Internet Appendix: Time Variation in the News-Returns Relationship

Paul Glasserman* Fulin Li† Harry Mamaysky‡

March 13, 2022

Contents

*Columbia Business School, pg20@columbia.edu.
†University of Chicago Booth School of Business, fli3@chicagobooth.edu.
‡Columbia Business School, hm2646@columbia.edu.

A1 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 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 A1.2 and Section A1.3 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).

A1.2 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.

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

(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.

A1.3 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.

A2 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 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.

The sentiment of article j is defined as

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

A2.2 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.

A3 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: ` ~ ! @ # $ % ^ * ( ) _ + − = [ ] \ { } | ; : " " , . / < > ?
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, 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.

A4 Trim 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

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

[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.

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 onwards, 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.

To mitigate the concern that these outliers 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 A11.

A5 Impulse response functions

We calculate impulse response functions to a sentiment shock using the local projection method of Jorda (2005). We run regressions (12) and (13) 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.[5] 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 (12) and (13).

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

[5]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.

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

A6 Tests of the news autocorrelation channel

We provide a formal derivation of the test statistic in Section 6.2. We start from a generic setting and derive a general argument, then apply the results to our setting.

A6.1 Derive the test statistic

Consider the following generic data generating process:

$$Y = \theta W + \xi \tag{A1}$$

$$W = \boldsymbol{\beta}' \boldsymbol{Z} + \eta \tag{A2}$$

Equations (A1) and (A2) imply that

$$Y = \boldsymbol{s}' \boldsymbol{Z} + \varepsilon, \boldsymbol{s} = \theta \boldsymbol{\beta}, \varepsilon = \theta \eta + \xi \tag{A3}$$

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

We want to test the null hypothesis $H_0 : \mathbb{E}[\xi | \boldsymbol{Z}] = 0$. 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 $\mathbb{E}[\eta | \boldsymbol{Z}] = 0$, we have $\mathbb{E}[\varepsilon | \boldsymbol{Z}] = 0$. Let $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{\beta}}$ denote the consistent estimates of $\boldsymbol{s}$ and $\boldsymbol{\beta}$ from OLS regressions (A3) and (A2), respectively. Then

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

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 (A3) and equation (A2), respectively.

$$Y^i_{t,u,v} = s_0 \times Sent^i_t + s_1 \times Sent^i_t \times Capacity_t + s_2 \times Capacity_t + \boldsymbol{\gamma}' \boldsymbol{X}^i_t + \varepsilon^i_{t,u,v} \quad \text{(A4)}$$

$$Sent^i_{t,u,v} = \beta_0 \times Sent^i_t + \beta_1 \times Sent^i_t \times Capacity_t + \beta_2 \times Capacity_t + \boldsymbol{\delta}' \boldsymbol{X}^i_t + \eta^i_{t+1} \quad \text{(A5)}$$

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 $g\left(\boldsymbol{\theta}\right) = \frac{s_0}{s_1} - \frac{\beta_0}{\beta_1}$. 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 A18 shows the β_0 estimates from the regression in (A5).

A6.2 Derive the p-value

Instead of using the Delta method to get the p-values for the test statistics,we propose the following simulation method.[6]

1. For each year y, run regressions (A4) and (A5), 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.

[6]The Delta method does not work well because the test statistic is far from 0. See Table 9.

3. Run the full panel regressions (A4) and (A5), 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 $\hat{\boldsymbol{C}}_1$ 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 $g_j^{sim} = \frac{\hat{s}_{0,j}}{\hat{s}_{1,j}} - \frac{\hat{\beta}_{0,j}}{\hat{\beta}_{1,j}}$, $\bar{g}^{sim} = \frac{1}{J}\sum_{j=1}^{J} g_j^{sim}$, $\hat{g} = \frac{\hat{s}_0}{\hat{s}_1} - \frac{\hat{\beta}_0}{\hat{\beta}_1}$.

