\right) = 0.$$

Rearranging we find

$$w_i = \left(\gamma_i \Sigma + \Lambda_i\right)^{-1} \left(\mu_i - P + \Lambda_i x\right). \tag{A2}$$

The price elasticity of demand is

$$\frac{\partial w_i}{\partial P} = -(\gamma_i \Sigma + \Lambda_i)^{-1}, \tag{A3}$$

so higher benchmarking penalty *or* higher risk aversion play a similar role of decreasing price elasticity. Note (A3) is also the elasticity of demand with respect to an investor's beliefs about future returns μ_i.

Market clearing requires

$$\sum_i \phi_i w_i = S,$$

where ϕ_i is the fraction of the population represented by the i-th investor with $\sum_i \phi_i = 1$, and $S \in \mathbb{R}^N$ is the supply of shares. Using (A2) we get

$$\sum_i \phi_i \left(\gamma_i \Sigma + \Lambda_i\right)^{-1} \left(\mu_i - P + \Lambda_i x\right) = S. \tag{A4}$$

After rearranging

$$\sum_i \phi_i \left(\gamma_i \Sigma + \Lambda_i\right)^{-1} \left(\mu_i + \Lambda_i x\right) - S = \sum_i \phi_i (\gamma_i \Sigma + \Lambda_i)^{-1} P.$$

Assume that Σ is a diagonal matrix with all entries equal to σ^2. And assume that the benchmarking penalty $\Lambda_i = \lambda_i I$ for $\lambda_i \in \mathbb{R}$ and I an $N \times N$ identity matrix. The market clearing condition for stock $j \in \{1, \ldots, N\}$ is therefore

$$\sum_i \frac{\phi_i}{\gamma_i \sigma^2 + \lambda_i} \left(\mu_{ij} + \lambda_i x_j\right) - S_j = P_j \sum_i \frac{\phi_i}{\gamma_i \sigma^2 + \lambda_i}.$$

The risk premium in the price is not central to the present analysis, so we set $S_j = x_j = 0, \forall j$. With this we get the following equation for the equilibrium price of security j

$$\sum_i \frac{\phi_i}{\gamma_i \sigma^2 + \lambda_i} \mu_{ij} = P_j \sum_i \frac{\phi_i}{\gamma_i \sigma^2 + \lambda_i},$$

from which we get

$$P_j = \left(\sum_i \frac{\phi_i}{\gamma_i \sigma^2 + \lambda_i} \right)^{-1} \sum_i \frac{\phi_i}{\gamma_i \sigma^2 + \lambda_i} \mu_{ij}. \tag{A5}$$

Since the supply of shares and the benchmark target x_j are zero, there is no risk discount in the price and P_j is simply the weighted average investor belief about future payoff of stock j.

We now specialize the equilibrium to three types of investors. A fraction ϕ_1 represents non-institutional investors. These investors have $\gamma_1 = 1$ (without loss of generality) and face no benchmarking constraints on portfolio holdings so $\lambda_1 = 0$. A fraction ϕ_2 of investors are financial intermediaries. They are less risk-averse than non-institutional investors, so $\gamma_2 = \gamma < 1$ and face no benchmarking restrictions, so $\lambda_2 = 0$. Shin (2009) shows that a VaR constraint leads to an equivalent optimization problem to (A1) where γ represents the shadow cost of the VaR constraint. We can therefore interpret γ for financial intermediaries as a measure of the degree to which they are constrained in their risk-taking activities. Finally, passive institutional investors, or indexers, have zero risk-aversion but face a benchmarking restriction $\lambda_3 > 1$, because the i-th indexer is required by mandate to not deviate too far away from its benchmark index x (assumed to be zero). The degree to which the i-th investor is constrained is given by $\gamma_i \sigma^2 + \lambda_i$, so assuming $\sigma^2 \approx 1$, we can say that intermediaries are less constrained than non-institutional investors, and non-institutional investors are less constrained than indexers. The key features of the three

Investor	Weight	Risk Aversion	Benchmarking	Constrained
Non-institutional	ϕ_1	$\gamma_1 = 1$	$\lambda_1 = 0$	Medium
Intermediaries	ϕ_2	$\gamma_2 < 1$	$\lambda_2 = 0$	Least
Passive	ϕ_3	$\gamma_3 = 0$	$\lambda_3 > 1$	Most

types of investors can be summarized as follows:

We now specify how each investor group updates beliefs based on news. Assume that I_j represents good news about security j, and that $\mu_{ij} = f_i(I_j, \dots)$ with $\partial\mu_{ij}/\partial I_j > 0$, where the $\dots$ indicate that beliefs can depend on other factors besides news. We assume that non-institutional investors and intermediaries update their beliefs in proportion to the information content τ of news, so

$$\frac{\partial \mu_{ij}}{\partial I_j} = f(\tau) \qquad \text{for } i \in \{1, 2\},$$

where $f(\tau) \in (0, 1)$ and is increasing in τ. Furthermore, we assume that $\sigma^2 = \sigma^2(\tau) \in (0, 1)$ for $i = 1, 2$ and that $\sigma^2(\tau)$ is decreasing in the information content τ of news.[5] Indexers don't update beliefs in response to news, so $\partial\mu_{3j}/\partial I_j = 0$. As in (A1), this reflects institutional constraints on the behavior of indexers.

Using the price from (A5) we find that the sensitivity of P_j to news is

$$\frac{\partial P_j}{\partial I_j} = \frac{\frac{\phi_1}{\sigma^2(\tau)} + \frac{\phi_2}{\gamma_2\sigma^2(\tau)}}{\frac{\phi_1}{\sigma^2(\tau)} + \frac{\phi_2}{\gamma_2\sigma^2(\tau)} + \frac{\phi_3}{\lambda_3}} f(\tau).$$

Note that the denominator is positive, and that $\phi_1 = 1 - \phi_2 - \phi_3$. Making this substitution and multiplying by σ^2 we find:

$$\begin{aligned} \frac{\partial P_j}{\partial I_j} &= \frac{1 - \phi_2 - \phi_3 + \frac{\phi_2}{\gamma_2}}{1 - \phi_2 - \phi_3 + \frac{\phi_2}{\gamma_2} + \frac{\phi_3}{\lambda_3}\sigma^2(\tau)} f(\tau), \\ &= \frac{1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 - \phi_3}{1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 + \frac{\sigma^2(\tau)-\lambda_3}{\lambda_3}\phi_3} f(\tau). \end{aligned} \tag{A6}$$

The following properties hold in equilibrium:

Proposition 1. *Price sensitivity to news increases with more intermediaries and decreases with more passive investors.*

[5] This would happen under normality if news I_j consisted of the dividend plus noise, and τ was the precision of the noise term – though this argument ignores the equilibrium effect of the impact of precision on the informativeness of prices. But I think it would still go through even in equilibrium.

Proposition 2. *Price sensitivity to news increases when intermediaries are less constrained, i.e., have lower γ.*

Proposition 3. *Price sensitivity to news increases as the information content of news grows.*

To check Proposition 3, observe that higher τ increases $f(\tau)$ and decreases $\sigma^2(\tau)$ and both effects tend to increase $\partial P_j / \partial I_j$.

To check Proposition 2, note that in the top expression in (A6) both the numerator and denominator increase by the same amount when γ falls; since both are positive and the numerator is smaller than the denominator, this increases the price sensitivity, i.e., for $x, y > 0$, $\frac{d}{dx}(\frac{x}{x+y}) = \frac{1}{x+y} - \frac{x}{(x+y)^2} = \frac{1}{x+y}(1 - \frac{x}{x+y}) > 0$.

In what follows, we drop $f(\tau)$ from (A6) since it's positive and does not affect the sign of the derivatives. To check Proposition 1, note that:

$$\frac{\partial}{\partial \phi_2}\left(\frac{\partial P_j}{\partial I_j}\right) = \frac{\frac{1-\gamma_2}{\gamma_2}}{1 + \frac{1-\gamma}{\gamma}\phi_2 + \frac{\sigma^2 - \lambda}{\lambda}\phi_3} - \frac{1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 - \phi_3}{(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 + \frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}\phi_3)^2} \frac{1-\gamma_2}{\gamma_2}.$$

Since the denominator is positive, the sign of this is the same as the sign of

$$\left(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 + \frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}\phi_3\right) - \left(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 - \phi_3\right) = \sigma^2(\tau)\phi_3 > 0.$$

To check that more indexers decrease the price sensitivity to news, note that

$$\frac{\partial}{\partial \phi_3}\left(\frac{\partial P_j}{\partial I_j}\right) = -\frac{1}{1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 + \frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}\phi_3} - \frac{1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 - \phi_3}{(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 + \frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}\phi_3)^2} \frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}.$$

Since the denominator is positive, the sign of this is the same as the sign of

$$-\left(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 + \frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}\phi_3\right) - \left(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 - \phi_3\right)\frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}.$$

Since $\lambda_3 > 1$ and $\sigma^2(\tau) < 1$ the sign of the second term above is positive and the entire expression is therefore less than

$$-\left(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 + \frac{\sigma^2(\tau) - \lambda_3}{\lambda_3}\phi_3\right) + \left(1 + \frac{1-\gamma_2}{\gamma_2}\phi_2 - \phi_3\right) = -\sigma^2(\tau)\phi_3 < 0.$$

A3.2 Two-period model

We extend the model to two periods, which allows us to make predictions on price *underreaction* to news, conditional on investor composition, intermediary constraints, and news informativeness. The key feature of the two-period model is based on Hong and Stein (1999, HS): we assume that non-institutional investors and intermediaries behave like *newswatchers* in HS: they "formulate their asset demands based on the static-optimization notion that they buy and hold until the liquidating dividend" and "they do not condition on current or past prices." HS refer to this as a "Walrasian equilibrium with private valuations," as opposed to a rational expectations equilibrium. HS motivate this behavior as a simple form of bounded rationality (see their discussion on page 2149). This assumption can be thought of as a reduced form version Sims' (2011) rational inattention. Since we are focused on price responses to public news over a short time interval, we are not concerned about overreaction, and so do not introduce HS's momentum traders.

We assume that a fraction $\theta \in [0,1]$ of the non-institutional and intermediary sector pays attention to stock j in period 0, and the remaining fraction $1-\theta$ pays attention to the stock in period 1. As in HS, time 0 investors remain in the market in period 1. Investors who pay attention to stock j receive the same public signal I_j. We interpret θ as the technological capacity constraint faced by investors. As technology improves, investors are able to follow more stocks, and with respect to stock j, a greater fraction of investors is able to follow the stock in period 0. Stock j pays a liquidating dividend in period 2. Our newswatchers optimize objective function (A1) using their information I_j with respect to the liquidating dividend. As in HS, they do not condition on prices. The $1-\theta$ fraction of investors who do not follow stock j in period 0 simply stay out of the market for j until period 1 – they do not have capacity to devote to following stock j in period 0. The indexers behave as before. We assume $\mu_{1j} = \mu_{2j} = \mu$ and $\mu_{3j} = 0$ for simplicity.

Given our assumptions, the period 1 price is the same as the price in the one-period model, and is given by (A5), i.e.,

$$P_{1j} = \frac{\frac{\phi_1}{\sigma^2}\mu + \frac{\phi_2}{\gamma_2\sigma^2}\mu}{\frac{\phi_1}{\sigma^2} + \frac{\phi_2}{\gamma_2\sigma^2} + \frac{\phi_3}{\lambda_3}} = \frac{\phi_1\mu + \frac{\phi_2}{\gamma_2}\mu}{\phi_1 + \frac{\phi_2}{\gamma_2} + \frac{\phi_3}{\lambda_3}\sigma^2}.$$

The period 1 price reflects *all* information. Given the HS assumptions, the period 0 equilibrium is identical to the period 1 equilibrium, except the fraction of non-institutional and intermediary investors is given by $\theta\phi_1$ and $\theta\phi_2$. The period 0 price is therefore

$$P_{0j} = \frac{\phi_1 \mu + \frac{\phi_2}{\gamma_2}\mu}{\phi_1 + \frac{\phi_2}{\gamma_2} + \frac{\phi_3}{\lambda_3}\frac{\sigma^2}{\theta}}. \tag{A7}$$

The HS newswatcher assumption leads to a very tractable equilibrium, and, as in their paper, there is price underreaction. To see this note that

$$P_{0j} = \alpha(\theta) P_{1j} \qquad \text{and} \qquad P_{1j} - P_{0j} = (1 - \alpha(\theta)) P_{1j} \tag{A8}$$

where

$$\alpha(\theta) = \frac{\phi_1 + \frac{\phi_2}{\gamma_2} + \frac{\phi_3}{\lambda_3}\sigma^2}{\phi_1 + \frac{\phi_2}{\gamma_2} + \frac{\phi_3}{\lambda_3}\frac{\sigma^2}{\theta}} \qquad \text{and} \qquad \alpha(\theta) \in [0, 1].$$

Note that $\alpha(\theta)$ is increasing in θ. Therefore, when news I_j arrives, the period 0 price reaction will be smaller than the full (period 1) price reaction, and the price change from period 0 to period 1, $P_{1j} - P_{0j}$, will be nonzero and will go in the same direction as the period 0 price response:

$$\frac{\partial}{\partial I_j}(P_{1j} - P_{0j}) > 0.$$

In light of (A8) the following proposition is immediate:

Proposition 4. *Propositions 1, 2, and 3 all apply to the period 0 price response to news and to the period 1 return $P_{1j} - P_{0j}$ in response to news.*

Finally, if technology improves, i.e., as θ increases, the model makes an unambiguous prediction:

Proposition 5. *As θ increases, the period 0 price response to news $\partial P_{0j}/\partial I_j$ increases, and the period 1 price change in response to news $\partial(P_{1j} - P_{0j})/\partial I_j$ decreases.*

This follows from (A8) and the fact that $\alpha(\theta)$ is increasing in θ.

Of course, allowing the period 1 investors to participate in period 0 trading while conditioning on the period 0 price of j, or allowing for arbitrageurs who can profit from understanding the dynamics of the model, would make our results less stark. But the main intuition of price underreaction and its dependence on technological constraints would remain.

A3.2.1 Period 0 and 1 holdings by intermediaries

From (A2), the demands of the three investor types for stock j are

$$\begin{aligned}
w_{1j} &= \frac{1}{\sigma^2}(\mu_j - P_j) \\
w_{2j} &= \frac{1}{\gamma_2\sigma^2}(\mu_j - P_j) \\
w_{3j} &= -\frac{1}{\lambda_3}P_j,
\end{aligned}$$

where we assume that $\mu_{1j} = \mu_{2j} = \mu_j > 0$ and $\mu_{3j} = 0$. Market clearing thus requires that

$$\theta\phi_1 w_{1j} + \theta\phi_2 w_{2j} + \phi_3 w_{3j} = S_j > 0.$$

Plugging in the above demands, P_j must satisfy

$$\theta\frac{\phi_1}{\sigma^2}(\mu_j - P_j) + \theta\frac{\phi_2}{\gamma_2\sigma^2}(\mu_j - P_j) - \frac{\phi_3}{\lambda_3}P_j = S_j.$$

Rearranging we find that

$$\begin{aligned}
P_j &= \left(\theta\frac{\phi_1}{\sigma^2} + \theta\frac{\phi_2}{\gamma_2\sigma^2}\right)\left(\theta\frac{\phi_1}{\sigma^2} + \theta\frac{\phi_2}{\gamma_2\sigma^2} + \frac{\phi_3}{\lambda_3}\right)^{-1}\mu_j - \left(\theta\frac{\phi_1}{\sigma^2} + \theta\frac{\phi_2}{\gamma_2\sigma^2} + \frac{\phi_3}{\lambda_3}\right)^{-1}S_j \\
&= \frac{\theta\phi_1\gamma_2\lambda_3 + \theta\phi_2\lambda_3}{\theta\phi_1\gamma_2\lambda_3 + \theta\phi_2\lambda_3 + \phi_3\gamma_2\sigma^2}\mu_j - \frac{\gamma_2\sigma^2\lambda_3}{\theta\phi_1\gamma_2\lambda_3 + \theta\phi_2\lambda_3 + \phi_3\gamma_2\sigma^2}S_j.
\end{aligned}$$

The period 0 demand of the intermediary sector $X_2(\theta) = \theta\phi_2 w_2$ is therefore given by

$$\begin{aligned}
X_{2j}(\theta) &= \theta\phi_2\frac{1}{\gamma_2\sigma^2}(\mu_j - P_j) \\
&= \theta\phi_2\frac{1}{\gamma_2\sigma^2}\frac{\phi_3\gamma_2\sigma^2}{\theta\phi_1\gamma_2\lambda_3 + \theta\phi_2\lambda_3 + \phi_3\gamma_2\sigma^2}\mu_j + \frac{\theta\phi_2\lambda_3}{\theta\phi_1\gamma_2\lambda_3 + \theta\phi_2\lambda_3 + \phi_3\gamma_2\sigma^2}S_j \\
&= \frac{\phi_2\phi_3}{\phi_1\gamma_2\lambda_3 + \phi_2\lambda_3 + \phi_3\gamma_2\sigma^2/\theta}\mu_j + \frac{\phi_2\lambda_3}{\phi_1\gamma_2\lambda_3 + \phi_2\lambda_3 + \phi_3\gamma_2\sigma^2/\theta}S_j.
\end{aligned}$$

Given that all quantities in $X_{2j}(\theta)$ are positive, it is easy to see that $\partial X_{2j}(\theta)/\partial\theta > 0$, and therefore $X_{2j}(1) - X_{2j}(\theta) > 0$ for $\theta < 1$. So the intermediary sector adds to its holdings of stock j in period 1. Furthermore, since $\partial\mu_j/\partial I_j > 0$, we will have that $\partial(X_{2j}(1) - X_{2j}(\theta))/\partial I_j > 0$, meaning that with good news, intermediaries increase their period 1 buying by even more.

