Why does news coverage predict returns?
Evidence from the underlying editor preferences for risky stocks

G. Schwenkler & H. Zheng*

February 18, 2022†

Abstract

We show that news coverage predicts future stock returns to equal extents because it drives attention flows and because it is a signal of risk. We establish our results by constructing measures of editor preferences for stocks that capture how much more often a stock appears in the news because of its risk characteristics. Editor preferences are highly time-varying and predictive of future returns. Long-short strategies based on our editor preferences achieve annualized alphas of up to 15%. Our paper validates recent theories that posit that editorial reporting points attention-constraint investors to risky assets that earn high future returns.

Keywords: News coverage, attention constraints, risk signals, predictability, natural language processing. JEL codes: G12, G14, G17.

*Schwenkler is at the Department of Finance, Santa Clara University Leavey School of Business. Zheng is at Fidelity Investments. Schwenkler is corresponding author. Email: gschwenkler@scu.edu, website.

†This previous version of this paper circulated under the title "Time-Varying Editor Preferences, Attention Constraints, and Asset Prices". We are grateful to Steven Kou, Evgeny Lyandres, and Andrea Vedolin; and seminar participants at Boston University and Santa Clara University for useful comments and suggestions.

1 Introduction

Attention and information processing are scarce resources in financial markets. Investor do not have the capacity to pay attention to all events that occur over time (Kahneman (1973), Peng and Xiong (2006), and others). It also takes investors time to process information (Peng (2005), Sims (2003), and others). News media plays an important role in guiding investor attention. In financial markets, stocks covered by the news often experience buying pressure and excessive trading by the part of attention-constrained investors; zsee Barber and Odean (2008), Engelberg and Parsons (2011), and Peress (2014), among others. Hillert et al. (2014) and Hillert and Ungeheuer (2021) show that this attention-driven trading behavior results in higher future returns and stronger momentum effects for highly covered firms. Ahern and Sosyura (2015) and Schwenkler and Zheng (2021) show that investors that face information processing constraints often overreact to information reported in the news, resulting in predictable return reversals. In summary, the existing literature suggests that highly covered stocks behave differently than poorly covered stocks because of a behavioral reason: investors with attention and information processing constraints misreact to information reported in financial news.[1]

This consensus was recently challenged by Chahrour et al. (2021) and Nimark and Pitschner (2019). In environments in which agents face attention constraints and news media are for-profit businesses, Chahrour et al. (2021) and Nimark and Pitschner (2019) show that it is an equilibrium outcome that agents delegate their collection of information to news publishers and that news editors choose to report about events that agents care about the most to attract their attention and transform them into paying readers. The most relevant events for risk-averse financial agents are those that represent risks for their investments. The theories of Chahrour et al. (2021) and Nimark and Pitschner (2019) imply that news coverage is predictive of future returns both because it drives investor attention flows, which can lead to misreactions through a behavioral channel, and because it points to risky assets with high expected returns. This latter channel is purely rational. In this paper, we ask: to which extent is news coverage predictive of future stock returns because of attention flow effects, and to which extent is it purely due to being a signal of risk?

To answer this question, we construct an empirical measure of editor preference that captures how much more likely it is that financial news editors choose to report about a stock

[1] A notable exception is Fang and Peress (2009), who show that no-coverage firms post high returns in order to appear more attractive to attention-constraint investors.

because of its risk characteristics. Editor preference is purely a measure of risk. It excludes demand consideration effects that would push financial news editor to report more frequently about larger or information-poor stocks (Mullainathan and Shleifer (2005), Fang and Peress (2009)). Our measure of editor preference differs from existing measures of coverage. Coverage is a measure of how often a stock appears in the news, while editor preference is a measure of how much more often a stock appears in the news because of its risk characteristics. Figure 1 highlights the differences between coverage and editor preference.