A7 Return response to news

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 A7 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 A7, 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).[7] 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.[8] The magnitude of the underreaction in the

[7]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.

[8]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.

most recent time period is similar to the full-sample coefficient of 0.884.[9]

Table A8 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 illiqudity 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 A7 slightly diminishes the role of *Sent* in most subperiods. Our full-sample results in Table A8 are even closer to TSM.

Table A9 shows the results for $Retrf$ and CAR of the ten-day ahead returns regressions. Table A10 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 2018); 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).

A8 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

[9]Murray, Xiao and Xia (2020) 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 A7 suggesting that the phenomena we examine may impact a broad class of return patterns.

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.[10] 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.[11]

Figure A7 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.[12] Table A6 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 each article.[13] We lag the sentiment window by three days because of potential uncertainty as to the accuracy of the earnings announcement date.[14] While an earnings event on trading day t will enter our sample only if company j was a member of the S&P500

[10]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.

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

[12]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 A12 (discussed below), as Table A13 in the Internet Appendix shows. Also, the results in Figure A6 (discussed in Section 5.3) are not impacted by the exclusion of these quarters.

[13]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.

[14]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.

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.[15] 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

$$SUE^i_{t+1} \text{ or } SAFE^i_{t+1} = s_0 \times Sent^i_t + \boldsymbol{\beta}' \boldsymbol{X}^i_t + \epsilon^i_t, \tag{A6}$$

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 A6. Standard errors for the earnings regressions are clustered by quarter.

Table A12 summarizes the results of this analysis.[16] 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.[17] The informativeness of news, as measured by the magnitudes and significance of the coefficients in Table A12, has varied over time. There is little evidence of either an upward or downward secular trend in the s_0 estimates.

As a robustness check of entropy as a measure of news informativeness, Panel D of Figure A6 shows the s_0 coefficient from annual versions of the SUE regression in (A6) plotted against annual average entropy. The two series are highly correlated, supporting our interpretation of both as measures of news informativeness.[18] Panel E shows the

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

[16] Table A14 of the Internet Appendix shows the complete full-sample regression results.

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

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

annual $CAR_{0,0}$ sensitivity to contemporaneous news plotted against the annual SUE-sentiment coefficient s_0 from (A6). 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 (A6) 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 (A6). 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.

A9 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.[19] Panel A of Figure A6 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

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

lagged news during high information periods. Panel B of Figure A6 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 A6 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.

References

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

Murray, S., H. Xiao, and Y. Xia, 2020, "Charting by machines," working paper.