A4 Magnitude of the effect

A4.1 Trading simulations

Tables A7 and A8 show that the results discussed in Section 5.4 are robust against different choices of *scale* in (18).

A4.2 Impulse response functions

We calculate impulse response functions to a sentiment shock using the local projection method of Jorda (2005). We run regressions (13) and (14) in the paper with the left hand side one-day returns or CARs on the event day t, day $t+1, t+2, \ldots, t+40$. The day t (contemporaneous) regression uses the 4pm–4pm sentiment, and excluded the contemporaneous abnormal return $CAR_{0,0}$ as an explanatory variable. We calculate the impulse response as the value of a hypothetical $100 portfolio invested for each day at that day's forecasted incremental return due to a unit sentiment shock. This assumes the sentiment shock under consideration has been orthogonalized to all other contemporaneous influences.

To calculate the cumulative baseline response for day h we add up all single day $Sent$ coefficients up to and including $t+h$, scaled by a one standard deviation sentiment shock.[6] Standard errors are calculated assuming each one-day return is independent, and using the one-day return standard errors (clustered by time) from the panel regressions in (13) and (14).

To calculate the price response to a sentiment shock conditional on a one standard deviation increase in intermediary capital or decrease in passive ownership we add the $Sent \times Capacity$ or subtract the $Sent \times Ownership$ interaction, scaled by a one standard deviation change in $Sent$ times a one standard deviation change in the interacting variable, to each day's forecasted marginal return. For calculating standard errors for conditional responses, we assume the marginal $Sent$ response and the interacted response are independent.

Figure A4 shows the impulse response function of future excess returns (panel A) and CARs (panel B) to a one standard deviation sentiment shock conditional on average passive/total ownership (solid line). Also shown is the impulse response conditional on a one standard deviation decrease in passive/total ownership (dashed line). Figure 7 in the

[6] Calculating the geometric return, i.e. $100 \times (1 + E[r_t]) \times (1 + E[r_{t+1}]) \times \cdots$ yields an almost identical result.

main body of the paper shows the responses to sentiment and sentiment interacted with intermediary capacity.

A5 Alternative news sources

A5.1 Comparing the coverage by different news sources

Figure A6 shows the differences in the composition of firms covered by the Dow Jones, Wall Street Journal, Financial Times, and Thomson Reuters. Figure A7 shows the differences in industry coverage across these news sources.

To plot the Average Market Cap in Figure A6, we compute the average of the log-transformed market capitalization of each company in each quarter. The average is weighted by the number of days where the company has at least one article in each of the four news sources. For the quarterly number of firm-day observations in Figure A6, we count the total number of company-day observations where the company has at least one article in one day in each of the four news sources for each quarter. For the remaining graphs in Figure A6, we fit the Fama-French (2015) five-factor model augmented with momentum for each company using daily returns over the full sample. For each news source, we calculate the average of their R-squareds and each of the coefficients in each quarter, weighted by the number of days where the company has at least one article in the given news source.

For Figure A7, we count two sets of company-day observations for each of the four news sources. First, we count the number of company-day observations as described in the quarterly firm-day observations in Figure A6. Second, we count the number of company-day observations in a given quarter about companies in each of the 12 industries based on classification from Ken French's website. We then plot the fraction of articles about companies in each industry by dividing the second count (company-day observations in a sector) by the first count (total company-day observations) for each of the four news sources.

A6 Robustness checks

A6.1 Earnings forecastability by news

TSM argue that news informativeness can be measured by the degree to which earnings surprises are forecastable by lagged news sentiment. We use this insight as a check of robustness of entropy as an indicator of news informativeness. As in TSM, we use two measures of earnings surprise: standardized unexpected earnings (SUE) and standardized analysts' forecast errors (SAFE). Our construction of SUE is explained in Section 2.2. We compute standardized analysts' forecast errors (SAFE) as the difference between actual earnings per share and the median of analyst forecasts made within the $[-30, -3]$ trading day window prior to the earnings announcement, divided by the standard deviation of unexpected earnings. We use a $[-30, -3]$ trading day window to avoid stale analyst forecasts and a potentially inaccurate earnings announcement date. We also include analyst forecast revisions and forecast dispersion as controls. Forecast revision is the sum of changes in the median analyst's forecast of earnings-per-share (EPS) scaled by the stock price at the end of the prior month, with the sum taken from the prior earnings announcement to the current one. Forecast dispersion is the standard deviation of EPS forecasts (either confirmed or revised) from the prior earnings announcement date to the current one, scaled by the same σ_q used to calculate SUE.[7] We winsorize SUE and SAFE at the 5% level for the earnings regressions, as we winsorize forecast dispersion and forecast revisions at the 1% level.[8]

Figure A8 plots the quarterly cross-sectional standard deviations of SUE and SAFE over time. The figure shows considerable time variation in these measures, indicating time variation in the baseline predictability of earnings. The standard deviation of earnings surprises peaks around the time of the global financial crisis.[9] Table A9 shows summary statistics of the firm-quarter earnings regression variables.

For our earnings regressions, we calculate news sentiment in the month prior to the earnings release. More specifically, our news sentiment measure is the average of the sentiment scores of individual articles mentioning company j within a $[-30, -3]$ trading day window prior to the earnings announcement date t, weighted by the number of words in

[7] Not using a three trading day lag with regard to forecast revisions and dispersion is conservative because it means our sentiment measure is lagged relative to the controls.

[8] Winsorization at the $X\%$ level means setting all observations above (below) the $100 - X/2$ $(X/2)$ percentile to that percentile's value.

[9] The spike in both series in 1Q2018 is due to the very low number of observations we have for that quarter. The spike in the cross-sectional standard deviation of SUE in 4Q2017 is due to the recognition of large, one-time gains (losses) on deferred tax liabilities (assets) as a result of the Tax Cut and Jobs Act of 2017. For example, in their 2017 Annual Report, the CME Group said that "2017 net income included a $2.6 billion net income tax benefit due to recognition of a reduction in deferred tax liabilities as a result of the Tax Cut and Jobs Act of 2017." This gain was recognized in their 4Q2017 earnings. In 4Q2017, the standard deviation of $SAFE$ shows no commensurate increase, as analyst expectations already incorporated these effects. Excluding 4Q2017 and 1Q2018 from our sample does not meaningfully affect our results in Table A10 (discussed below), as Table A11 shows. Also, the results in Figure A9 (discussed in Section 5.3) are not impacted by the exclusion of these quarters.

each article.[10] We lag the sentiment window by three days because of potential uncertainty as to the accuracy of the earnings announcement date.[11] While an earnings event on trading day t will enter our sample only if company j was a member of the S&P500 index on day t, we will use articles about j in the $[t-30, t-3]$ trading day window even if the company was not a member of the S&P500 on those days, as long as the articles satisfy the ≤ 7 RICs and ≥ 25 word requirements.[12] Our return controls in the earnings regressions are from trading day $t-2$ and the $[t-30, t-3]$ trading day window prior to the earnings announcement date t. Our other control variables are from the month prior to the earnings announcement month.

Our earnings regressions take the form

$$SUE^i_{t+1} \text{ or } SAFE^i_{t+1} = s_0 \times Sent^i_t + \boldsymbol{\beta}'\boldsymbol{X}^i_t + \epsilon^i_t, \tag{A9}$$

using quarterly data. The sentiment measure $Sent^i_t$ is stock i's average sentiment in the month preceding the announcement date of quarter $t+1$ earnings, as described in Section 2. We use the same controls $\boldsymbol{X}^i_t$ in both regressions, except that we include lagged SUE (but not lagged SAFE) in the SUE regression, and we include lagged SAFE (but not lagged SUE) in the SAFE regression. The controls are the most recently available observations in the month prior to the announcement date of quarter $t+1$ earnings. The other controls and their summary statistics are shown in Table A9. Standard errors for the earnings regressions are clustered by quarter.

Table A10 summarizes the results of this analysis.[13] The table reports the $Sent^i_t$ coefficient s_0 for the SUE and SAFE regressions for the same time periods we used in our return regressions. First, the results confirm that for the full time period and in most subperiods, sentiment is a significant predictor of earnings, and the fact that SAFE is forecastable by lagged sentiment indicates that analysts do not fully incorporate the information in news sentiment into their forecasts.[14] The informativeness of news, as measured by the magnitudes and significance of the coefficients in Table A10, has varied over time. There is little evidence of either an upward or downward secular trend in the

[10] We also ran the analysis in Section 5.3 using an equally-weighted $[-30, -3]$ trading day news sentiment measure. The results were qualitatively similar. We use the word-weighting to be consistent with TSM.

[11] TSM point out that "Compustat earnings announcement dates may not be exact." Though we use announcement dates from I/B/E/S we follow the TSM convention to be conservative.

[12] Restricting the analysis to articles only on days when company j was a member of the S&P500 index does not change the results.

[13] Table A12 of the shows the complete full-sample regression results.

[14] Prior work has found evidence for both underreaction and overreaction to news by analysts; see Abarbanell and Bernard (1992) and Easterwood and Nutt (1999).

s_0 estimates.

A6.1.1 Annual entropy analysis

Our basic mode of analysis is to compare the sentiment coefficient in annual regressions of one-day abnormal returns – for $CAR_{1,1}$ in equation (2) and for $CAR_{0,0}$ in equation (3) – against an annual measure of news informativeness.[15] Panel A of Figure A9 shows the s coefficient from an annual regression of $CAR_{0,0}$ on contemporaneous sentiment and control variables plotted against the average entropy of all articles that appeared in that year. There is an economically and statistically significant relationship between average article informativeness and the magnitude of the contemporaneous return response to news. This is strongly supportive of the news information hypothesis.

That more informative news flow has a larger contemporaneous price effect is not surprising. But how does this relate to stock underreaction to news? Sims (2003) and a large subsequent literature propose that investors have a limited capacity to process information. This information capacity constraint should become more binding when there is more information to process. With a more binding constraint, market participants take longer to react to value-relevant news, and stock prices should therefore react more to lagged news during high information periods. Panel B of Figure A9 shows the $CAR_{0,0}$ sentiment coefficient from Panel A, but this time plotted against the sentiment coefficient from the $CAR_{1,1}$ regression on lagged sentiment from (2). Years when prices have relatively large reactions to one-day lagged news are also years when stock prices are very responsive to contemporaneous news. This is indirect supportive of the limited capacity hypothesis.

Panel C of Figure A9 offers further evidence for the hypothesis. It shows the sentiment coefficients from annual $CAR_{1,1}$ regressions plotted against annual average entropy. There is an economically and statistically significant relationship between the tendency of stocks to underreact to news (and thus for returns to load positively on one-day lagged news) and our entropy measure of informativeness. In time periods of more informative news flow, stocks have stronger reactions to contemporaneous news and also have stronger reactions to lagged news.

[15] The results for $CAR_{1,10}$ are qualitatively similar to the results for $CAR_{1,1}$.

A6.1.2 Annual entropy analysis and earnings forecastability

As a robustness check of entropy as a measure of news informativeness, Panel D of Figure A9 shows the s_0 coefficient from annual versions of the SUE regression in (A9) plotted against annual average entropy. The two series are highly correlated, supporting our interpretation of both as measures of news informativeness.[16] Panel E shows the annual $CAR_{0,0}$ sensitivity to contemporaneous news plotted against the annual SUE-sentiment coefficient s_0 from (A9). In years when news is more informative about earnings surprises, stock prices have a stronger reaction to contemporaneous news. Together, Panels D and E support our interpretation that both entropy and the earnings coefficient s_0 from (A9) proxy for the information content of news.

To test whether news-earnings informativeness and the degree of stock-news underreaction are related, Panel F plots the sentiment coefficient from the $CAR_{1,1}$ regression in (2) against the sentiment coefficient s_0 from the earnings regression in (A9). There is no relationship between the informativeness of news for earnings and the degree of stock-news underreaction. Apparently, the portion of news that is informative about future earnings gets quickly absorbed into prices (Panel E) and there is little left for future prices to react to (Panel F). The correlation of entropy and s_0 (Panel D) suggests that entropy captures components of news flow that are relevant for near-term earnings. But entropy also captures other components of news flow, and these latter components seem to be related to price underreaction to news. A better understanding of the components of news flow is an interesting area for future study; Glasserman et al. (2020) is a step in this direction.

A6.2 Short-selling constraints

When a company experiences surprisingly bad news, some market participants may short its stock, anticipating and contributing to a decline in the stock price. However, in a mechanism described by Miller (1977), they may not be able to sell short the desired amount of stock if doing so is costly, and this constraint may slow the process by which bad news gets incorporated in the stock price, causing an underreaction. If short-sale constraints fully explain the underreaction, we should (1) observe more underreaction in stocks with more binding short-sale constraints, and (2) only observe underreaction when bad news comes out. Since short-sale constraints are plausibly related to intermediary capital and to the presence of institutional owners in a stock, such constraints are a

[16] The annual SAFE s_0 coefficient is also positively correlated with annual average entropy.

tempting explanation for our finding of a systematic relationship between intermediary capital and institutional ownership and underreaction.

To test these two hypotheses, we group stocks by the tightness of their short-sale constraints and the tone of the news, then examine which group exhibits more underreaction. We use measures of short interest and institutional ownership to proxy for the short-sale constraints. Asquith, Pathak and Ritter (2005) posit that short interest captures the short-sale demand, and institutional ownership is a proxy for the supply of lendable shares. Stocks with the highest short interest and the lowest institutional ownership will have the most binding short-sale constraints. The short interest variable (SI) is defined in Section 2.2. For institutional ownership, we use the residual measure (RI) introduced by Nagel (2005), which adjusts for size. We first perform a logit transformation on the institutional ownership variable (IO) from Section 2.2. Then for each quarter, we regress the transformed variable on log market cap and squared log market cap. The RI measure is defined as the residual from this regression. Nagel (2005) argues that, because institutional ownership and firm size are highly correlated, sorting on IO is akin to sorting on size. To capture the effect of IO on short-sale constraints, it is therefore necessary to take out the firm size effect.[17]

The first hypothesis implies that stocks with higher SI and lower RI should exhibit more underreaction. So we first sort stocks by SI and RI. Specifically, for each month, we obtain the median SI and the median RI across all stock-day observations within that month. Using these cutoffs, we double sort the stocks by SI and RI independently. In each bucket, we run the main specification in equation (2). We are interested in the coefficient s, which captures the stock price underreaction to news.

Table A13 panel A shows the coefficient estimates. We focus on the responses of cumulative abnormal returns, though the excess return results are similar. Over both the one-day horizon and the ten-day horizon, stocks with high SI and *high* RI exhibit the largest magnitude of underreaction, and the underreaction is highly significant at the 1% level. These stocks have less binding short-sale constraints than those with high SI and low RI, yet they show more underreaction. In fact, low SI and low RI stocks, again not the short-sale constrained group, also show more underreaction than the high SI and low RI ones. This contradicts the first prediction of the short-sale constraint story.

We then test the second prediction on the asymmetric response to good news versus bad news. To that end, we obtain the median sentiment, across all stock-day observations

[17] In unreported results, we use IO instead of RI in the double and triple sorts below, and also try dependent sorts instead of independent sorts. The results remain qualitatively the same.

within a month. Using this sentiment cutoff and the SI and RI cutoffs from above, we triple sort stocks by sentiment, SI, and RI independently, and run the main specification in equation (2) for each bucket. In Table A14, panel A shows the coefficient estimates for the low sentiment buckets, and panel B shows the results for the the high sentiment buckets.

Over a one-day horizon, we observe a highly significant coefficient of 0.976 for cumulative abnormal returns in the low sentiment, high SI, low RI bucket, which is the bucket with the most binding short-sale constraint, as well as bad news. This is consistent with the prediction that stocks with tight short-sale constraints underreact to bad news, but not to good news. But we also see evidence of underreaction in the low sentiment, low SI, low RI grouping, and this effect is unlikely to be caused by short-sale constraints since these stocks are not heavily shorted.

Furthermore, the triple sort results are not robust to different return horizons. At the ten-day horizon, we observe the strongest $CAR_{1,10}$ response in the high sentiment, high SI, high RI bucket, with a coefficient of 4.149 (significant at the 10% level). And we observe a similar response in the low sentiment, high SI, high RI bucket, which is not short-sale constrained. For other groups where short sale constraints are most binding (those with high SI and low RI), we do not observe significant underreaction nor asymmetric responses across different news tone. These results argue against the second hypothesis.