Using our measure of editor preference to disentangle the informational content of news coverage, we show that around half of the power of news coverage to predict returns is due to attention flow effects, and the other half is due to risk. Our results indicate that high coverage firms post higher returns than the riskiest firms in the market, consistent with the attention flow channel of Barber and Odean (2008), Engelberg and Parsons (2011),Hillert et al. (2014), and Hillert and Ungeheuer (2021). We also show that low coverage firms post unexpectedly high returns, consistent with the investor recognition channel of Fang and Peress (2009). Our results empirically show that news coverage is predictive of future returns both because of behavioral misreactions and because it is a signal of risk, and that both of these channels contribute almost equally to the predictive power of news coverage.

Extending the exiting literature, we show that editor preference is a more predictive signal of future returns than the news coverage from which we extract it because it aggregates information about pricing-relevant risks in a timely and comprehensive way. We estimate editor preferences using coverage in business news articles published in the New York Times between 1995 and 2015. We find that editor preferences vary significantly over time. For example, we find that financial news editors preferred to report about firms with low cumulative returns going into the financial crisis of 2008. In the few years after the crisis, editors preferred to report about firms with negative exposure to the momentum and size factors, and high idiosyncratic volatilities. We show that the time-variation in the estimated editor preferences is highly predictive of future stock returns.

More precisely, we show that firms with high editor preference in one month earn higher subsequent returns than firms with low preference. This holds both in nominal terms as well as in risk-adjusted returns after accounting for the risk factors of Fama and French (1993), Carhart (1997), and Frazzini and Pedersen (2014). Our findings are robust to controlling for sentiment, which suggests that the predictive power of editor preference is different than the

predictive power of sentiment established by Baker and Wurgler (2006), Garcia (2013), Niessner and So (2018), Tetlock (2007), and others. A trading strategy that exploits the predictive power of editor preferences has an annualized alpha that reaches from 8.3% (when restricted to the largest 500 firms in the economy in any given month) to 15.4% (in the widest cross section of the largest 5,000 firms in any month). We find that no single risk factor explains this anomaly. Instead, we show that the anomaly arises because editor preferences aggregate a wide range of time-varying risk characteristics in a timely way.

The results of this paper validate the theories of Chahrour et al. (2021) and Nimark and Pitschner (2019). They also highlight the predictive power of editorial selection in financial news media and complement recent findings by Bybee et al. (2020), Cong et al. (2019), Kelly et al. (2021), and Larsen et al. (2021) on the predictive power of editorial selection in macroeconomic news media.

This paper is organized as follows. Section 2 introduces the concept of editor preference and how we measure it empirically. Section 3 analyzes the performance of long-short strategies implied by editor preferences. Section 4 studies the informational content of editor preference, while Section 5 dissects the informational content of news coverage in relation to editor preferences. Section 6 concludes. The appendix describes our data collection approach.

2 Measuring editor preferences

We define the monthly editor preference for a firm as the conditional expectation of how often the firm appears in the news of that month given its contemporaneous risk characteristics. We measure editor preference in excess of size and analyst coverage because we know that larger firms appear more often in the news than smaller firms, and that news coverage often serves as a substitute for analyst coverage (see Engelberg and Parsons (2011), Mullainathan and Shleifer (2005), and Solomon and Soltes (2012)). We also measure editor preference in excess of the prior coverage history given that coverage decisions are highly sticky (Schwenkler and Zheng (2019)). Our definition of editor preference differs from news coverage definitions that have previously been used in the literature; see Fang and Peress (2009), Hillert et al. (2014), Hillert and Ungeheuer (2021), and others. Coverage measures how often an asset appeared in the news. In contrast, editor preference measures how much more often a firm appeared in the news because of how risky it is. Coverage can be interpreted as the output of a news selection function $\mathcal{S}$ as in

Nimark and Pitschner (2019) in which the editor preference for certain risks is an input among size, analyst coverage, and prior coverage history:

$$\text{News Coverage} = \mathcal{S}\left(\text{Size}, \text{Analyst Coverage}, \text{Prior Coverage}, \text{Editor Preference(Risks)}\right) \quad (1)$$

Figure 1 highlights the differences between our measure of editor preference and coverage. The intersection of covered and high editor preference stocks represents the most *newsworthy* stocks in the language of Nimark and Pitschner (2019). These are the stocks that investors should pay attention to because they are highly risky.