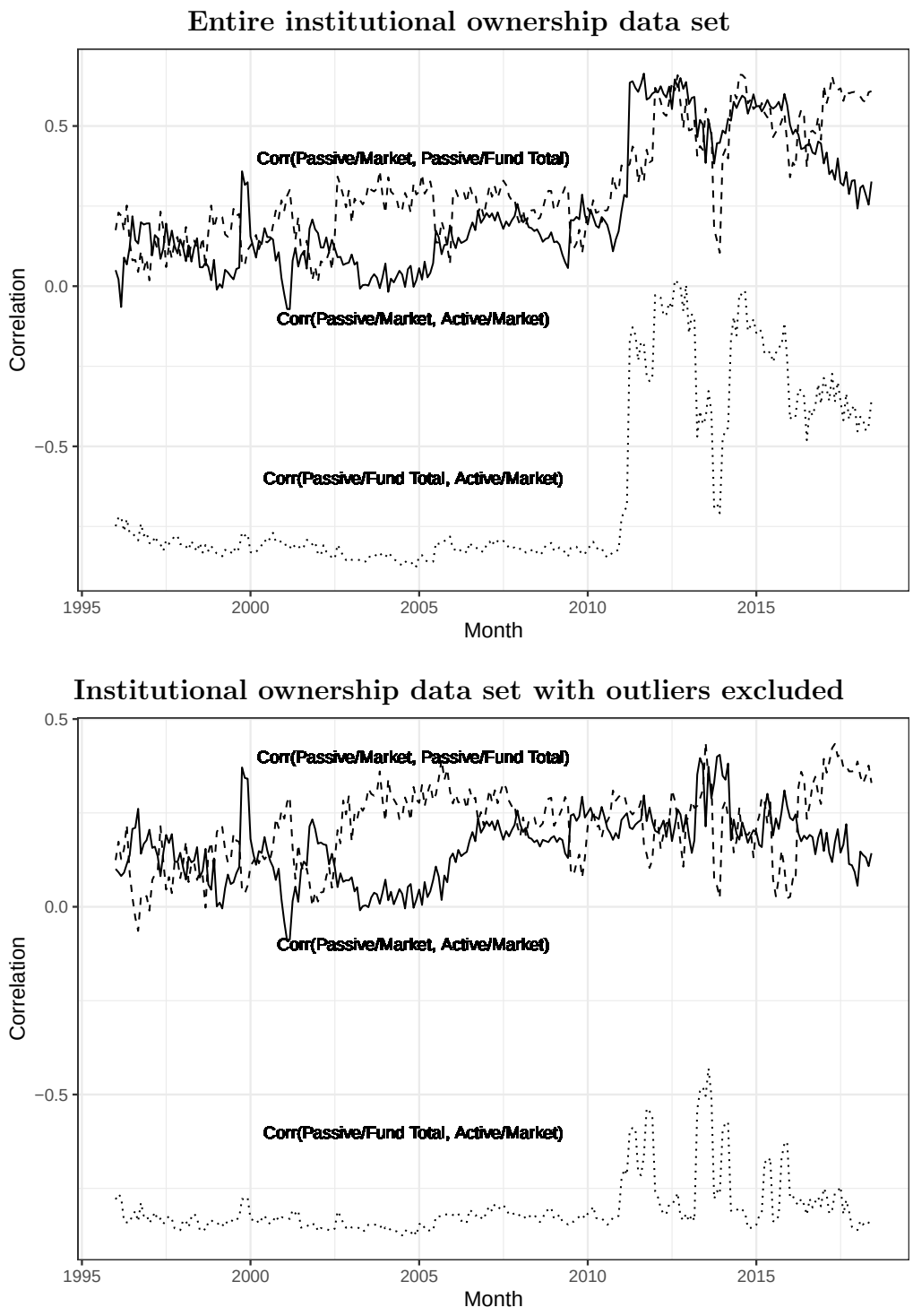


Fig. A1. Time series of cross-sectional ownership correlations. Within each month, this chart shows the cross-sectional correlations of our three ownership measures. The top panel shows the results for the full data set. The bottom panel shows the results when excluding the bottom 2.5% of each series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