A6.3 Serial correlation of news flow

Market participants may underreact to news because they do not fully understand the data generating process. In this section, we consider a particular aspect of the data generating process – the autocorrelation of news. Wang, Zhang, and Zhu (2018) document news sentiment momentum at a monthly frequency, and Huang, Tan, and Wermers (2020) show that news tone is highly persistent within-day and across consecutive days. Suppose market participants are unaware of this positive autocorrelation and simply assume independence in news tone. Investors will then respond to today's news unaware that tomorrow's news will likely have a similar tone. When tomorrow's news arrives it will surprise investors and cause a stock price reaction, even though this should have been forecastable using news from today. There will appear to be "underreaction" of prices to news, but this underreaction will operate entirely through the forecastability of tomorrow's news by today's news. We refer to this as the *news autocorrelation hypothesis.*

To test this hypothesis, we start from the following two panel regressions,

$$\begin{aligned} Y^i_{t,u,v} &= s_0 \times Sent^i_t + s_1 \times Sent^i_t \times \xi^i_t + s_2 \times \xi^i_t + \boldsymbol{\gamma}'\boldsymbol{X}^i_t + \varepsilon^i_{t,u,v}, \\ Sent^i_{t,u,v} &= \beta_0 \times Sent^i_t + \beta_1 \times Sent^i_t \times \xi^i_t + \beta_2 \times \xi^i_t + \boldsymbol{\delta}'\boldsymbol{X}^i_t + \eta^i_{t+1}, \end{aligned} \tag{A10}$$

where $Y^i_{t,u,v}$ is the excess or cumulative abnormal returns from day $t+u$ to $t+v$, $Sent^i_{t,u,v}$ is the average sentiment from day $t+u$ to $t+v$, ξ^i_t is either $Capacity_t$ or $Ownership^i_t$, and $\boldsymbol{X}_t$ is a vector of controls. With this notation, $Sent^i_t$ is the same as $Sent^i_{t,0,0}$. The β_0 coefficient in (A10) is roughly 0.25 for one-day ahead sentiment and 0.18 for ten-day ahead sentiment, and is highly significant in both cases (see Table A15). We therefore would like to understand whether this predictability in news sentiment is responsible for the predictability in price underreaction.

If the predictability of $Y^i_{t,u,v}$ results from the predictability of $Sent^i_{t,u,v}$, then s_0 should be a multiple of β_0, and s_1 should be the same multiple of β_1. To see this, consider a return process where

$$Y^i_{t,u,v} = a + b \times Sent^i_{t,u,v} + e^i_{t,u,v}, \tag{A11}$$

where the noise term is independent of $Sent^i_{t,u,v}$ and all time t information. Then the top equation in (A10) would follow from the sentiment process in the bottom equation.[18] Given (A11), news autocorrelation fully explains stock price underreaction, and thus the ratios of estimated coefficients $\hat{s}_0/\hat{s}_1$ and $\hat{\beta}_0/\hat{\beta}_1$ should be close to each other. In fact, this test has power against more general specifications than shown in (A11).

With $\hat{\boldsymbol{\theta}} = (\hat{s}_0, \hat{s}_1, \hat{\beta}_0, \hat{\beta}_1)$, we thus arrive at the test statistic

$$g(\hat{\boldsymbol{\theta}}) \equiv \frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1}.$$

Under the news autocorrelation hypothesis, $g(\hat{\boldsymbol{\theta}}) \xrightarrow{p} 0$. If the ratios $\hat{s}_0/\hat{s}_1$ and $\hat{\beta}_0/\hat{\beta}_1$ are far from each other, then we can reject this hypothesis, and thus conclude that news autocorrelation does not fully explain the stock price underreaction.

Panel A of Table A16 shows the test statistics and the simulated p-values for the above regressions when ξ^i_t equals intermediary capacity. The details of the simulation are discussed in Section A6.3.2. Across different combinations of the CAR variables and the intermediary capacity measures, the test statistics are always significant at the 5% level. Hence, we can reject the null hypothesis that news autocorrelation is the only channel through which stock prices underreact to news. In Panel B of Table A16, we set ξ^i_t equal

[18] See Section A6.3.1 for a precise derivation.

to our mutual fund ownership variables from Section 5.2 and redo the analysis. We reject the null for ten-day cumulative abnormal returns for all three ownership variables, and we reject the null hypothesis for one-day returns for the $Passive/Market$ ownership measure, and for $Active/Market$ measure for CAR (the one CAR non-rejection has a p-value of 0.1077).

These results indicate that news autocorrelation and market participants' lack of awareness of this correlation do not fully explain the dependence of stock price underreaction on our intermediary and ownership interaction variables.

A6.3.1 Deriving the test statistic

We provide a formal derivation of the test statistic in Section A6.3. We start from a generic setting and derive a general argument, then apply the results to our setting.

Consider the following generic data generating process:

$$Y = \theta W + \xi \tag{A12}$$

$$W = \boldsymbol{\beta}' \boldsymbol{Z} + \eta \tag{A13}$$

Equations (A12) and (A13) imply that

$$Y = \boldsymbol{s}' \boldsymbol{Z} + \varepsilon, \boldsymbol{s} = \theta \boldsymbol{\beta}, \varepsilon = \theta \eta + \xi \tag{A14}$$

In what follows, assume that $\mathbb{E}\left[\eta | \boldsymbol{Z}\right] = 0$, $\theta \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^k, k \geq 2$.

We want to test the null hypothesis $H_0 : \mathbb{E}\left[\xi | \boldsymbol{Z}\right] = 0$. The null hypothesis says that $\boldsymbol{Z}$ affects Y only through W, there are no other channels through which $\boldsymbol{Z}$ could affect Y. Under H_0 and given the assumption $\mathbb{E}\left[\eta | \boldsymbol{Z}\right] = 0$, we have $\mathbb{E}\left[\varepsilon | \boldsymbol{Z}\right] = 0$. Let $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{\beta}}$ denote the consistent estimates of $\boldsymbol{s}$ and $\boldsymbol{\beta}$ from OLS regressions (A14) and (A13), respectively. Then

$$\begin{aligned} \hat{\boldsymbol{s}} &\xrightarrow{p} \boldsymbol{s} = \operatorname{Var}\left(\boldsymbol{Z}\right)^{-1} \operatorname{Cov}\left(\boldsymbol{Z}, Y\right) = \theta \boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}} &\xrightarrow{p} \boldsymbol{\beta} = \operatorname{Var}\left(\boldsymbol{Z}\right)^{-1} \operatorname{Cov}\left(\boldsymbol{Z}, W\right) \end{aligned}$$

which implies

$$\frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1} \xrightarrow{p} \frac{s_0}{s_1} - \frac{\beta_0}{\beta_1} = 0$$

Hence,

$$H_0 : \mathbb{E}\left[\xi|\boldsymbol{Z}\right] = 0 \implies \frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1} \xrightarrow{p} 0$$

If we find that $\frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1}$ is far from 0, then we reject the $H_0 : \mathbb{E}\left[\xi|\boldsymbol{Z}\right] = 0$, and we conclude that there are other channels through which $\boldsymbol{Z}$ affects Y.

Now we map this generic setting to our paper. $Y = Y^i_{t,u,v}$ is the Retrf or CAR variable over horizon $[t+u, t+v]$ for stock i. $W = Sent^i_{t+1}$. $\boldsymbol{Z} = \left(Sent^i_t, Sent^i_t \times Capacity_t, Capacity_t, (\boldsymbol{X}^i_t)'\right)'$. If we find that $\frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1}$ is far from 0, we conclude that the news autocorrelation channel does not fully explain the stock price underreaction.

So we can run the following two regressions to obtain consistent estimates of $\boldsymbol{s}$ and $\boldsymbol{\beta}$. These two regressions correspond to equation (A14) and equation (A13), respectively.

$$Y^i_{t,u,v} = s_0 \times Sent^i_t + s_1 \times Sent^i_t \times Capacity_t + s_2 \times Capacity_t + \boldsymbol{\gamma}'\boldsymbol{X}^i_t + \varepsilon^i_{t,u,v} \quad \text{(A15)}$$

$$Sent^i_{t,u,v} = \beta_0 \times Sent^i_t + \beta_1 \times Sent^i_t \times Capacity_t + \beta_2 \times Capacity_t + \boldsymbol{\delta}'\boldsymbol{X}^i_t + \eta^i_{t+1} \quad \text{(A16)}$$

Let $\boldsymbol{\theta} = (s_0, s_1, \beta_0, \beta_1)'$, $\hat{\boldsymbol{\theta}} = \left(\hat{s}_0, \hat{s}_1, \hat{\beta}_0, \hat{\beta}_1\right)'$. Define $g(\boldsymbol{\theta}) = \frac{s_0}{s_1} - \frac{\beta_0}{\beta_1}$. Then the test statistic is $g\left(\hat{\boldsymbol{\theta}}\right)$. The null hypothesis is $H_0 : g\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{p} 0$.

To get a sense of the persistence in the $Sent^i_{t,u,v}$ variable, Table A15 shows the β_0 estimates from the regression in (A16).

A6.3.2 Deriving the p-value

Instead of using the Delta method to get the p-values for the test statistics,we propose the following simulation method.[19]

1. For each year y, run regressions (A15) and (A16), keep the coefficient estimates $\left(\hat{s}_{0,y}, \hat{s}_{1,y}, \hat{\beta}_{0,y}, \hat{\beta}_{1,y}\right)$.

2. Compute the Pearson correlation matrix for $\left(\hat{s}_{0,y}, \hat{s}_{1,y}, \hat{\beta}_{0,y}, \hat{\beta}_{1,y}\right)$ using the annual coefficient estimates from Step 1.

3. Run the full panel regressions (A15) and (A16), keep the coefficient estimates $\left(\hat{s}_0, \hat{s}_1, \hat{\beta}_0, \hat{\beta}_1\right)$. Also keep the estimated covariance matrix of the coefficients. Let $\hat{\boldsymbol{C}}_1$

[19] The Delta method does not work well because the test statistic is far from 0. See Table A16.

denote the estimated covariance of $(\hat{s}_0, \hat{s}_1)$, and $\hat{\boldsymbol{C}}_2$ denote the estimated covariance of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$.

4. Compute the covariance between $(\hat{s}_0, \hat{s}_1)$ and $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, using the estimated correlation matrix from Step 2 and the standard errors of $\left(\hat{s}_0, \hat{s}_1, \hat{\beta}_0, \hat{\beta}_1\right)$ from Step 3. Let $\hat{\boldsymbol{C}}_3$ denote that covariance matrix.

5. Draw $J = 1,000,000$ observations from a multivariate normal distribution with mean $\left(\hat{s}_0, \hat{s}_1, \hat{\beta}_0, \frac{\hat{s}_1}{\hat{s}_0}\hat{\beta}_0\right)$ and covariance matrix $\begin{pmatrix} \hat{\boldsymbol{C}}_1 & \hat{\boldsymbol{C}}_3 \\ \hat{\boldsymbol{C}}_3' & \hat{\boldsymbol{C}}_2 \end{pmatrix}$. Let $\left(\hat{s}_{0,j}, \hat{s}_{1,j}, \hat{\beta}_{0,j}, \hat{\beta}_{1,j}\right)$ denote the j-th draw.

6. Compute the p-value of the test statistic as the fraction of draws that satisfy $\left|g_j^{sim} - \bar{g}^{sim}\right| > |\hat{g} - \bar{g}^{sim}|$, where $g_j^{sim} = \frac{\hat{s}_{0,j}}{\hat{s}_{1,j}} - \frac{\hat{\beta}_{0,j}}{\hat{\beta}_{1,j}}$, $\bar{g}^{sim} = \frac{1}{J}\sum_{j=1}^{J} g_j^{sim}$, $\hat{g} = \frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1}$.

A6.4 Return response to news by subperiod

We partition the data into three five-year subperiods starting in 1996, one four-year subperiod at the end of our sample, as well as two subperiods which were classified as NBER recessions, in light of Garcia's (2013) finding of a changing news-returns relationship over the business cycle. The subperiods were selected by first identifying NBER recessions, and then splitting the remaining data into equal-sized windows. We chose subperiods prior to running any regressions and did not change them subsequently. Table A17 shows the results of the regression in (2) over the full sample with $u = v = 1$, as well as over the different subperiods.

In Table A17, we see that the news-returns relationship was stronger in the earlier parts of the sample, with sentiment coefficients of 1.595 (1996–2000), 1.255 (2001), and 0.861 (2002–2006).[20] The predictability of returns by sentiment rises slightly during the financial crisis period of 2007–2009 to 0.963 (significant at the 10% level), then drops sharply to 0.244 in the post-crisis years 2010–2014, and returns to 0.733 (significant at the 1% level) in the most recent time period of 2015–2018.[21] The magnitude of the underreaction in the most recent time period is similar to the full-sample coefficient of

[20] Of these, only 1.255 is not significant because it represents only the 2001 recession year, and is therefore associated with a high standard error.

[21] Our finding that single-name predictability did not sharply increase in the financial crisis contrasts with the finding in Garcia (2013) that news predictability for index returns is most pronounced during recessions.

0.884.[22]

Table A18 shows the results of the specification in (2) run with the original TSM control variables augmented with our two volatility controls. Here we use share turnover instead of illiquidity as we do in our main specification. Share turnover is defined as trading volume divided by the number of shares outstanding. Share turnover on day t is the average share turnover in the $[t-84, t-21]$ trading day window. The inclusion of IO, SI and log illiquidity as control variables in Table A17 slightly diminishes the role of *Sent* in most subperiods. Our full-sample results in Table A18 are even closer to TSM.

Table A19 shows the results for $Retrf$ and CAR of the ten-day ahead returns regressions. Table A20 is a summary of regression (2) for one- and ten-day ahead returns and of regression (3) for full-day and 4pm-9:30am sentiment. All four regressions are run over the full sample and over subperiods. The top panel shows results for $Retrf$ and the bottom panel shows results for CAR. A brief summary of the results: there is evidence of forecastability at the ten-day ahead horizon; the contemporaneous reactions of prices to news are much higher than the reaction of prices to lagged news, as has been documented in the prior literature (TSM, Heston and Sinha 2017 and Ke, Kelly, and Xiu 2021); the results of the contemporaneous 4pm-9:30am news regressions are very similar to the results of the full-day news regressions; there is no negative relationship over the subperiods between *Sent* coefficients in the lagged news regressions in (2) and the contemporaneous news regression in (3).

A6.5 Trimming ownership variables

To mitigate the concern that the ownership variable outliers discussed in Section A2.3 drive our ownership interaction results, we rerun the ownership interaction regressions in Table 5 but using the 2.5% trimmed ownership series, and confirm that the results are qualitatively unchanged, as can be seen in Table A21.

[22] Murray, Xiao and Xia (2023) examine the degree to which a recurrent neural network can forecast stock returns using lagged returns. They examine the performance of their strategy in subperiods (e.g., 1995-2004, 2005-2014, 2015-2019) that are similar to ours. The profitability of their strategy is high in 1995-2004 and 2015-2019, and low in the middle period 2005-2014. And the profitability in the most recent period is not as high as in the initial period. The time variation in their forecastability results is very close to our findings in Table A17 suggesting that the phenomena we examine may impact a broad class of return patterns.

A6.6 The role of earnings announcements

TSM show that the sentiment of articles containing any word beginning with "earn" is more strongly associated with contemporaneous returns and is a stronger predictor of future returns than sentiment of articles that do not contain earnings-related words (though the latter remains statistically and economically important). To ensure that our results are not driven by articles about earnings, we run a version of the specifications in (2) and (3) that drops all event days that take place either on earnings announcement days, or on the trading day following the earnings announcement.[23] Dropping these two-day announcement periods reduces the number of observations in our full-sample regression by roughly 10%.

Table A22 is the analogue of Table A20 but after the two-day announcement windows are dropped. Table A20 shows the Table 1 results for the full sample, as well as results by subperiods as explained in Section A6.4. The magnitudes of the contemporaneous coefficients drop in the absence of earnings-related news, but remain economically and statistically important. The lagged sentiment coefficients for one- and ten-day ahead returns are roughly comparable. The impact of news on future returns hardly changes when earnings days are excluded from the sample and, thus, earnings-related news are not the main drivers of our results.

A6.7 Controlling for volatility

Ang et al. (2006) showed that stocks with high idiosyncratic volatility earn "abysmally" low excess returns. It is possible therefore that the reason negative sentiment forecasts low returns is because it is associated with high idiosyncratic or systematic volatility. We include $CAR_{0,0}^2$ (idiosyncratic variance) and VIX (systematic volatility) in all our regressions to control for this possibility. Table A23 shows the results of the specification in (2) when we remove the volatility controls. Compared to the baseline results in Table A20, the forecasting ability of $Sent$ for both excess returns and CARs is basically unchanged in the absence of the volatility controls.