2.1 Empirical approach

We estimate editor preferences from business news articles published in the New York Times. We identify firm mentions in these news articles using the natural language processing methodology of Schwenkler and Zheng (2019). We then run cross-sectional regressions over rolling three-month horizons of the degree of monthly news coverage on common firm features: size, leverage, past return, idiosyncratic volatility, analyst coverage, analyst forecast dispersion, analyst upgrades and downgrades, factor betas, and an indicator of non-coverage in the prior month. Our empirical measure of editor preference is given by the regression-implied projection of news coverage on all features other than size, analyst coverage, and the prior non-coverage indicator. We focus our analyses on the largest 500 firms each month.

2.1.1 News data

We obtain daily news for the time between January 1, 1995, and December 30, 2015, through The New York Times' API. We only keep articles from the "Business" and "Business Day" sections and filter out any company announcements. We obtain a sample of 140,227 articles in total. Table 1 provides summary statistics of the news articles.

For every article in our data, we apply the firm identification methodology introduced in Schwenkler and Zheng (2019) to recognize any mentions of firms. This methodology has two basic steps. It first uses a natural language processing (NLP) toolkit provided called *coreNLP* to identify and classify any named entities appearing in the articles (see Manning et al. (2014) and Arnold (2017)). For those entities classified as organizations, it follow a series of rules to filter out firms and match them with the Compustat/CRSP database. The approach of Schwenkler and Zheng (2019) has more than 87% matching accuracy.

Table 1 gives some basic statistics of our firm identification results. Notice that one company can be mentioned multiple times within an article. We believe that the number of mentions contains useful information. If two companies are mentioned in an article while the former one shows up more times, then in our analysis it will have a higher mention score than the latter one instead of being equal. This difference separates our paper from other text-based asset pricing studies such as Hillert et al. (2014), Scherbina and Schlusche (2015), and Chahrour et al. (2021) to some extent. Those studies generally rely on firm tags assigned to articles by the data provider that sells the data. The tags refer to the firms that are most frequently mentioned in an article. As a result, there should not be a big difference between our method and a tag-based method to identify firm mentions. However, our approach has a fundamental advantage over a tag-based approach that relies on a commercial data provider: our methodology can be implemented essentially free of costs. The New York Times API is freely accessible, and the codes that implement the methodology of Schwenkler and Zheng (2019) are available here. The only cost associated with implementing our methodology and replicating our results is the time investment to download data and write code.

The total number of mentions for a firm in our data is heavy tailed. To control for this issue, we consider in our analysis a measure of coverage rather than mentions given by:

$$\text{coverage}_{i,t} = \ln(1 + \text{mentions}_{i,t}),$$

where i and t are the firm and time indices.

2.1.2 Firm data

We obtain data from CRSP, Compustat, and I/B/E/S; Appendix A provides details of our data collection approach. We only consider firms that have stocks listed on the NYSE, NASDAQ, or AMEX. Each month we select the 500 largest firms by market capitalization for which we have complete observations.[2] We obtain a sample consisting of 1,391 unique firms. Panel (a) of Table 2 reports summary statistics and Figure 2 displays sector distributions in our sample. We see that our sample overweights the manufacturing and information sectors, and understates the financial sector when compared to the CRSP / Compustat universe over our sample period.

We divide the firms into two groups: a covered group, which encompasses all firms that are mentioned at least once in a month in our sample, and a non-covered group of remaining

[2] In several extensions in Section 4.2, we also show that our results generalize to smaller firms.

firms. Panels (b) and (c) of Table 2 report summary statistics of the covered and non-covered group of firms, and Figure 2 reports their sector distributions. We observe that our firm sample overweights the manufacturing sector and underweights the financial sector relative to the whole CRSP / Compustat sample. There are 710 firms that are never covered in our sample. We observe that covered firms are on average larger than non-covered firms. We also observe that non-covered firms experience fewer analyst recommendation changes (either upgrades or downgrades). Figure 3 shows the proportion of the largest 500 firms in a month that are covered by the news. On average, around 20 to 35% of the firms in our sample fall in the covered group. There is one outlier in February 2015, which appears to be a data issue with the New York Times API.