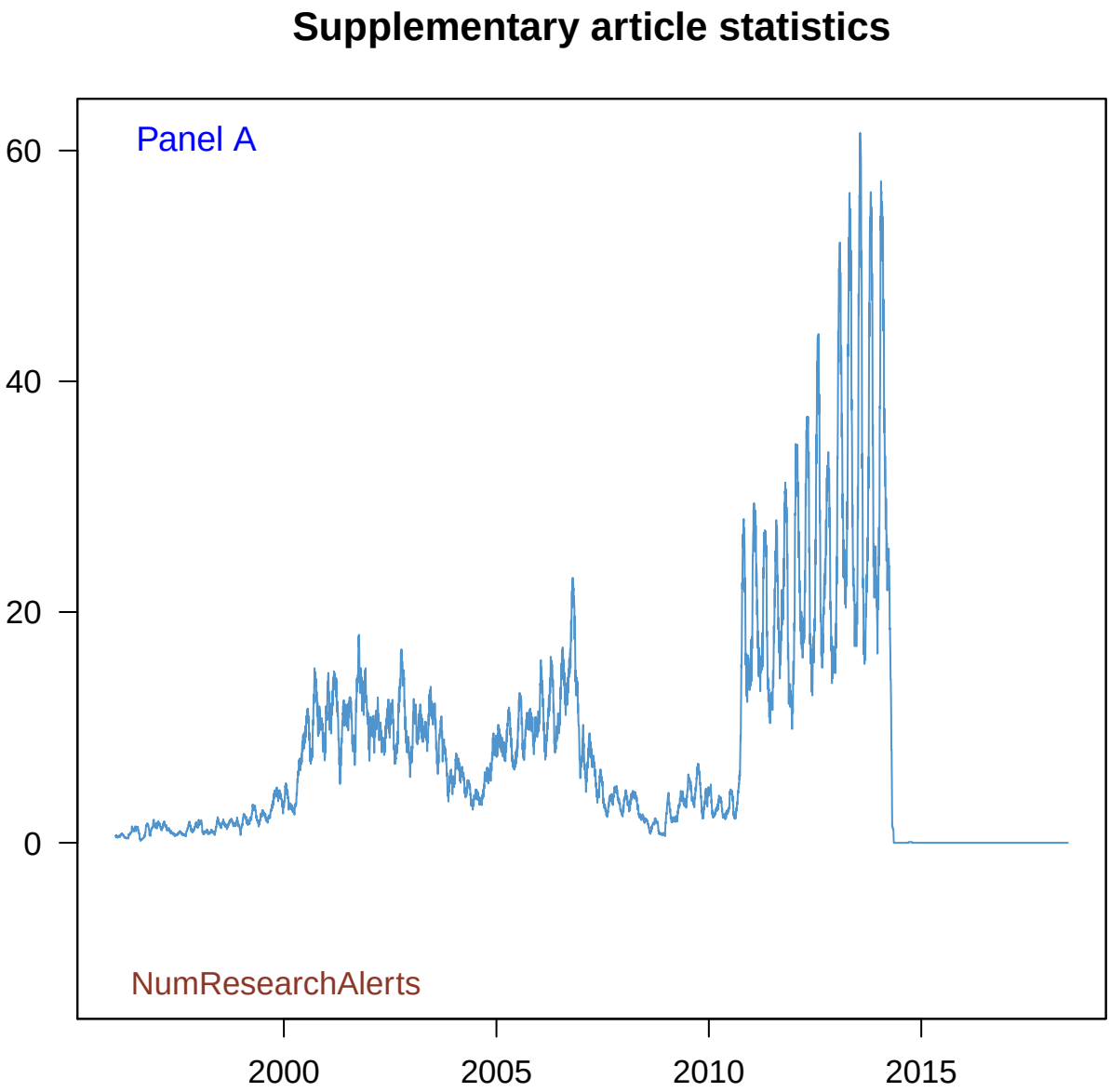


Fig. A2. This chart shows the daily number of articles with headlines containing "RESEARCH ALERT-" (case insensitive match).