A6.8 The VIX as an interaction variable

To control for the possibility that intermediary capacity, ownership, and entropy simply proxy for the impact of investor perceptions of risk on the sentiment coefficient in (13,

[23] We drop both days because we are uncertain whether the earnings announcement takes place before or after the market close on the announcement day.

14, 15), we run a specification analogous to these but use the VIX as the interaction variable for sentiment. Table A24 shows these results. In all cases, the VIX has a positive influence on the impact of sentiment on contemporaneous returns, and a negative (and usually insignificant) influence on the impact of sentiment on future returns. These results are fundamentally different from the intermediary capacity, ownership, and entropy interaction results in Tables 4, 5, and 6, where the impact of the interaction variable on the sentiment coefficient has the same sign for contemporaneous and future returns. The VIX, therefore, cannot be the underlying driver of our results.

References

Abarbanell, J. and V. Bernard, 1992, "Tests of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behavior," *Journal of Finance*, 47 (3), 1181–1207.

Asquith, P., Pathak, P.A. and Ritter, J.R., 2005,"Short interest, institutional ownership, and stock returns," *Journal of Financial Economics*, 78(2), 243–276.

Das, S. and M. Chen, 2007, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," *Management Science*, 53 (9), 1375–1388.

Easterwood, J. and S. Nutt, 1999, "Inefficiency in analysts' earnings forecasts: Systematic misreaction or systematic optimism?" *Journal of Finance*, 54 (5), 1777–1797.

Fama, E. and K. French, 2015, "A five-factor asset pricing model," *Journal of Financial Economics*, 16, 1–22.

Hoechle, D., 2007, "Robust standard errors for panel regressions with cross-sectional dependence," *The Stata Journal*, 7 (3), 281–312.

Hong, H. and J. Stein, 1999, "A unified theory of underreaction, momentum trading, and overreaction in asset markets," *Journal of Finance*, 54, 2143–84.

Huang, A., H. Tan, and R. Wermers, 2020, "Institutional trading around corporate news: Evidence from textual analysis," *Review of Financial Studies*, 33 (10), 4627–4675.

Jorda, O., 2005, "Estimation and inference of impulse responses by local projection," *American Economic Review*, 95 (1), 161–182.

Ke, Z., B. Kelly, and D. Xiu, 2021, "Predicting returns with text data," working paper.

Miller, E., 1977, "Risk, uncertainty, and divergence of opinion," *Journal of Finance*, 32 (4), 1151–1168.

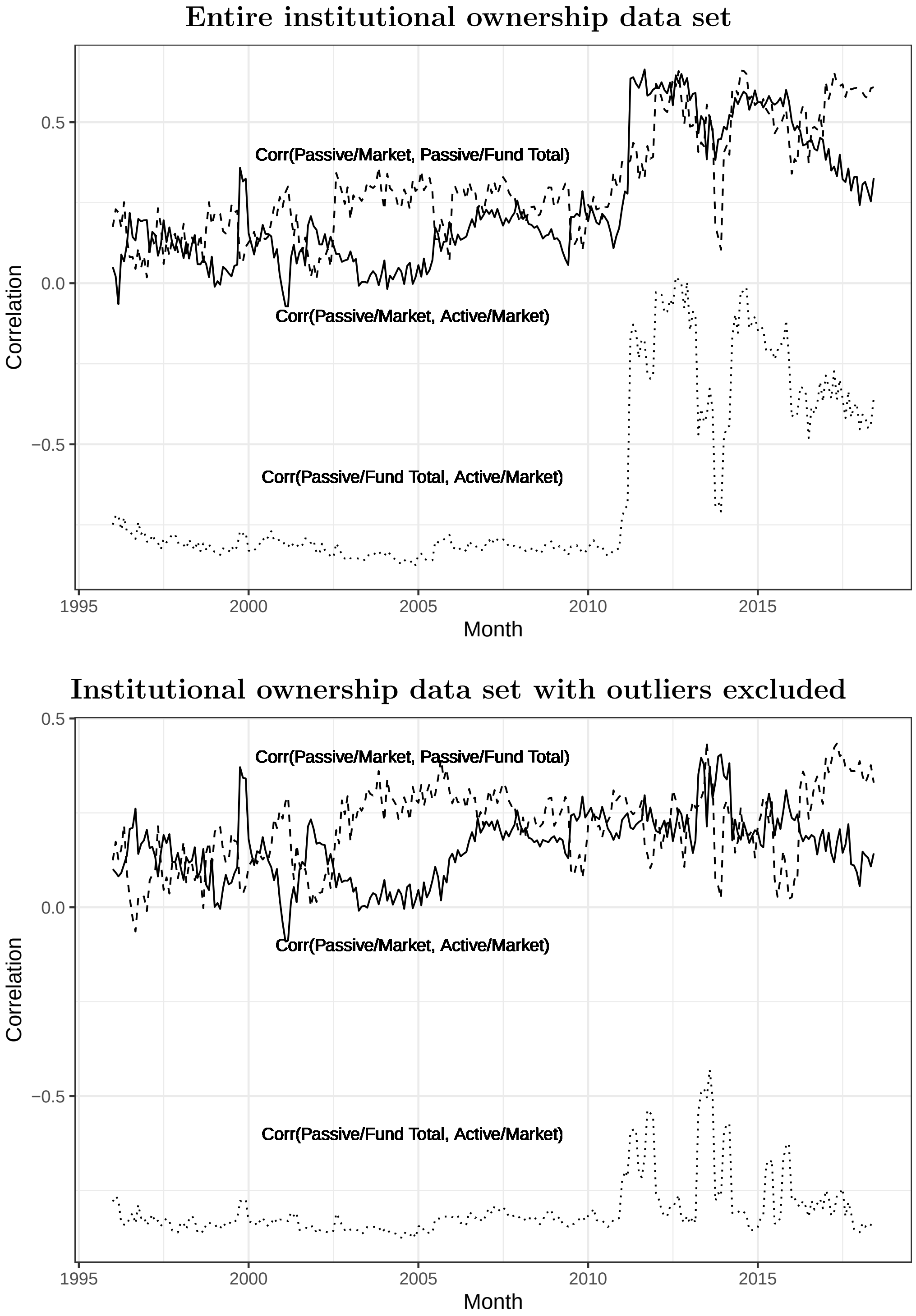
Murray, Xiao and Xia, 2023, “Charting by machines,” working paper.

Nagel, S., 2005, “Short sales, institutional investors and the cross-section of stock returns,” *Journal of Financial Economics* 78, 277–309.

Shin, H.S., 2009, *Risk and Liquidity*, Oxford University Press.

Sims, C., 2011, “Rational inattention and monetary economics,” Chapter 4 in *Handbook of Monetary Economics*, Vol. 3A, Elsevier.

Wang, Y., B. Zhang, and X. Zhu, 2018, “The momentum of news,” working paper.



Entire institutional ownership data set

Institutional ownership data set with outliers excluded

Fig. A1. Time series of cross-sectional ownership correlations. Within each month, this chart shows the cross-sectional correlations of our three ownership measures. The top panel shows the results for the full data set. The bottom panel shows the results when excluding the bottom 2.5% of each series within each month.

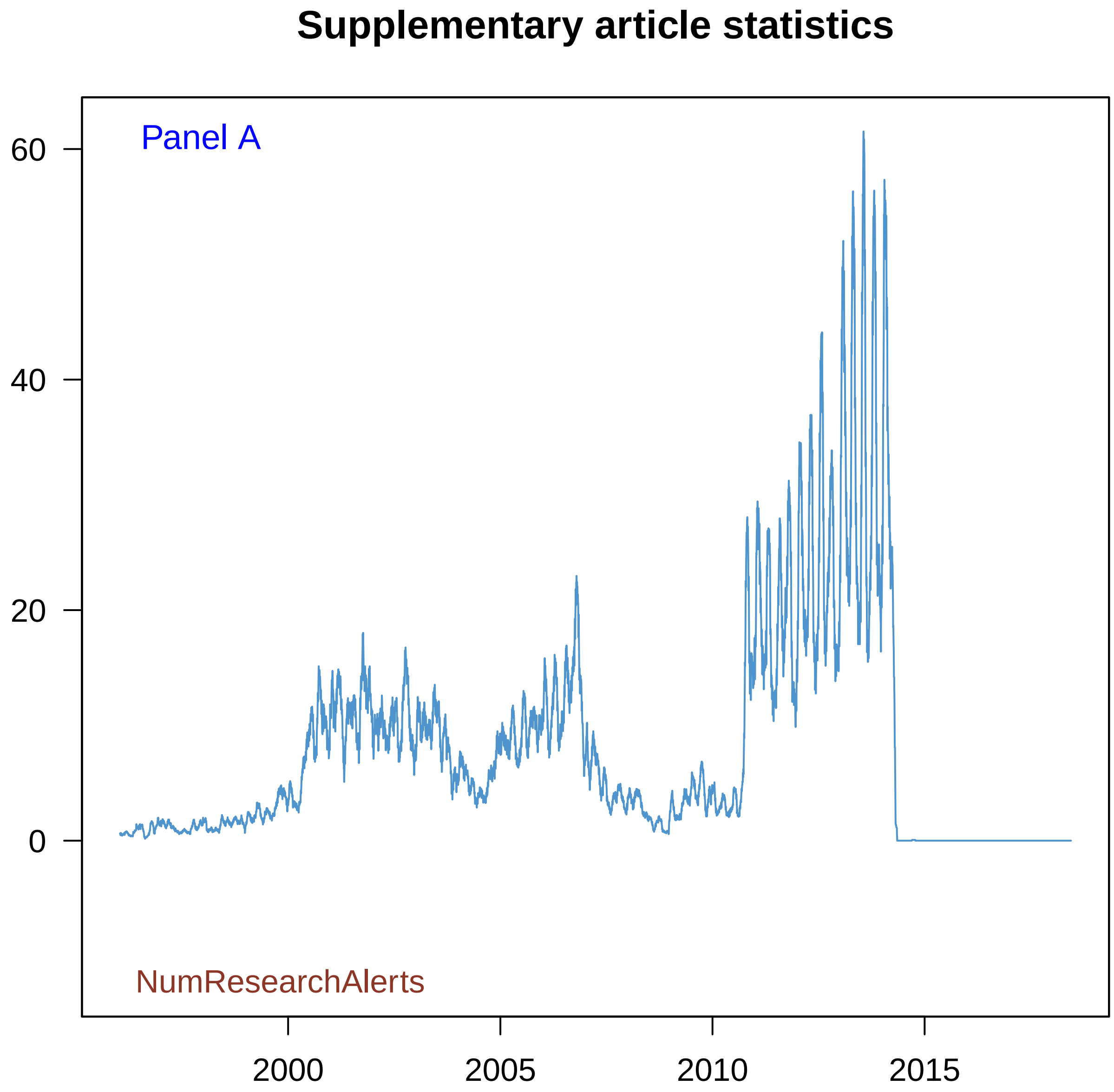


Fig. A2. This chart shows the daily number of articles with headlines containing “RESEARCH ALERT-” (case insensitive match).

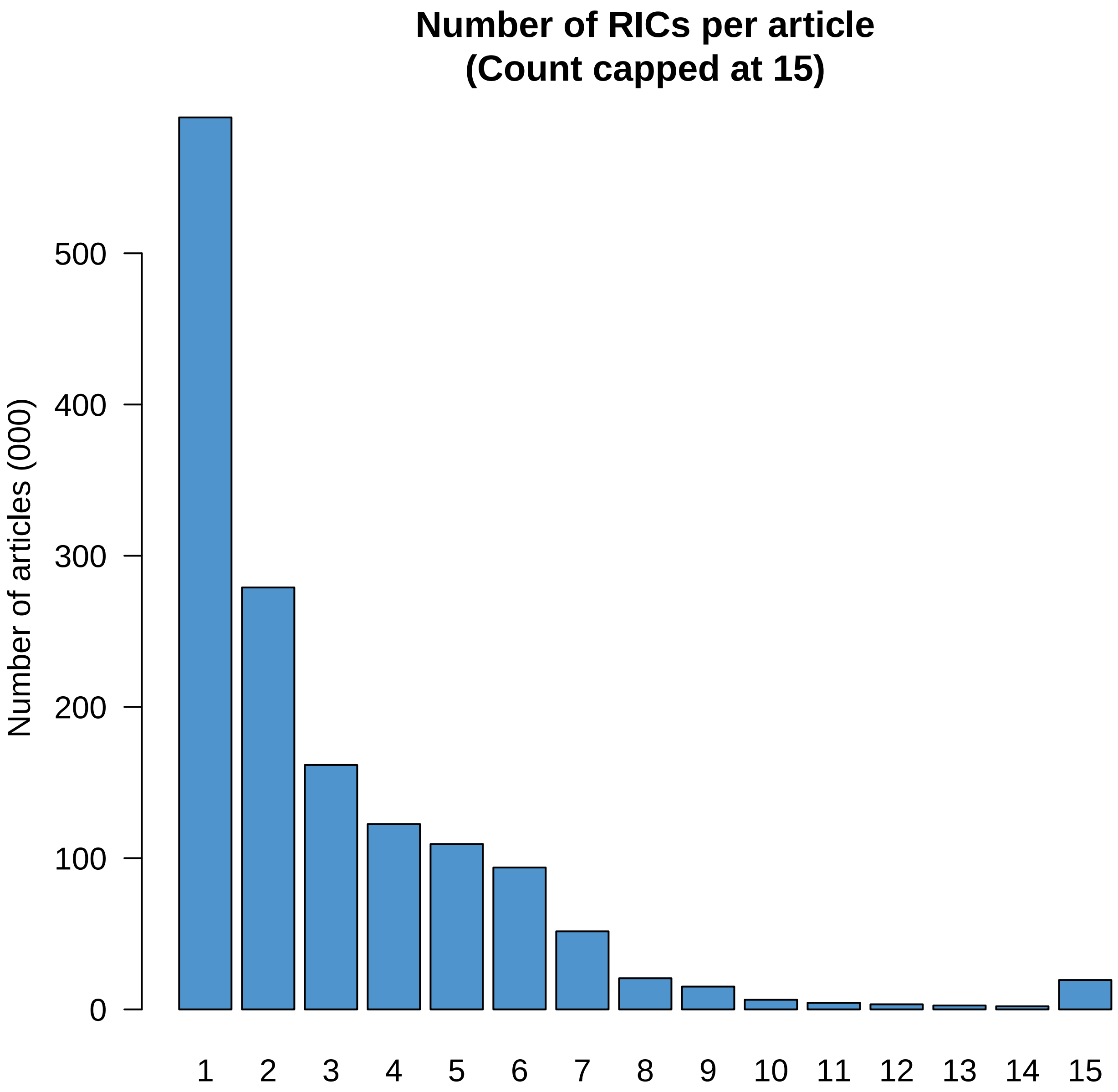
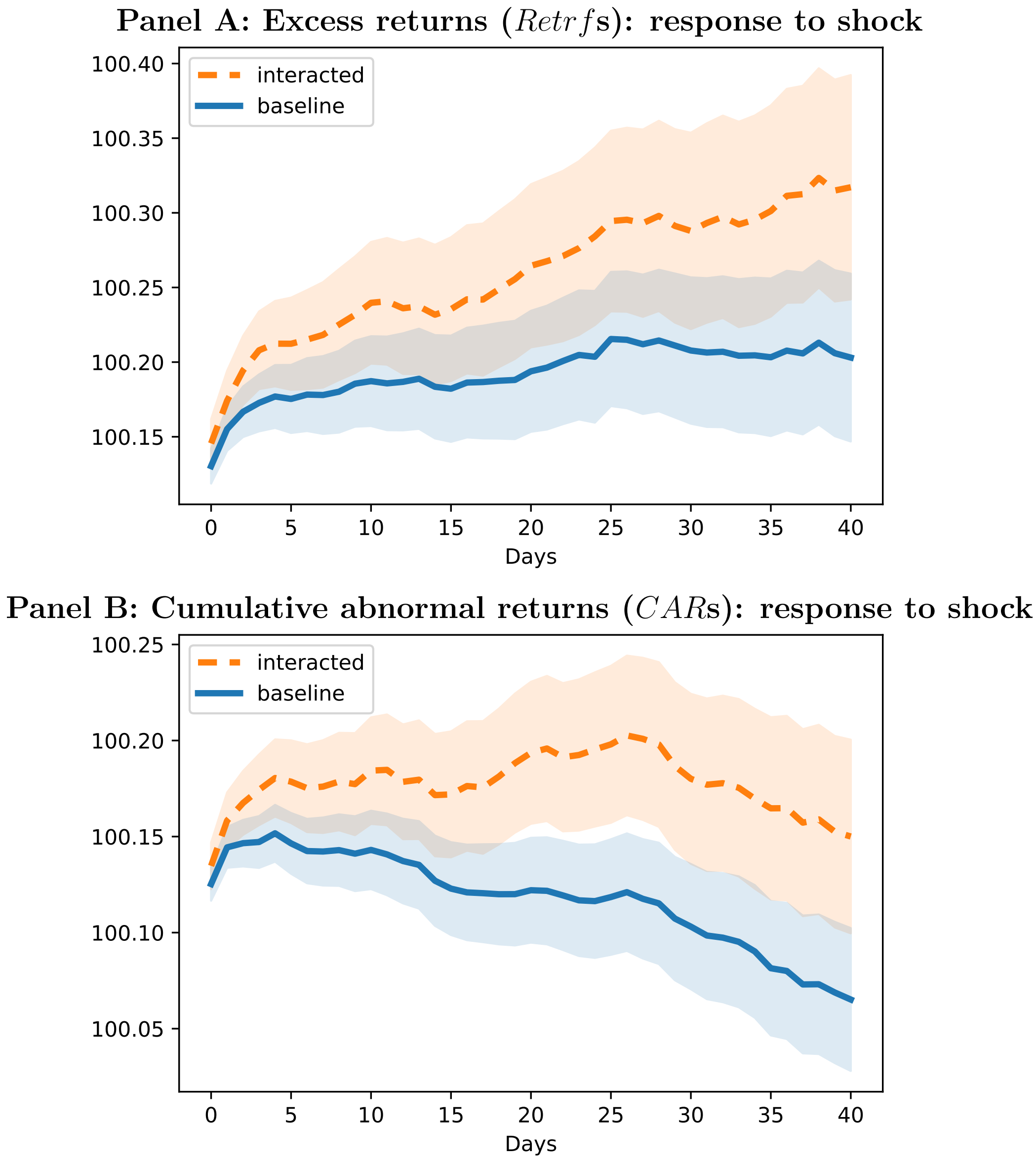


Fig. A3. This figure shows the histogram of the number of RICs (Reuters company identifier) per article. The y-axis is labeled with the number of articles in each RICs bucket, in thousands.

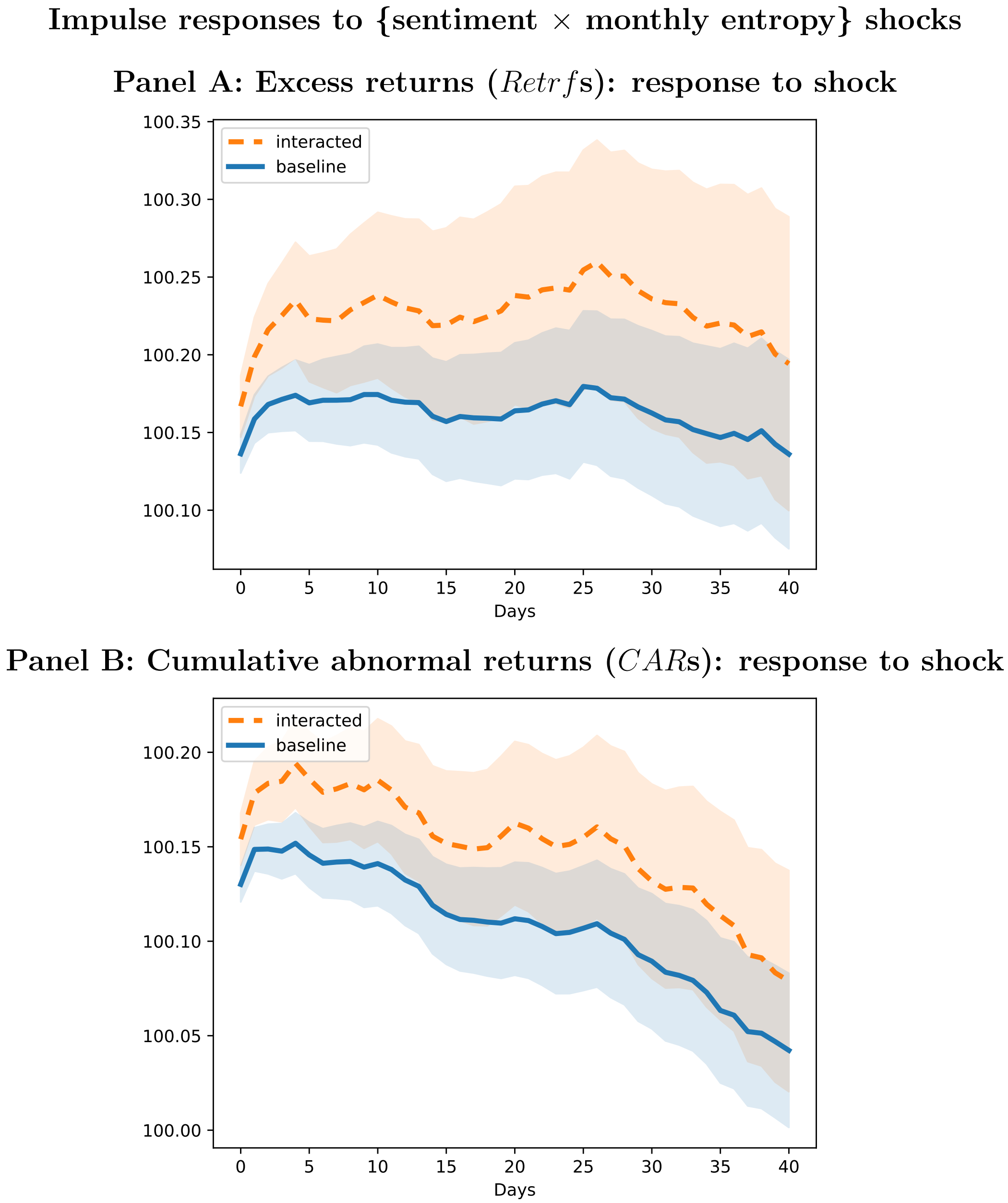
Impulse responses to {sentiment × passive/total ownership} shocks



Panel A: Excess returns ($Retrfs$): response to shock

Panel B: Cumulative abnormal returns ($CARs$): response to shock

Fig. A4. Impulse response functions estimated using the local projection method of Jorda (2005). The figure shows the baseline response (labeled *baseline*) of future excess returns and cumulative abnormal returns ($CARs$) to a one standard deviation sentiment shock, as well as the response conditional on a one-standard deviation decrease in passive/total ownership (labeled *interacted*). The starting price level on day -1 is 100. Day 0 is the news event day. The x-axis is in number of days. The top panel shows cumulative excess returns, and the bottom panel shows $CARs$. The cumulative responses show the arithmetic sums of one-day returns; the geometric cumulative returns are almost identical. Standard errors are based off time-clustered panel regressions of one-day ahead future returns on lagged sentiment, and assume independence of one-day returns across time, and between the baseline and the conditinal responses. The shaded regions represent 2 standard error bands around the impulse response.



Impulse responses to {sentiment × monthly entropy} shocks

Panel A: Excess returns ($Retrfs$): response to shock

Panel B: Cumulative abnormal returns (CARs): response to shock

Fig. A5. Impulse response functions estimated using the local projection method of Jorda (2005). The figure shows the baseline response (labeled *baseline*) of future excess returns and cumulative abnormal returns (CARs) to a one standard deviation sentiment shock, as well as the response conditional on a one-standard deviation increase in monthly entropy (labeled *interacted*). The starting price level on day -1 is 100. Day 0 is the news event day. The x-axis is in number of days. The top panel shows cumulative excess returns, and the bottom panel shows CARs. The cumulative responses show the arithmetic sums of one-day returns; the geometric cumulative returns are almost identical. Standard errors are based off time-clustered panel regressions of one-day ahead future returns on lagged sentiment, and assume independence of one-day returns across time, and between the baseline and the conditinal responses. The shaded regions represent 2 standard error bands around the impulse response.

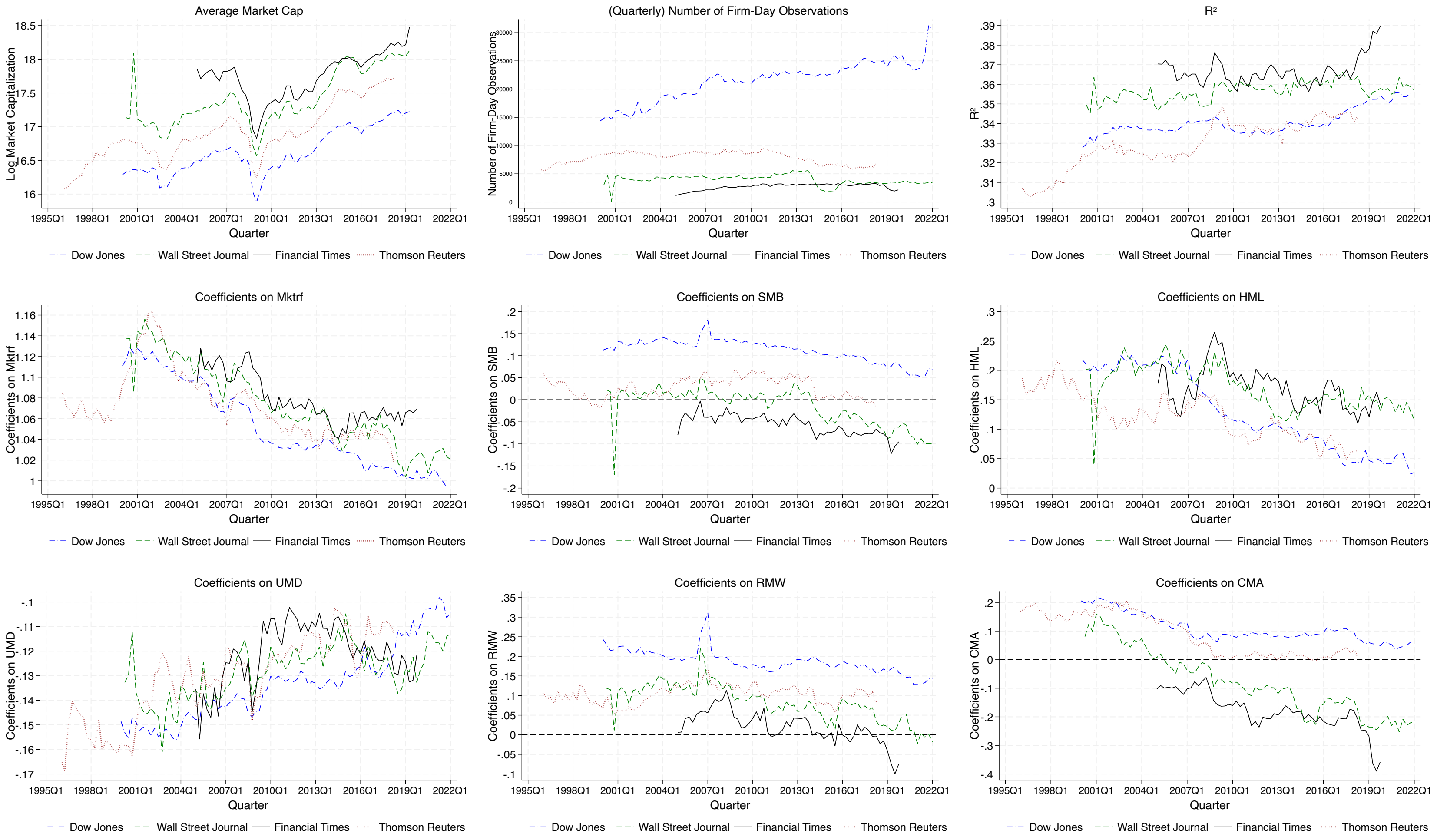
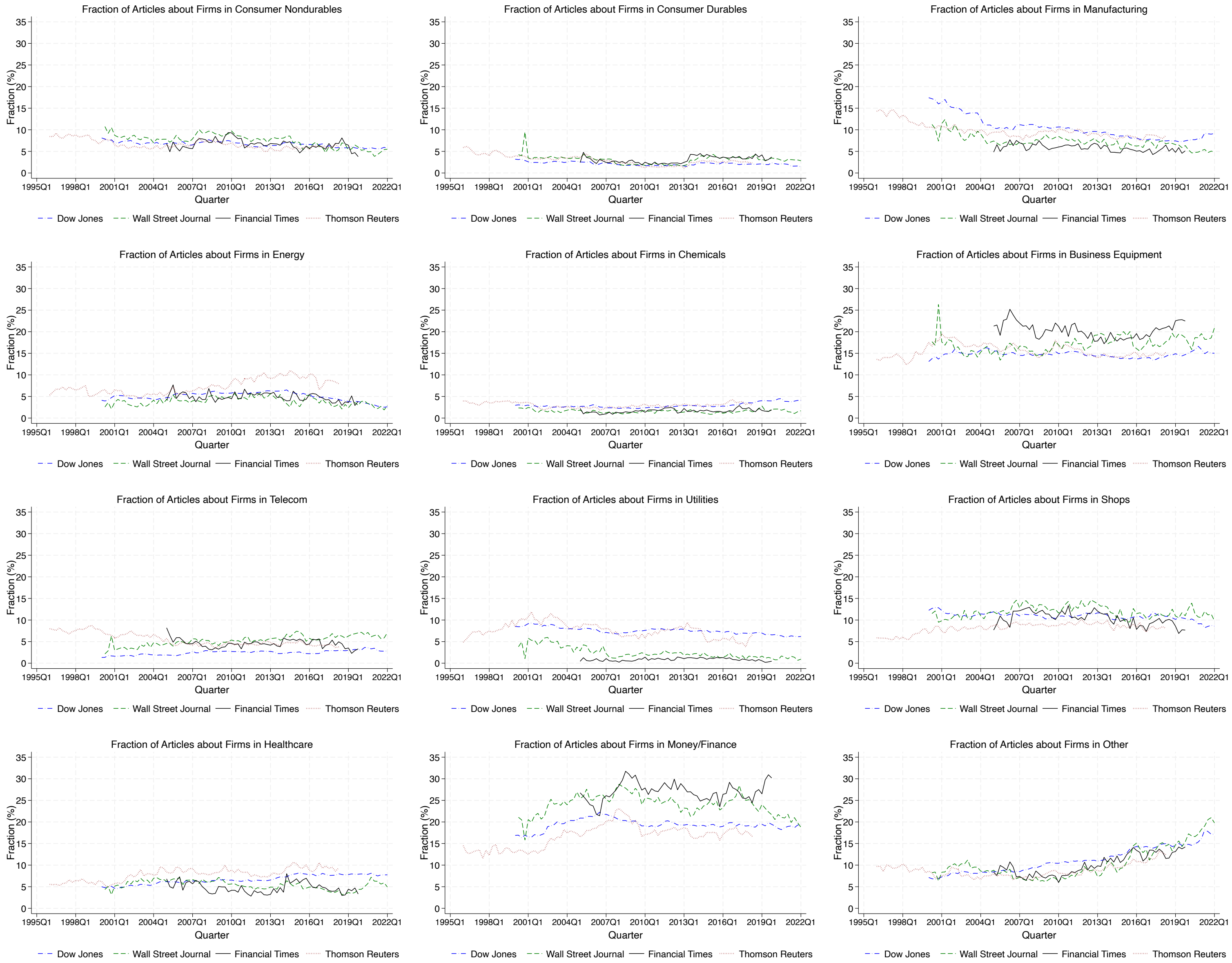


Fig. A6. Analysis of composition effects for Dow Jones, Wall Street Journal, Financial Times, and Reuters data sources. The methodology to generate these graphs is explained in Section A5.1.



Fraction of Articles about Firms in Consumer Nondurables

Fraction of Articles about Firms in Consumer Durables

Fraction of Articles about Firms in Manufacturing

Fraction of Articles about Firms in Energy

Fraction of Articles about Firms in Chemicals

Fraction of Articles about Firms in Business Equipment

Fraction of Articles about Firms in Telecom

Fraction of Articles about Firms in Utilities

Fraction of Articles about Firms in Shops

Fraction of Articles about Firms in Healthcare

Fraction of Articles about Firms in Money/Finance

Fraction of Articles about Firms in Other

Fig. A7. Analysis of industry coverage for Dow Jones, Wall Street Journal, Financial Times, and Reuters data sources. The methodology to generate these graphs is explained in Section A5.1.