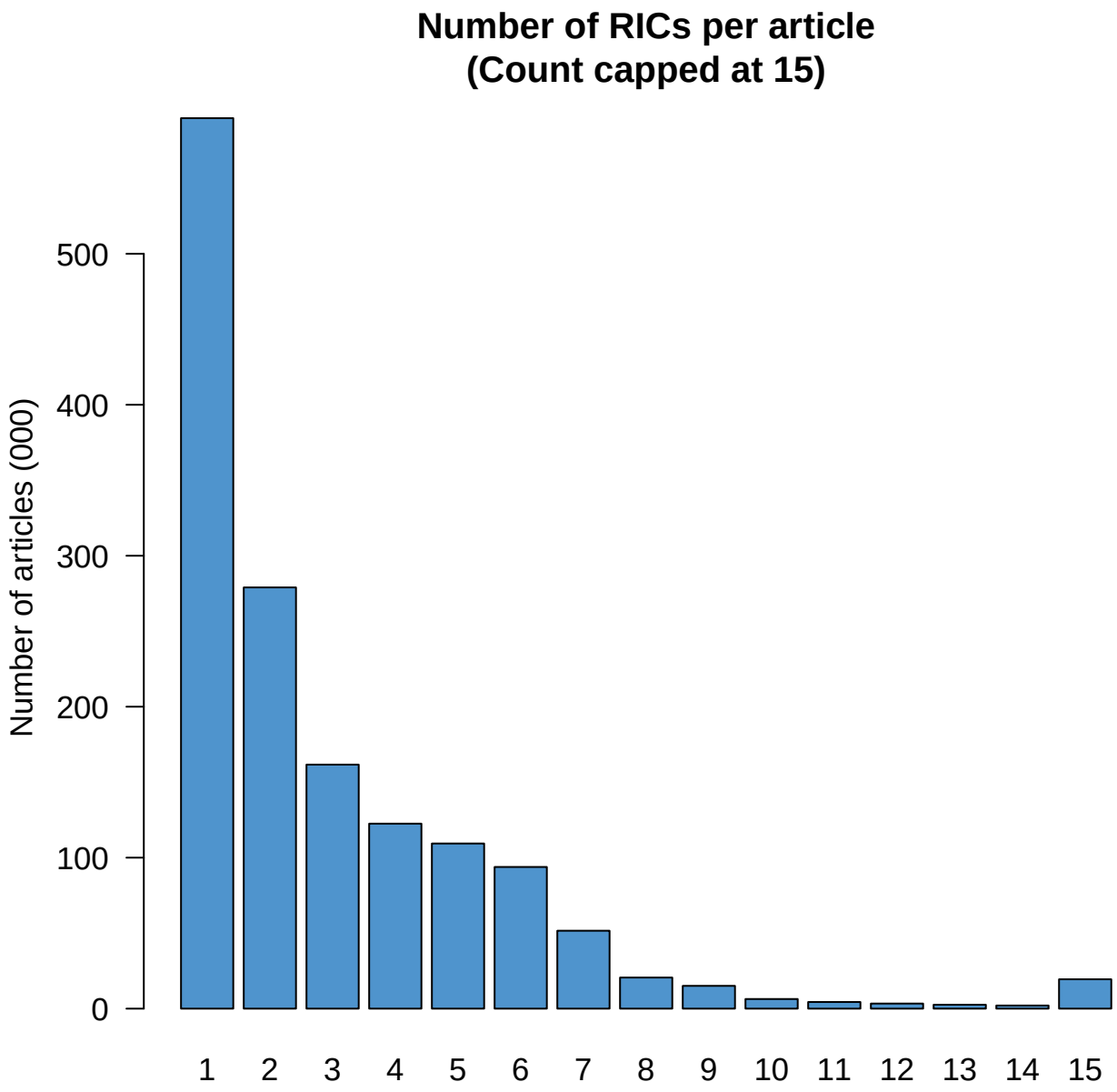
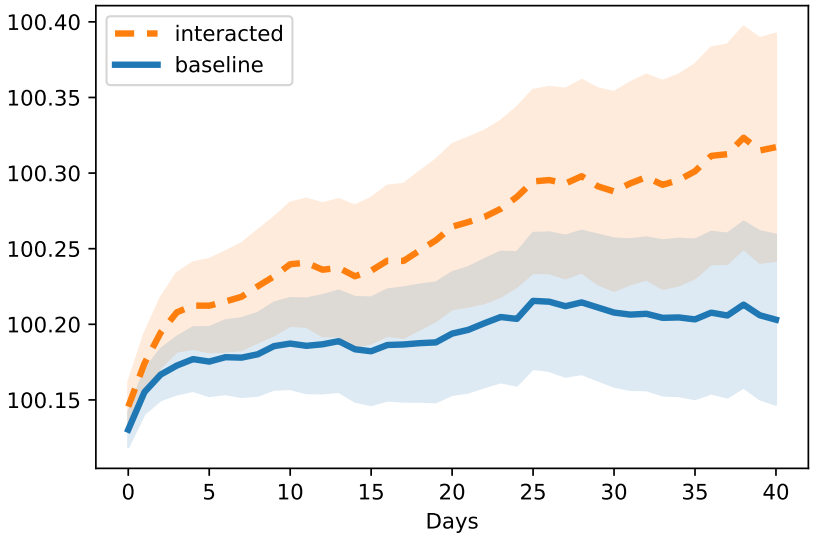


Fig. A3. This figure shows the histogram of the number of RICs (Reuters company identifier) per article. The y-axis is labeled with the number of articles in each RICs bucket, in thousands.