Quarterly cross-sectional standard deviation of SUE and SAFE

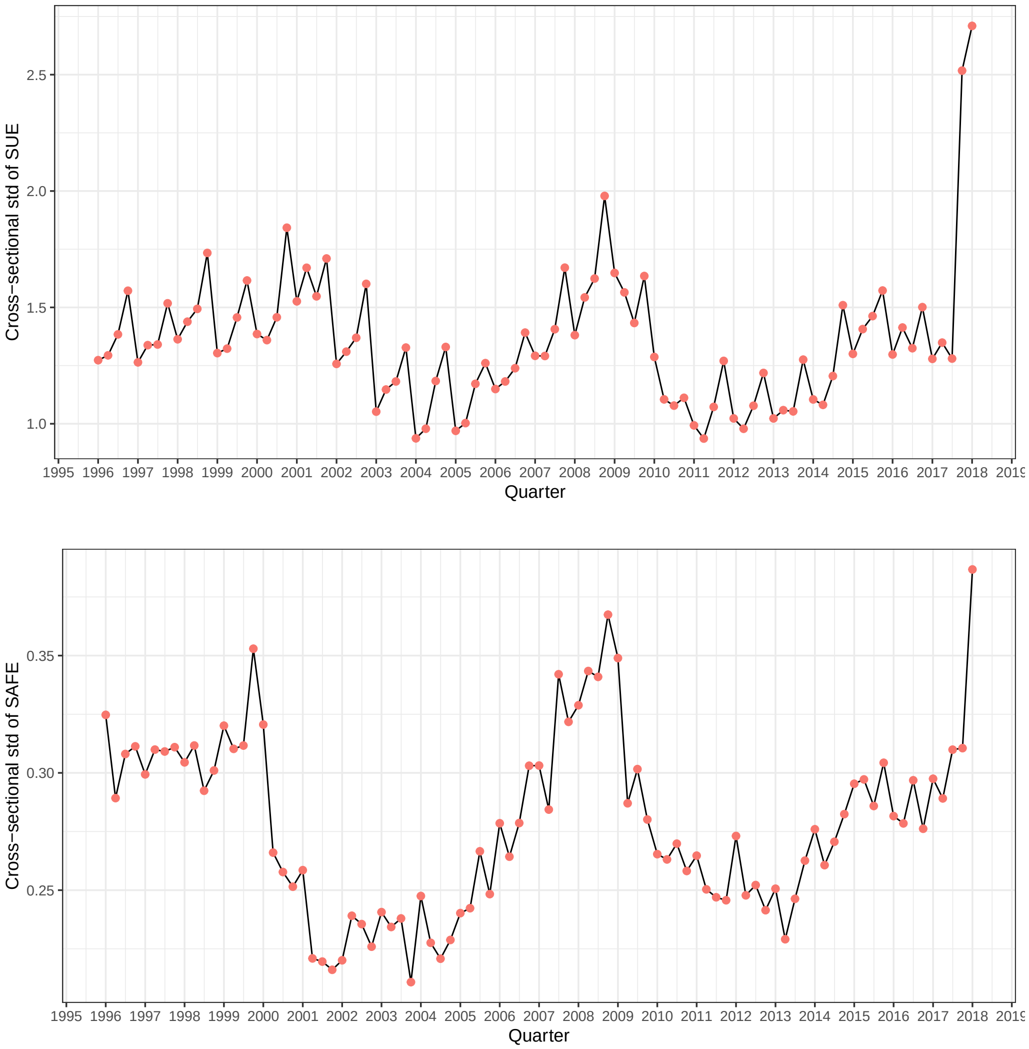
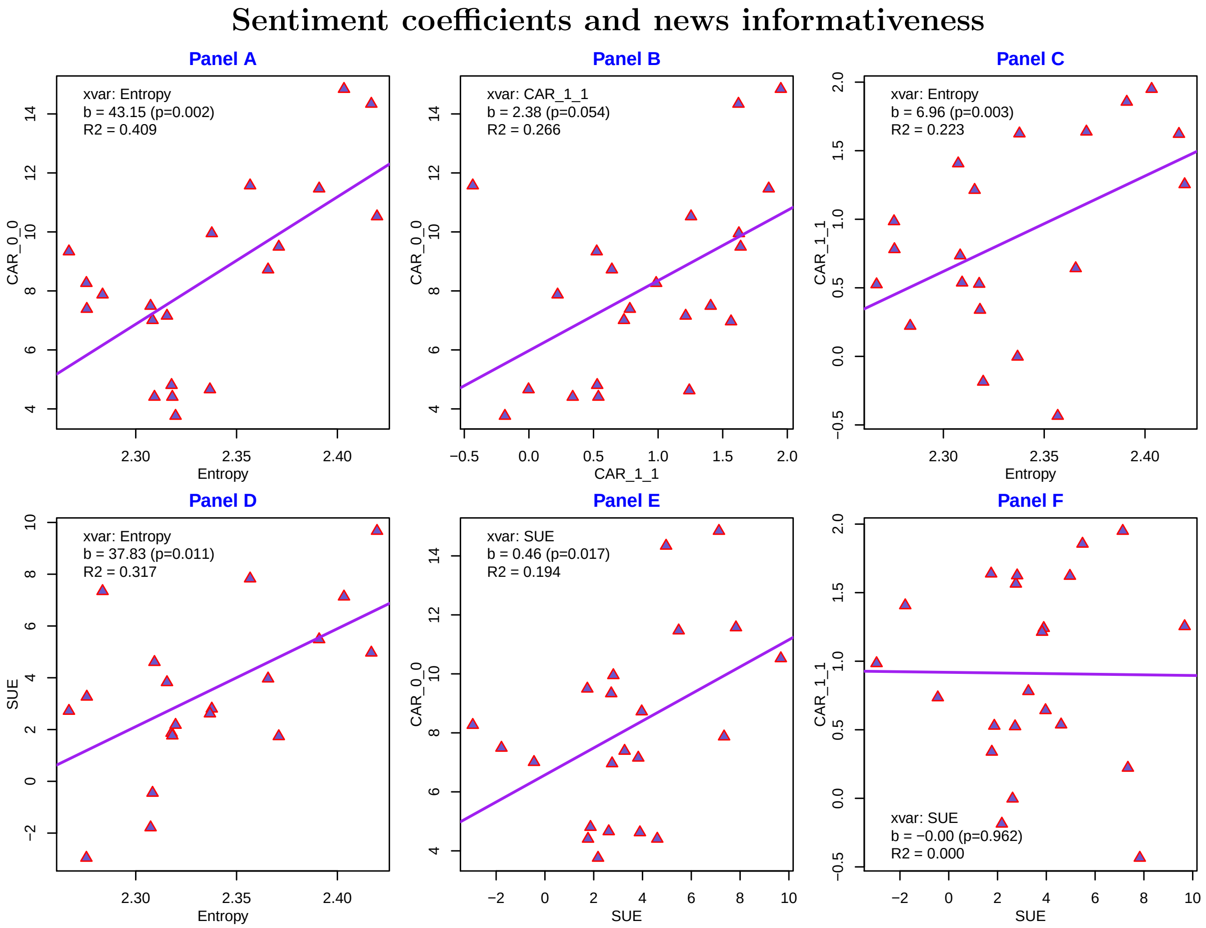


Fig. A8. The top panel shows the quarterly cross-sectional standard deviation of SUE. The bottom panel shows the quarterly cross-sectional standard deviation of $SAFE$.



Sentiment coefficients and news informativeness

Panel A

Panel B

Panel C

Panel D

Panel E

Panel F

Fig. A9. Panel A shows the sentiment coefficients from annual regressions of returns $CAR_{0,0}$ on contemporaneous sentiment (eq. 3) plotted against annual average entropy. Panel B shows the sentiment coefficients from the $CAR_{0,0}$ regression plotted against the sentiment coefficients from an annual regression of returns $CAR_{1,1}$ on one-day lagged sentiment (eq. 2). Panel C shows the $CAR_{1,1}$ sentiment coefficients plotted against annual average entropy. Panel D plots the sentiment coefficient from annual regressions of SUE on lagged monthly sentiment (eq. A9) against annual average entropy. Panel E plots the annual $CAR_{0,0}$ sentiment coefficients against the annual SUE sentiment coefficients. Panel F plots the annual $CAR_{1,1}$ sentiment coefficients against the annual SUE sentiment coefficients. Each point in the table corresponds to a single year of the sample. Each chart also shows the R^2 of the best fitting regression line (shown in purple) between the y- and x-variables, as well as the slope coefficient and p-value of the regression, with standard errors calculated using White's heteroscedasticity correction.

Table A1
Abbreviations.

Abbreviation	Replacement	Abbreviation	Replacement
SYS	SYSTEMS	UTILS	UTILITIES
MFG	MANUFACTURING	CHEM	CHEMICAL
WLDWD	WORLDWIDE	INTL	INTERNATIONAL
SVCS	SERVICES	INDS	INDUSTRIES
PPTY	PROPERTY	INVS	INVESTORS
RETRMENT	RETIREMENT	DEPT	DEPARTMENT
RLTY	REALTY	TR	TRUST
MGMT	MANAGEMENT	RES	RESOURCES
NETWRKS	NETWORKS	SOLS	SOLUTIONS
EXCH	EXCHANGE	HLDG	HOLDING
REST	RESORTS	MACHS	MACHINES
LTG	LIGHTING	LABS	LABORATORIES
RESH	RESEARCH	FRAG	FRAGRANCES
INFO	INFORMATION		

Table A2
Direct replacement.

Method	Words
Direct	INC, CORP, CO, GROUP, LTD, PLC, HOLDINGS, COMPANY, COMPANIES, COS, HLDGS, GRP, 2ND, COR, GP, LLC

Table A3
Recursive replacement: *Variant-1*.

Method	Words
Recursive	NEW, DEL, DE, NY, VA, WIS, GA, AG, MA, NC, NEV, NJ, OH, PA, TX, WA, NV, BRIDGEPORT, IND, AMER, LIMITED, KANSAS

Table A4
Recursive replacement: *Variant-2*.

Method	Words
Recursive	INTERNATIONAL, ENERGY, FINANCIAL, INDUSTRIES, L, SYSTEMS, RESOURCES, SERVICES, TECHNOLOGIES, TECHNOLOGY, INTL, POWER, ELECTRIC, HOLDING, SVCS, SERVICE, OF, INDS, UTILITIES, SYS, ENERGIES, UTILS, INSURANCE, LT, HLDG, RES

Table A5
Replaced patterns in augmented article body. X denotes a numerical character, ␣denotes a space character, and [␣]∗denotes zero or more space characters.

Pattern	Replacement	Pattern	Replacement
Panel A: year and month			
19XX 19XX.XX 19XX-XX	␣␣y␣␣	20XX 20XX.XX 20XX-XX	␣␣y␣␣
Panel B: entity names			
s&p	snp	s␣&␣p	snp
standard␣&␣poor’s	snp	standard␣and␣poor’s	snp
snp␣500	snp500	dow␣jones␣industrial␣average	djia
new␣york␣stock␣exchange	nyse	london␣stock␣exchange	ftse
stock␣exchange␣of␣hong␣kong␣	sehk	australian␣stock␣exchange␣	asx
fannie␣mae	fnma	freddie␣mac	fdmc
federal␣reserve	fed	securities␣and␣exchange␣commission	sec
chief␣executive␣officer␣	ceo	chief␣financial␣officer␣	cfo
chief␣operating␣officer␣	coo	chief␣investment␣officer␣	cio
vice␣president␣	vp	international␣monetary␣fund␣	imf
u.n.	un		
Panel C: numerical strings			
XXXXXXXXXX	␣␣bn␣␣	XXXXXXX	␣␣mn␣␣
X[␣]∗billion	␣␣bn␣␣	X[␣]∗million	␣␣mn␣␣
Panel D: punctuation marks between sentences			
? ! . : ;	∗∗∗		
Panel E: punctuation marks within sentences			
“”# $ % & ‘’( ) ∗+ − \<= >@ [ ] ˆ`{|}˜	␣		

Table A6
The table shows the years of coverage of each news archive, the number of firm-day observations from each news source for S&P 500 companies on the day the articles were written, as well as the number of firm-day observations that overlap with our Thomson Reuters data set. The article-PERMNO mapping procedure is explained in Section A1.2.

Number of firm-day observations across different news archives

	Years covered	Number firm-day (FD) observations	Number TR overlapping FD observations
Thomson Reuters	1996 – 2018	706,000	—
Dow Jones	2000 – 2022	1,909,660	435,887
Wall Street Journal	2000 – 2022	346,245	162,865
Financial Times	2005 – 2019	163,278	94,114

Table A7
Each row shows daily alphas, in basis points (bps), from the trading strategy explained in Section 5.4. The alphas are relative to the Fama and French (2015) five factor model with momentum. The columns correspond to different values of the *keep* variable in (A2). The columns without the TC label assume zero transaction costs; the ones with a TC label assume transaction costs equal 3 bps per unit of turnover (round-trip transaction). The rows correspond to different conditioning variables (none, intermediary capitalization, active ownership, and entropy, respectively) that impact the gross size of the long-short strategy via (18), with the *scale* variable set to 0.165. The numbers in parentheses represent p-values with standard errors calculated using Newey-West with lags equal to the floor of $4(N/100)^{2/9}$ where N is the number of observations in the sample (see Hoechle 2007).

News trading strategy six-factor alphas (bps per day) with $scale = 0.165$

Condition	Keep=1	Keep=0.33	Keep=1 TC	Keep=0.33 TC
None	7.678	3.287	2.399	1.666
	(0.000)	(0.000)	(0.037)	(0.011)
CR	9.202	4.331	3.537	2.593
	(0.000)	(0.000)	(0.006)	(0.000)
Pct Active	8.645	3.844	3.087	2.138
	(0.000)	(0.000)	(0.019)	(0.005)
Entropy	7.543	3.244	2.693	1.754
	(0.000)	(0.000)	(0.024)	(0.010)

Table A8

Each row shows daily alphas, in basis points (bps), from the trading strategy explained in Section 5.4. The alphas are relative to the Fama and French (2015) five factor model with momentum. The columns correspond to different values of the *keep* variable in (A2). The columns without the TC label assume zero transaction costs; the ones with a TC label assume transaction costs equal 3 bps per unit of turnover (round-trip transaction). The rows correspond to different conditioning variables (none, intermediary capitalization, active ownership, and entropy, respectively) that impact the gross size of the long-short strategy via (18), with the *scale* variable set to 0.66. The numbers in parentheses represent p-values with standard errors calculated using Newey-West with lags equal to the floor of $4(N/100)^{2/9}$ where N is the number of observations in the sample (see Hoechle 2007).

News trading strategy six-factor alphas (bps per day) with $scale = 0.66$

Condition	Keep=1	Keep=0.33	Keep=1 TC	Keep=0.33 TC
None	7.678	3.287	2.399	1.666
	(0.000)	(0.000)	(0.037)	(0.011)
CR	13.772	7.466	6.909	5.360
	(0.000)	(0.000)	(0.001)	(0.000)
Pct Active	11.545	5.517	5.134	3.547
	(0.000)	(0.000)	(0.008)	(0.002)
Entropy	7.136	3.115	2.833	1.789
	(0.000)	(0.000)	(0.060)	(0.043)

Table A9
Summary statistics for the earnings regressions. All statistics are calculated by pooling single-name data across all companies in our sample. This includes only the time periods during which these companies were members of the S&P 500 index.

Summary statistics for earnings regressions

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
SUE (5% Win)	40,000	−0.042	1.397	−4.320	−0.508	0.564	3.271
SAFE (5% Win)	36,812	0.097	0.281	−0.568	−0.007	0.177	0.991
Sent	35,839	−0.011	0.016	−0.250	−0.019	0.000	0.111
Forecast Dispersion (1% Win)	40,053	0.146	0.172	0.000	0.040	0.185	1.217
Forecast Revisions (1% Win)	40,289	−0.001	0.003	−0.028	−0.0003	0.000	0.007
$\mathrm{CAR}_{-2,-2}$	40,478	0.029	1.918	−37.216	−0.803	0.806	55.365
$\mathrm{CAR}_{-30,-3}$	40,477	−0.061	9.234	−82.174	−4.477	4.198	209.534
Short Interest (%)	38,817	3.188	3.575	0.000	1.193	3.776	77.120
Institutional Ownership (%, 1% Win)	40,320	71.423	19.075	0.962	61.612	84.583	111.719
log(Market Cap)	40,425	23.152	1.162	19.079	22.377	23.831	27.481
IHS(Book/Market) (1% Win)	38,274	0.448	0.303	−0.109	0.227	0.612	1.583
log(Illiquidity)	40,468	−22.466	1.387	−27.596	−23.361	−21.589	−13.853
α	40,467	0.015	0.116	−0.976	−0.046	0.069	1.222

Table A10
These regressions include as controls: lagged SUE or SAFE, analyst forecast dispersion, analyst forecast revisions, lagged abnormal returns $CAR_{-2,-2}$ and $CAR_{-30,-3}$, short interest, institutional ownership, log market capitalization, the IHS transform of book to market, log illiquidity, and the past year's alpha from our six factor model. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

SUE **and** $SAFE$ **forecastability by** $SENT$

	1996-2018	1996-2000	2001	2002-2006	2007-2009	2010-2014	2015-2018
SUE	3.733***	3.971***	9.67***	4.323***	3.182**	3.144***	0.015
SAFE	0.557***	0.369	1.399***	0.743***	0.967**	0.379*	0.557**

Table A11
These regressions include as controls: lagged SUE or SAFE, analyst forecast dispersion, analyst forecast revisions, lagged abnormal returns $CAR_{-2,-2}$ and $CAR_{-30,-3}$, short interest, institutional ownership, log market capitalization, the IHS transform of book to market, log illiquidity, and the past year's alpha from our six factor model. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

SUE and $SAFE$ forecastability by $SENT$ excluding 4Q2017 and 1Q2018

	1996-2017 Q3	1996-2000	2001	2002-2006	2007-2009	2010-2014	2015-2017 Q3
SUE	4.078***	3.971***	9.67***	4.323***	3.182**	3.144***	1.678
SAFE	0.552***	0.369	1.399***	0.743***	0.967**	0.379*	0.528**

Table A12

Forecasting regressions for SUE and SAFE. These regressions include as controls: lagged SUE or SAFE, analyst forecast dispersion, analyst forecast revisions, lagged abnormal returns $CAR_{-2,-2}$ and $CAR_{-30,-3}$, short interest, institutional ownership, log market capitalization, the IHS transform of book to market, log illiquidity, and the past year's alpha from our six factor model. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

SUE and SAFE forecasting regressions from 1996 to 2018

	Dependent variable:	
	SUE	SAFE
Constant	0.385	-0.169^{***}
Sent	3.733^{***}	0.557^{***}
Lag(SUE)	0.263^{***}	
Lag(SAFE)		0.213^{***}
Forecast Dispersion	-0.601^{***}	0.238^{***}
Forecast Revisions	57.751^{***}	7.491^{***}
$\text{CAR}_{-2,-2}$	0.012^{***}	0.002^{***}
$\text{CAR}_{-30,-3}$	0.005^{***}	0.002^{***}
Short Interest (%)	-0.010^{***}	-0.003^{***}
IO (%)	-0.001	0.0003^{***}
log(Market Cap)	-0.032	-0.018^{***}
IHS(Book/Market)	0.034	-0.051^{***}
log(Illiquidity)	-0.024	-0.029^{***}
α	1.394^{***}	0.034^{**}
Observations	31,581	29,733
Adjusted R^2	0.137	0.122
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table A13

We obtain the monthly median short interest (SI) and median residualized ownership (RI), and use these cutoffs to double sort stock-day observations by SI and RI (independently). For each bucket, we run the main specification in equation (2). The coefficient estimates $\hat{s}$ and standard errors are reported in panel A. The average SI and average RI for each bucket and the number of observations are shown in panel B. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Return predictability by SI and RI

Panel A

		Coefficients		Standard Errors	
		Low RI	High RI	Low RI	High RI
Low SI	$\text{CAR}_{0,0}$	6.377***	11.028***	(0.258)	(0.466)
	$\text{CAR}_{0,0}$ (4pm-9:30am)	5.193***	7.894***	(0.31)	(0.58)
	$\text{CAR}_{1,1}$	0.953***	0.563	(0.225)	(0.385)
	$\text{CAR}_{1,10}$	0.174	-0.533	(0.704)	(1.191)
High SI	$\text{CAR}_{0,0}$	5.666***	11.673***	(0.21)	(0.483)
	$\text{CAR}_{0,0}$ (4pm-9:30am)	3.81***	9.317***	(0.231)	(0.609)
	$\text{CAR}_{1,1}$	0.73***	1.577***	(0.181)	(0.369)
	$\text{CAR}_{1,10}$	0.416	4.109***	(0.529)	(1.075)

Panel B

	Average SI		Average RI		Nobs	
	Low RI	High RI	Low RI	High RI	Low RI	High RI
Low SI	0.013	0.013	-0.917	0.564	182826	242336
High SI	0.051	0.058	-1.061	1.21	122740	123514

Table A14

We obtain the monthly median sentiment (Sent), median short interest (SI) and median residualized ownership (RI), and use these cutoffs to triple sort stock-day observations by sentiment, SI and RI (independently). For each bucket, we run the main specification in equation (2). The coefficient estimates $\hat{s}$ are reported in panels A and B, where A includes the buckets with below median sentiment and B includes the buckets with above median sentiment. Panels A and B also report the corresponding standard errors. The average Sent, average SI and average RI, and number of observations for each bucket are reported in panel C. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Return predictability by Sent, SI and RI

Panel A: Low Sent

		Coefficients		Standard Errors	
		Low RI	High RI	Low RI	High RI
Low SI	$CAR_{0,0}$	3.959***	6.601***	(0.399)	(0.73)
	$CAR_{0,0}$ (4pm-9:30am)	2.12***	1.021	(0.474)	(0.893)
	$CAR_{1,1}$	0.689*	0.275	(0.362)	(0.629)
	$CAR_{1,10}$	0.326	2.609	(1.147)	(2.002)
High SI	$CAR_{0,0}$	3.572***	7.299***	(0.351)	(0.788)
	$CAR_{0,0}$ (4pm-9:30am)	0.832*	3.074***	(0.446)	(1.004)
	$CAR_{1,1}$	0.976***	0.444	(0.282)	(0.616)
	$CAR_{1,10}$	1.244	4.042**	(0.81)	(1.818)

Panel B: High Sent

		Coefficients		Standard Errors	
		Low RI	High RI	Low RI	High RI
Low SI	$CAR_{0,0}$	4.53***	8.761***	(0.657)	(1.147)
	$CAR_{0,0}$ (4pm-9:30am)	1.071	3.747***	(0.784)	(1.408)
	$CAR_{1,1}$	0.36	0.639	(0.631)	(0.986)
	$CAR_{1,10}$	-1.556	2.149	(1.798)	(2.819)
High SI	$CAR_{0,0}$	4.845***	9.427***	(0.56)	(0.967)
	$CAR_{0,0}$ (4pm-9:30am)	1.31**	3.875***	(0.581)	(1.245)
	$CAR_{1,1}$	0.674	0.802	(0.515)	(0.867)
	$CAR_{1,10}$	1.291	4.149*	(1.457)	(2.488)

Panel C

	Average Sent				Average SI			
	Low Sent		High Sent		Low Sent		High Sent	
	Low RI	High RI	Low RI	High RI	Low RI	High RI	Low RI	High RI
Low SI	-0.026	-0.025	0.004	0.003	0.013	0.013	0.014	0.013
High SI	-0.027	-0.028	0.004	0.004	0.053	0.06	0.05	0.056
	Average RI				Nobs			
	Low Sent		High Sent		Low Sent		High Sent	
	Low RI	High RI	Low RI	High RI	Low RI	High RI	Low RI	High RI
Low SI	-0.871	0.568	-0.963	0.561	91692	123648	91134	118688
High SI	-1.082	1.181	-1.038	1.236	62510	58076	60230	65438

Table A15

This table shows the estimated β_0 from (A16). For the $\{1, 1\}$ regressions the dependent variable is the next day's sentiment $Sent^i_{t,1,1}$, and for the $\{1, 10\}$ regressions the dependent variable is the average sentiment measured over the next 10 days $Sent^i_{t,1,10}$. In both cases, the independent variable is the time t sentiment $Sent^i_t$ (or $Sent^i_{t,0,0}$). *** indicates significance at the 1% level or better. Note that for a given i and j, (A16) is the same for $\text{Retrf}_{i,j}$ and $\text{CAR}_{i,j}$.

News autocorrelation coefficient β_0 from (A16)

Panel A: News autocorrelation channel conditional on intermediary capacity								
	Capacity							
	CR (daily)		CR (monthly)		CR (quarterly)		Lev (quarterly)	
	$\hat{\beta}_0$	s.e.	$\hat{\beta}_0$	s.e.	$\hat{\beta}_0$	s.e.	$\hat{\beta}_0$	s.e.
$\text{Retrf}_{1,1}$	0.2503***	0.0029	0.2462***	0.0028	0.2462***	0.0028	0.2457***	0.0028
$\text{Retrf}_{1,10}$	0.1809***	0.002	0.1776***	0.0019	0.1776***	0.0019	0.1772***	0.0019
$\text{CAR}_{1,1}$	0.2503***	0.0029	0.2462***	0.0028	0.2462***	0.0028	0.2457***	0.0028
$\text{CAR}_{1,10}$	0.1809***	0.002	0.1776***	0.0019	0.1776***	0.0019	0.1772***	0.0019

Panel B: News autocorrelation channel conditional on mutual fund ownership						
	Ownership					
	Passive/Market		Active/Market		Passive/Fund Total	
	$\hat{\beta}_0$	s.e.	$\hat{\beta}_0$	s.e.	$\hat{\beta}_0$	s.e.
$\text{Retrf}_{1,1}$	0.2459***	0.0028	0.247***	0.0028	0.2462***	0.0028
$\text{Retrf}_{1,10}$	0.1772***	0.0019	0.1786***	0.0019	0.1772***	0.0019
$\text{CAR}_{1,1}$	0.2459***	0.0028	0.247***	0.0028	0.2462***	0.0028
$\text{CAR}_{1,10}$	0.1772***	0.0019	0.1786***	0.0019	0.1772***	0.0019

Table A16

This table reports the test statistic $g(\hat{\boldsymbol{\theta}}) \equiv \frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1}$, where the coefficient estimates $(\hat{s}_0, \hat{s}_1, \hat{\beta}_0, \hat{\beta}_1)$ come from the specification in (A10). These control vector $\boldsymbol{X}_t$ in these regressions contains: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, $SI(\%)$, $IO(\%)$, log($Market\ Cap$), $IHS(Book/Market)$, $log(Illiquidity)$, lagged α, $CAR^2_{0,0}$ and VIX. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Panel A: News autocorrelation channel conditional on intermediary capacity

Panel A: News autocorrelation channel conditional on intermediary capacity								
	Capacity							
	CR (daily)		CR (monthly)		CR (quarterly)		Lev (quarterly)	
	$g(\hat{\boldsymbol{\theta}})$	p-value	$g(\hat{\boldsymbol{\theta}})$	p-value	$g(\hat{\boldsymbol{\theta}})$	p-value	$g(\hat{\boldsymbol{\theta}})$	p-value
$\text{Retrf}_{1,1}$	-48.3473**	0.0127	-82.6166***	0.0024	-74.6312***	0.0031	-1237.6356	0.8358
$\text{Retrf}_{1,10}$	-54.8816***	0.0058	-135.7281***	0.0018	-114.6694***	0.005	310.9865***	8e-04
$\text{CAR}_{1,1}$	-46.5655**	0.0116	-81.4448***	0.0022	-73.3033***	0.0039	-3658.4688***	0.0042
$\text{CAR}_{1,10}$	-57.1644***	2e-04	-137.7646***	0	-117.1208***	1e-04	304.0765**	0.0109

Panel B: News autocorrelation channel conditional on mutual fund ownership

Panel B: News autocorrelation channel conditional on mutual fund ownership						
	Ownership					
	Passive/Market		Active/Market		Passive/Fund Total	
	$g(\hat{\boldsymbol{\theta}})$	p-value	$g(\hat{\boldsymbol{\theta}})$	p-value	$g(\hat{\boldsymbol{\theta}})$	p-value
$\text{Retrf}_{1,1}$	-922.3345***	0.0065	-14.8314	0.9493	268.3963*	0.0683
$\text{Retrf}_{1,10}$	-121.5613**	0.0111	-33.5105**	0.0276	264.3765**	0.0112
$\text{CAR}_{1,1}$	-947.6541**	0.0108	98.5824*	0.087	237.8326	0.1077
$\text{CAR}_{1,10}$	-116.6313***	0.0036	-41.4574***	0.0051	282.6225***	6e-04

Table A17

1-day ahead forecasting regressions. $Retrf_{i,j}$ ($CAR_{i,j}$) refers to the excess return (abnormal return) that includes days $t+i, \ldots, t+j$ where t is the event date. Returns are measured in percent.

One-day ahead return regressions

	Dependent variable:													
	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$
	1996-2018		1996-2000		2001		2002-2006		2007-2009		2010-2014		2015-2018	
Constant	0.126	0.136	-0.150	0.194	1.107^{*}	0.870^{*}	0.187	0.266	1.013	0.396	-0.035	0.089	-0.400	-0.278
Sent	1.192^{***}	0.914^{***}	2.129^{***}	1.584^{***}	0.566	1.314	1.160^{***}	0.899^{***}	1.174	1.156^{**}	0.598^{**}	0.227	0.480	0.719^{***}
$\text{CAR}_{0,0}$	0.001	0.001	-0.001	0.002	0.017	0.020	0.011	0.009	-0.015	-0.017	0.006	0.005	0.004	-0.002
$\text{CAR}_{-1,-1}$	-0.004	-0.008	-0.028^{***}	-0.024^{***}	0.003	0.001	-0.012	-0.011	0.016	0.0001	0.002	0.0005	0.001	-0.004
$\text{CAR}_{-2,-2}$	-0.009^{*}	-0.006	-0.009^{*}	-0.005	0.002	-0.0005	-0.004	-0.006	-0.016	-0.008	-0.004	-0.003	-0.007	-0.009
$\text{CAR}_{-30,-3}$	-0.001	-0.001	-0.0003	-0.0001	0.00002	0.001	-0.004^{**}	-0.003^{**}	0.001	-0.001	-0.001	-0.001	-0.003^{*}	-0.003^{**}
$\text{CAR}^2_{0,0}$	0.0005	0.0005	-0.001	-0.0005	-0.003^{*}	-0.003^{**}	0.0001	0.0001	0.001^{**}	0.001^{**}	-0.001^{*}	-0.001^{*}	0.001	0.0002
VIX	0.006	0.001	0.016^{*}	0.004^{**}	0.021	0.005	0.002	0.001	0.012	0.002	0.008	0.00000	0.012	-0.0004
SUE	0.011^{*}	0.007^{***}	0.0001	0.002	0.020	0.007	0.004	0.004	0.006	0.010	0.008	0.009^{**}	0.016^{***}	0.011^{***}
Short Interest (%)	-0.006^{*}	-0.004^{*}	-0.005	-0.006	-0.015	-0.006	-0.007	-0.004	-0.015	-0.007	-0.0003	-0.001	0.0001	0.001
IO (%)	-0.0002	-0.0002	0.001	0.0004	-0.001	-0.002	0.001	-0.0002	-0.005^{**}	-0.002^{*}	0.0002	0.00003	0.001^{*}	0.001^{*}
log(Market Cap)	-0.033	-0.014^{*}	0.026	0.008	-0.033	-0.079	-0.021	-0.013	-0.175^{*}	-0.014	0.015	-0.033^{**}	-0.022	-0.011
IHS(Book/Market)	0.024	0.0003	0.022	0.011	-0.029	0.006	0.072^{**}	0.015	0.004	-0.072	0.030	0.007	0.014	0.022
log(Illiquidity)	-0.026	-0.009	0.036	0.022	0.033	-0.048	-0.013	-0.003	-0.138	-0.005	0.018	-0.029^{**}	-0.031	-0.021^{*}
α	-0.015	0.048	0.260^{**}	0.226^{**}	0.124	0.225	0.018	0.020	-0.349	-0.143	0.129	0.048	-0.049	0.050
Observations	618,367	618,367	111,817	111,817	26,383	26,383	144,277	144,277	97,376	97,376	154,433	154,433	84,081	84,081
Adjusted R^2	0.001	0.0004	0.002	0.001	0.006	0.007	0.001	0.001	0.003	0.003	0.001	0.0003	0.001	0.0005

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table A18

This table replicates the return forecastability results from Tetlock, Saar-Tsechansky, and Macskassy (2008). 1-day ahead forecasting regressions. $Retrf_{i,j}$ ($CAR_{i,j}$) refers to the excess return (abnormal return) that includes days $t+i, \ldots, t+j$ where t is the event date. Returns are measured in percent. These regressions include as controls: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, $SI(\%)$, $IO(\%)$, log($Market\ Cap$), IHS($Book/Market$), $log(Illiquidity)$, lagged α, $CAR^2_{0,0}$ and VIX. The row label (4pm-9:30am) indicates that $Sent$ has been measured from the prior day's close to the event day's market open. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Replication of return results from Tetlock, Saar-Tsechansky, and Macskassy (2008)

	Dependent variable:													
	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$	$\text{Retrf}_{1,1}$	$\text{CAR}_{1,1}$
	1996-2018		1996-2000		2001		2002-2006		2007-2009		2010-2014		2015-2018	
Constant	-0.045	0.080	0.039	0.384^{**}	0.218	0.600	0.176	0.143	-0.090	-0.338	-0.053	0.123	-0.203	-0.080
Sent	1.258^{***}	0.961^{***}	2.125^{***}	1.674^{***}	0.228	1.140	1.424^{***}	0.976^{***}	1.407	1.316^{**}	0.559^{**}	0.224	0.464	0.696^{***}
$\text{CAR}_{0,0}$	0.0001	-0.001	-0.003	0.0001	0.017	0.012	0.006	0.004	-0.016	-0.017	0.006	0.005	0.003	-0.002
$\text{CAR}_{-1,-1}$	-0.006	-0.010^{**}	-0.025^{***}	-0.024^{***}	-0.028	-0.012	-0.012	-0.012	0.016	-0.00003	0.002	0.0004	0.002	-0.003
$\text{CAR}_{-2,-2}$	-0.011^{**}	-0.007^{*}	-0.013^{**}	-0.008	0.001	-0.001	-0.008	-0.009	-0.017	-0.009	-0.004	-0.003	-0.007	-0.009
$\text{CAR}_{-30,-3}$	-0.001	-0.001	-0.0004	-0.0004	-0.002	-0.001	-0.004^{**}	-0.003^{**}	0.001	-0.0004	-0.001	-0.001	-0.003^{*}	-0.003^{**}
$\text{CAR}^2_{0,0}$	0.0004	0.0004	-0.001^{*}	-0.001	-0.003^{**}	-0.003^{**}	0.0001	0.0001	0.001^{**}	0.001^{**}	-0.001	-0.001^{*}	0.0005	0.0001
VIX	0.006	0.001	0.017^{**}	0.004^{**}	0.031	0.008	0.001	0.001	0.010	0.002	0.008	-0.001	0.011	-0.001
α	-0.006	0.096^{*}	0.179^{*}	0.179^{**}	0.085	0.296	-0.015	0.070	-0.218	-0.131	0.110	0.074	-0.024	0.046
SUE	0.014^{**}	0.009^{***}	0.006	0.007	0.028	0.016	0.006	0.004	0.011	0.010	0.008	0.009^{**}	0.015^{***}	0.010^{***}
log(Market Cap)	-0.005	-0.002	-0.015	-0.009	-0.045	-0.023	-0.010	0.001	-0.016	0.011	-0.006	-0.002	0.008	0.008
IHS(Book/Market)	0.030	-0.010	-0.017	-0.033	-0.040	-0.008	0.073^{**}	0.018	-0.007	-0.067	0.028	0.008	0.003	0.010
log(Share Turnover)	-0.013	0.008	-0.001	0.040^{**}	-0.023	0.033	-0.012	0.032^{**}	-0.044	-0.012	-0.019	0.012	0.023	0.018
Observations	647,078	647,078	125,136	125,136	30,367	30,367	150,999	150,999	98,390	98,390	156,317	156,317	85,869	85,869
Adjusted R^2	0.001	0.0005	0.002	0.001	0.006	0.005	0.001	0.001	0.003	0.003	0.001	0.0003	0.001	0.0004