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

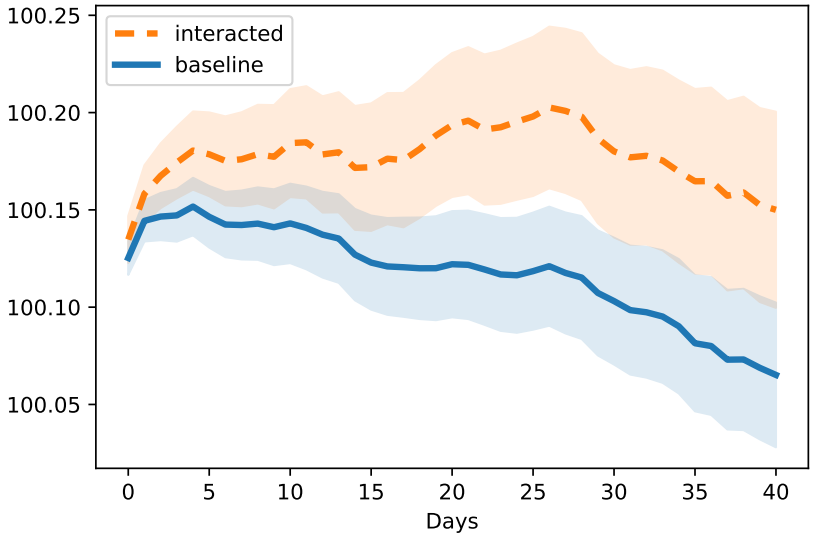
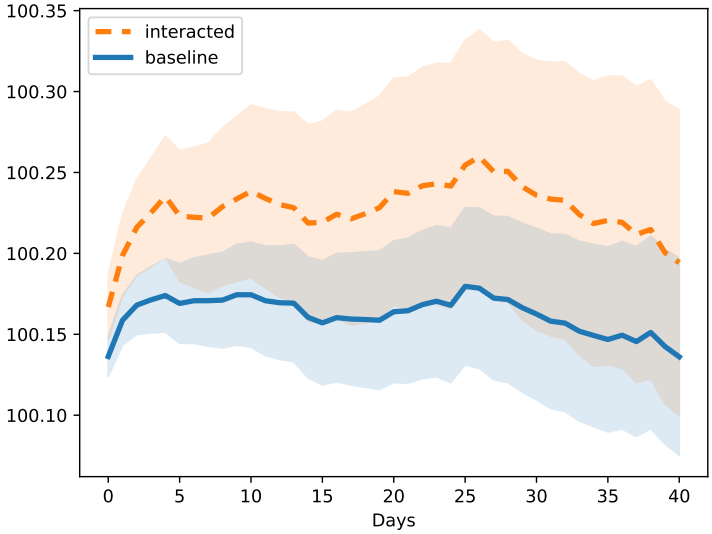


Fig. A4. Impulse response functions estimated using the local projection method of Jorda (2005). The figure shows the baseline response (labeled *baseline*) of future excess returns and cumulative abnormal returns ($CARs$) to a one standard deviation sentiment shock, as well as the response conditional 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