Note: *p<0.1; **p<0.05; ***p<0.01

Table A19

10-day ahead forecasting regressions. $Retrf_{i,j}$ ($CAR_{i,j}$) refers to the excess return (abnormal return) that includes days $t+i,\ldots,t+j$ where t is the event date. Returns are measured in percent. These regressions include as controls: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, SI(%), IO(%), log($Market\ Cap$), IHS($Book/Market$), log($Illiquidity$), lagged α, $CAR^2_{0,0}$ and VIX. The row label (4pm-9:30am) indicates that $Sent$ has been measured from the prior day's close to the event day's market open. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Ten-day ahead return regressions

	Dependent variable:													
	$Retrf_{1,10}$	$CAR_{1,10}$	$Retrf_{1,10}$	$CAR_{1,10}$	$Retrf_{1,10}$	$CAR_{1,10}$	$Retrf_{1,10}$	$CAR_{1,10}$	$Retrf_{1,10}$	$CAR_{1,10}$	$Retrf_{1,10}$	$CAR_{1,10}$	$Retrf_{1,10}$	$CAR_{1,10}$
	1996-2018		1996-2000		2001		2002-2006		2007-2009		2010-2014		2015-2018	
Constant	3.396^{***}	1.975^{***}	0.114	1.828^{***}	4.102^{**}	7.289^{***}	2.493^{**}	1.968^{***}	18.527^{***}	6.089^{***}	1.146^{*}	1.140^{***}	-3.316^{***}	-2.005^{***}
Sent	2.793^{***}	0.821^{*}	4.604^{***}	2.748^{***}	-0.528	1.867	2.989^{***}	2.632^{***}	4.210^{*}	0.096	1.280	-0.813	-1.796^{*}	-0.979
$CAR_{0,0}$	-0.040^{***}	-0.038^{***}	-0.013	-0.007	-0.035	-0.015	-0.032^{*}	-0.032^{*}	-0.111^{***}	-0.102^{***}	0.005	-0.010	-0.009	-0.020
$CAR_{-1,-1}$	-0.051^{***}	-0.049^{***}	-0.027^{**}	-0.017	-0.030	-0.025	-0.079^{***}	-0.066^{***}	-0.086^{**}	-0.099^{**}	-0.0001	-0.002	-0.047^{**}	-0.045^{**}
$CAR_{-2,-2}$	-0.065^{***}	-0.059^{***}	-0.043^{***}	-0.027^{*}	-0.006	0.040	-0.079^{***}	-0.068^{**}	-0.128^{***}	-0.132^{***}	-0.003	-0.001	-0.008	-0.030
$CAR_{-30,-3}$	-0.005^{*}	-0.007^{***}	-0.005	0.001	-0.020^{**}	-0.013	-0.028^{***}	-0.021^{***}	0.010	-0.006	-0.004	-0.005	-0.012^{**}	-0.020^{***}
$CAR^2_{0,0}$	0.002	0.003^{***}	-0.001	-0.001	-0.003	-0.002	0.005^{***}	0.006^{***}	0.002	0.004^{**}	-0.0003	0.001	0.003^{**}	0.002^{*}
VIX	0.020	0.002	0.084^{***}	0.008	0.339^{***}	0.042^{***}	0.018	0.004	0.024	0.006	0.027	-0.005^{**}	0.118^{***}	0.007
SUE	0.039^{**}	0.021^{***}	-0.035^{*}	-0.028	0.134^{***}	0.008	-0.025	-0.012	-0.038	0.069^{***}	0.059^{***}	0.026^{**}	0.073^{***}	0.056^{***}
Short Interest (%)	-0.027^{***}	-0.010^{*}	-0.008	-0.015	-0.081^{***}	-0.052^{**}	-0.020	0.017	-0.090^{***}	-0.025	0.018^{*}	-0.005	0.023	0.017
IO (%)	-0.002^{**}	-0.002^{***}	0.001	0.0001	0.001	-0.018^{***}	0.007^{***}	0.0002	-0.048^{***}	-0.024^{***}	0.0003	-0.0005	0.004^{***}	0.003^{***}
log(Market Cap)	-0.218^{***}	-0.121^{***}	0.188^{*}	0.038	0.127	-0.231	-0.266^{*}	-0.113	-0.936^{***}	-0.091	0.322^{***}	-0.273^{***}	-0.078	-0.005
IHS(Book/Market)	0.218^{***}	-0.001	0.235^{*}	0.258^{**}	0.453^{**}	0.729^{***}	0.433^{***}	-0.010	0.321	-0.452^{***}	0.122^{*}	0.032	-0.027	-0.053
log(Illiquidity)	-0.084	-0.046^{**}	0.276^{***}	0.136^{*}	0.715^{***}	0.085	-0.146	-0.031	-0.292	0.086	0.376^{***}	-0.234^{***}	-0.148^{**}	-0.072^{**}
α	-0.360^{*}	-0.052	1.470^{***}	1.676^{***}	-0.350	1.206^{**}	-0.318	-0.345	-1.871^{***}	-2.029^{***}	1.198^{***}	0.622^{***}	-1.056^{**}	-0.449
Observations	618,369	618,369	111,817	111,817	26,383	26,383	144,278	144,278	97,376	97,376	154,433	154,433	84,082	84,082
Adjusted R^2	0.002	0.002	0.004	0.001	0.042	0.007	0.006	0.006	0.010	0.010	0.002	0.001	0.012	0.003

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table A20
These regressions include as controls: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, $SI(\%)$, $IO(\%)$, $\log(Market\ Cap)$, $IHS(Book/Market)$, $log(Illiquidity)$, lagged α, $CAR_{0,0}^2$ and VIX. The $\text{Retrf}_{0,0}$ and $\text{CAR}_{0,0}$ regressions omit the $\text{CAR}_{0,0}$ control. The row label (4pm-9:30am) indicates that *Sent* has been measured from the prior day's close to the event day's market open. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Return predictability

		1996-2018	1996-2000	2001	2002-2006	2007-2009	2010-2014	2015-2018
$\text{Retrf}_{0,0}$	Sent	9.184***	9.842***	12.611***	9.694***	12.709***	5.117***	9.022***
$\text{Retrf}_{0,0}$	Sent (4pm-9:30am)	6.18***	6.01***	8.783***	7.47***	7.076***	3.888***	5.87***
$\text{Retrf}_{1,1}$	Sent	1.192***	2.129***	0.566	1.16***	1.174	0.598**	0.48
$\text{Retrf}_{1,10}$	Sent	2.793***	4.604***	-0.528	2.989***	4.21*	1.28	-1.796*
$\text{CAR}_{0,0}$	Sent	8.086***	9.209***	10.562***	9.084***	9.715***	4.404***	8.251***
$\text{CAR}_{0,0}$	Sent (4pm-9:30am)	5.949***	5.718***	7.594***	7.445***	6.96***	3.946***	5.495***
$\text{CAR}_{1,1}$	Sent	0.914***	1.584***	1.314	0.899***	1.156**	0.227	0.719***
$\text{CAR}_{1,10}$	Sent	0.821*	2.748***	1.867	2.632***	0.096	-0.813	-0.979

Table A21

These regressions include as controls: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, $SI(\%)$, $IO(\%)$, $\log(Market\ Cap)$, $IHS(Book/Market)$, $log(Illiquidity)$, lagged α, $CAR^2_{0,0}$ and VIX. The row label (4pm-9:30am) indicates that *Sent* has been measured from the prior day's close to the event day's market open. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels. Each ownership series in these regressions has been trimmed to exclude in each month the bottom 2.5% of observations.

Mutual fund ownership effects on sentiment predictability (trimmed ownership)

Return regressions

		Mutual Fund Ownership (%)		
		Passive/Market	Active/Market	Passive/Fund Total
$\text{Retrf}_{0,0}$	Sent	9.306***	9.212***	9.216***
	Sent×Ownership	-0.071	0.204***	-0.053***
$\text{Retrf}_{0,0}$	Sent (4pm-9:30am)	6.272***	6.194***	6.277***
	Sent (4pm-9:30am)×Ownership	-0.061	0.213***	-0.059***
$\text{Retrf}_{1,1}$	Sent	1.17***	1.186***	1.178***
	Sent×Ownership	-0.08	0.017	-0.013
$\text{Retrf}_{1,10}$	Sent	2.77***	2.861***	2.776***
	Sent×Ownership	-0.498***	0.216**	-0.137***

CAR regressions

		Mutual Fund Ownership (%)		
		Passive/Market	Active/Market	Passive/Fund Total
$\text{CAR}_{0,0}$	Sent	8.169***	8.081***	8.088***
	Sent×Ownership	-0.048	0.188***	-0.045***
$\text{CAR}_{0,0}$	Sent (4pm-9:30am)	6.002***	5.935***	6.004***
	Sent (4pm-9:30am)×Ownership	-0.007	0.189***	-0.036**
$\text{CAR}_{1,1}$	Sent	0.925***	0.926***	0.932***
	Sent×Ownership	-0.047	0.032	-0.018*
$\text{CAR}_{1,10}$	Sent	0.975**	0.932**	0.894**
	Sent×Ownership	-0.346***	0.15**	-0.115***

Table A22
These regressions include as controls: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, $SI(\%)$, $IO(\%)$, $\log(Market\ Cap)$, $IHS(Book/Market)$, $log(Illiquidity)$, lagged α, $CAR^2_{0,0}$ and VIX. The row label (4pm-9:30am) indicates that *Sent* has been measured from the prior day's close to the event day's market open. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels. These specifications drop all event days that fall on earnings announcement days, or on subsequent business days.

Return predictability (dropped earnings days)

		1996-2018	1996-2000	2001	2002-2006	2007-2009	2010-2014	2015-2018
$\text{Retrf}_{0,0}$	Sent	7.684***	8.972***	11.802***	7.967***	10.834***	3.832***	6.137***
$\text{Retrf}_{0,0}$	Sent (4pm-9:30am)	4.832***	5.382***	8.283***	5.951***	5.045***	2.543***	3.574***
$\text{Retrf}_{1,1}$	Sent	1.131***	2.047***	0.301	1.075***	1.021	0.691**	0.345
$\text{Retrf}_{1,10}$	Sent	3.012***	4.731***	-0.202	2.931***	4.075	1.524*	-1.624
$\text{CAR}_{0,0}$	Sent	6.599***	8.323***	9.809***	7.4***	7.752***	3.207***	5.332***
$\text{CAR}_{0,0}$	Sent (4pm-9:30am)	4.637***	5.133***	7.089***	5.966***	4.898***	2.716***	3.088***
$\text{CAR}_{1,1}$	Sent	0.87***	1.444***	0.854	0.875***	1.115**	0.29	0.619**
$\text{CAR}_{1,10}$	Sent	0.85*	2.709**	2.576	2.581***	-0.28	-0.798	-1.102

Table A23
These regressions include as controls: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, $SI(\%)$, $IO(\%)$, $\log(Market\ Cap)$, $IHS(Book/Market)$, $log(Illiquidity)$, and lagged α. The row label (4pm-9:30am) indicates that *Sent* has been measured from the prior day's close to the event day's market open. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels. These specifications *do not include* the VIX and the squared lagged CARs as explanatory variables.

Return predictability (no volatility controls)

		1996-2018	1996-2000	2001	2002-2006	2007-2009	2010-2014	2015-2018
$\text{Retrf}_{0,0}$	Sent	9.593***	9.956***	14.579***	10.2***	13.775***	5.484***	9.177***
$\text{Retrf}_{0,0}$	Sent (4pm-9:30am)	6.552***	6.116***	10.493***	7.96***	7.9***	4.328***	5.931***
$\text{Retrf}_{1,1}$	Sent	1.002***	2.038***	0.314	1.118***	0.359	0.526*	0.42
$\text{Retrf}_{1,10}$	Sent	2.225***	4.104***	-5.802	2.291**	2.673	1.009	-2.373**
$\text{CAR}_{0,0}$	Sent	7.993***	9.083***	11.76***	9.099***	9.558***	4.452***	8.215***
$\text{CAR}_{0,0}$	Sent (4pm-9:30am)	5.835***	5.584***	8.499***	7.454***	6.701***	3.972***	5.468***
$\text{CAR}_{1,1}$	Sent	0.861***	1.58***	1.32	0.879***	0.926*	0.233	0.718***
$\text{CAR}_{1,10}$	Sent	0.627	2.771***	1.286	2.187**	-0.62	-0.768	-1.031

Table A24
These regressions include as controls: constant, $CAR_{0,0}$, $CAR_{-1,-1}$, $CAR_{-2,-2}$, $CAR_{-30,-3}$, SUE, $SI(\%)$, $IO(\%)$, $\log(Market\ Cap)$, $IHS(Book/Market)$, $log(Illiquidity)$, lagged α, $CAR^2_{0,0}$ and VIX. The row label (4pm-9:30am) indicates that *Sent* has been measured from the prior day's close to the event day's market open. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

VIX effects on sentiment predictability

Return regressions

		VIX
$\text{Retrf}_{0,0}$	Sent	0.477
	Sent×VIX	0.425***
$\text{Retrf}_{0,0}$	Sent (4pm-9:30am)	2.534*
	Sent (4pm-9:30am)×VIX	0.178**
$\text{Retrf}_{1,1}$	Sent	2.395
	Sent×VIX	-0.059
$\text{Retrf}_{1,10}$	Sent	9.571**
	Sent×VIX	-0.331

CAR regressions

		VIX
$\text{CAR}_{0,0}$	Sent	3.26***
	Sent×VIX	0.235***
$\text{CAR}_{0,0}$	Sent (4pm-9:30am)	3.953***
	Sent (4pm-9:30am)×VIX	0.098**
$\text{CAR}_{1,1}$	Sent	1.053*
	Sent×VIX	-0.007
$\text{CAR}_{1,10}$	Sent	6.501***
	Sent×VIX	-0.277***

Table A25

We test the null hypothesis H_0 from (21) that regression results are indistinguishable using sentiment from the Thomson Reuters unrestricted sample versus sentiment from Thomson Reuters but restricted to firm-day observations which overlap with those of three alternative news sources. A "–" indicates that the coefficient estimates from the unrestricted Thomson Reuters archive and from the Thomson Reuters archive restricted to the same firm-day observations as an alternative news source have different signs but are not statistically different; stars without an ✗ indicate the coefficients are statistically different but their signs are the same. Statistically significant qualitative differences – requiring different signs *and* statistically different coefficients – are indicated by an ✗. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Comparing results using Thomson Reuters sentiment for the unrestricted sample versus Thomson Reuters sentiment for the restricted samples

		Full	Intermediary				Mutual Fund Ownership			Entropy			
			CR (Daily)	CR (Monthly)	CR (Quarterly)	Lev (Quarterly)	Passive/Market	Active/Market	Passive/Fund Total	Daily	Monthly	Quarterly	Annual
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Retrf$_{0,0}$	DJ												
	WSJ								*				
	FT	***	–	–		***	–	***					
Retrf$_{1,1}$	DJ						–		–				
	WSJ						–						
	FT				–			–	–				
Retrf$_{1,10}$	DJ							*					
	WSJ									–		–	
	FT									–	**	*	–
CAR$_{0,0}$	DJ						–			✗*	*		
	WSJ	***											
	FT					**	✗**	***		✗**	**		
CAR$_{1,1}$	DJ						–						
	WSJ					–	–						
	FT				–	–		–	–				
CAR$_{1,10}$	DJ	–						*					
	WSJ	–				✗***	✗*			–		–	–
	FT	–	**	**	**	✗**	✗**		–	✗**		✗**	✗**