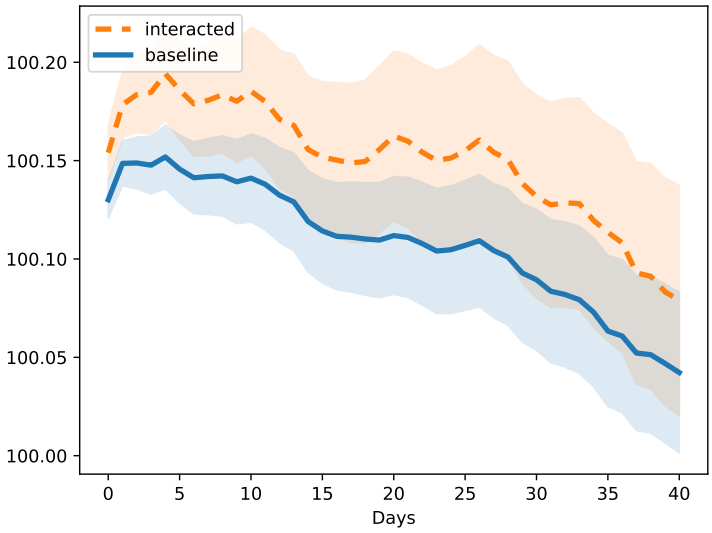


Fig. A5. Impulse response functions estimated using the local projection method of Jorda (2005). The figure shows the baseline response (labeled *baseline*) of future excess returns and cumulative abnormal returns (CARs) to a one standard deviation sentiment shock, as well as the response conditional 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.

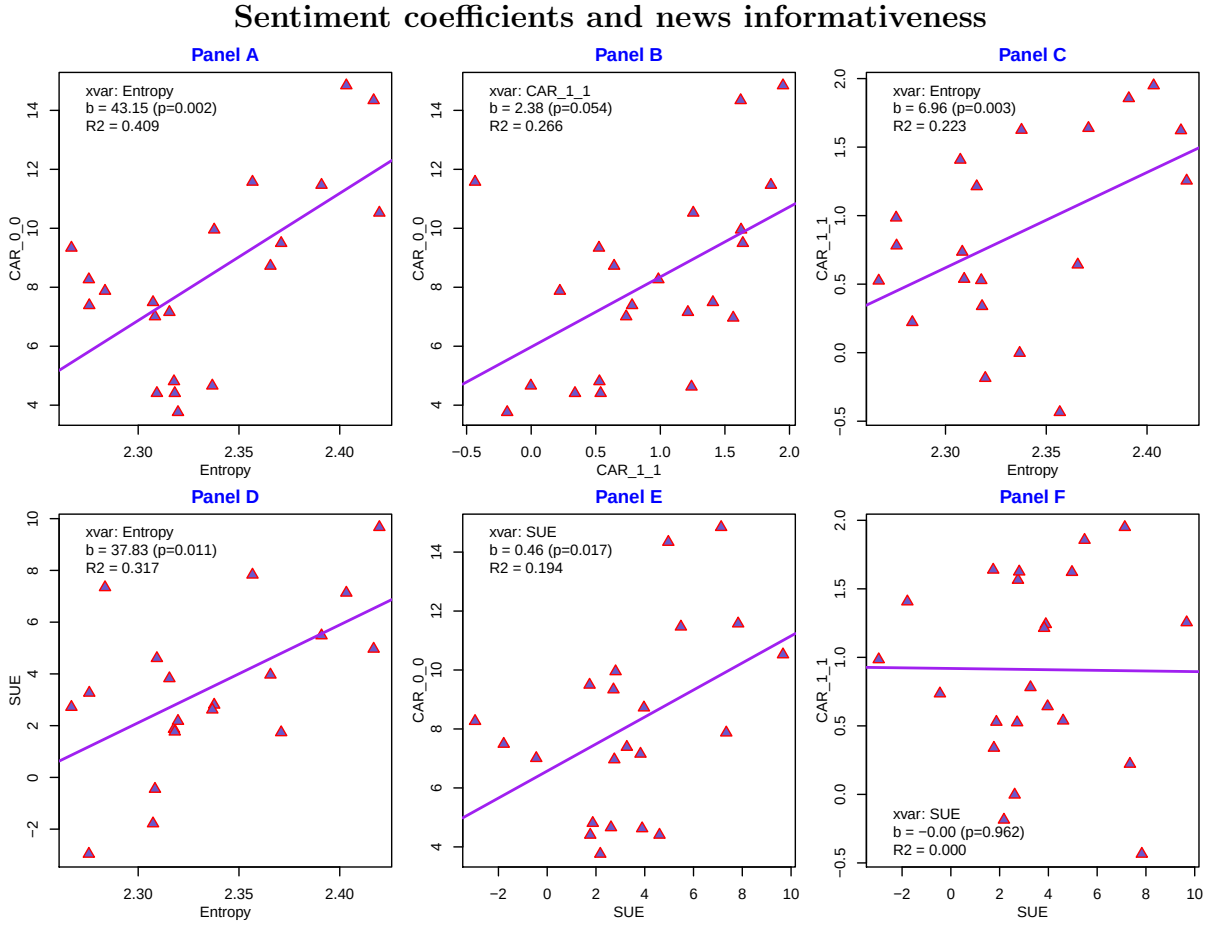


Fig. A6. Panel A shows the sentiment coefficients from annual regressions of returns $CAR_{0,0}$ on contemporaneous sentiment (eq. 3) plotted against annual average entropy. Panel B shows the sentiment coefficients from the $CAR_{0,0}$ regression plotted against the sentiment coefficients from an annual 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. A6) 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.

Quarterly cross-sectional standard deviation of SUE and SAFE

Fig. A7. The top panel shows the quarterly cross-sectional standard deviation of SUE. The bottom panel shows the quarterly cross-sectional standard deviation of $SAFE$.

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			
“”# $ % & ‘’( ) ∗+ − \<= >@ [ ] ^ ` {|}˜	␣		

Table A6
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
$CAR_{-2,-2}$	40,478	0.029	1.918	−37.216	−0.803	0.806	55.365
$CAR_{-30,-3}$	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 A7

1-day ahead forecasting regressions. $Retrf_{i,j}$ ($CAR_{i,j}$) refers to the return (abnormal return) that includes days $t+i, \ldots, t+j$ where t is the event date. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

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**
$\text{CAR}_{0,0}^2$	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 A8

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 return (abnormal return) that includes days $t+i,\ldots,t+j$ where t is the event date. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels. 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 A9

10-day ahead forecasting regressions. $Retrf_{i,j}$ ($CAR_{i,j}$) refers to the return (abnormal return) that includes days $t+i, \ldots, t+j$ where t is the event date. Standard errors are clustered by time. The *, **, and *** indicate significance at the 10%, 5%, and 1% levels. 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:													
	$\text{Retrf}_{1,10}$	$\text{CAR}_{1,10}$	$\text{Retrf}_{1,10}$	$\text{CAR}_{1,10}$	$\text{Retrf}_{1,10}$	$\text{CAR}_{1,10}$	$\text{Retrf}_{1,10}$	$\text{CAR}_{1,10}$	$\text{Retrf}_{1,10}$	$\text{CAR}_{1,10}$	$\text{Retrf}_{1,10}$	$\text{CAR}_{1,10}$	$\text{Retrf}_{1,10}$	$\text{CAR}_{1,10}$
	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
$\text{CAR}_{0,0}$	-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
$\text{CAR}_{-1,-1}$	-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^{**}
$\text{CAR}_{-2,-2}$	-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
$\text{CAR}_{-30,-3}$	-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^{***}
$\text{CAR}^2_{0,0}$	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 A10

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 $\text{Retrf}_{0,0}$ 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 A11

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 A12

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 A13

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 A14

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 A15

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 A16

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 A17

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 A18
This table shows the estimated β_0 from (A5). 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, (A5) is the same for $\text{Retrf}_{i,j}$ and $\text{CAR}_{i,j}$.

News autocorrelation coefficient β_0 from (A5)

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