2.1.3 Empirical measurement of editor preference

The existing literature documents that the news tends to report about large firms and firms low analyst coverage (Engelberg and Parsons (2011), Fang and Peress (2009), Hillert et al. (2014), and Solomon and Soltes (2012)). This reporting bias is likely driven by demand consideration concerns (Mullainathan and Shleifer (2005), Schwenkler and Zheng (2019)): larger firms are popular with investors, and the news serves as a substitute for analyst coverage. The focus of our paper is to understand editor preferences in excess of these demand consideration effects. Because of this, we estimate time-varying editor preferences as follows. We first estimate the drivers of news coverage through monthly cross-sectional regressions. We then estimate editor preferences as the projection of news coverage on several firm features in excess of size and analyst coverage.

We run the following cross-sectional regression for each month t:

$$\begin{aligned}\text{coverage}_{i,s} = \alpha_s^{(t)} + \sum_k \beta_k^{(t)} \times \text{characteristic}_{k,i,s} + \delta^{(t)} \times \text{noncov}_{i,s}^{(t)} \\ + \gamma_S^{(t)} \times \ln\left(\text{size}_{i,s}\right) + \gamma_A^{(t)} \times \ln\left(\#\text{analysts}_{i,s}\right) + \epsilon_{i,s}. \end{aligned} \tag{2}$$

We pool the most recent three months ($s \in \{t-2, t-2, t\}$) to ensure that we have enough data for identification purposes and include month fixed effects ($\alpha_s^{(t)}$). The regressors include market capitalization ($\text{size}_{i,s}$), the number of analysts that cover a firm in a given month ($\#\text{analysts}_{i,s}$) as well as several firm characteristics ($\text{characteristic}_{k,i,s}$) that may affect news coverage: the cumulative return in past three months, the firm's idiosyncratic volatility, leverage ratio, the dispersion of analyst forecasts of earnings-per-share (EPS), the number of analyst upgrades and

downgrades, and firms' factor loadings in the Carhart (1997) four-factor model.[3] To control for the fact that coverage decisions are sticky (see Schwenkler and Zheng (2019)), we also include an indicator that marks whether a firm was not covered in the prior month:

$$\text{noncov}_{i,s}^{(t)} = \mathbf{1}_{\{\text{coverage}_{i,s-1}=0\}}.$$

This control allows us to include in our regressions both covered and non-covered firms (which have zeros on the left-hand side). We standardize all dependent and independent variables other than the prior-month non-coverage indicator on a monthly basis with their cross-sectional means and standard deviations.

Figure 4 displays the t-statistics of the estimates of the coefficients $\gamma_A^{(t)}$, $\gamma_S^{(t)}$, and $\delta^{(t)}$ over time. We observe that the estimates of $\gamma_S^{(t)}$ are always positive and statistically significant, validating the findings of Engelberg and Parsons (2011), Fang and Peress (2009), Hillert et al. (2014), and Solomon and Soltes (2012) that indicate that size is a significant driver of news coverage. In contrast to these aforementioned studies, we generally find an insignificant association between news coverage and analyst coverage; i.e., the estimate of $\gamma_A^{(t)}$ is only sporadically significant. This association is highly time-varying, however, and can switch sings. We also find that the estimate of $\delta^{(t)}$ is always negative and significant, highlighting the stickiness of non-coverage decisions (Schwenkler and Zheng (2019)).

Figure 5 shows the estimates of $\beta_k^{(t)}$ over time. We observe that the estimate are significantly large over subperiods of time. For example, a firm's 3-month cumulative return negatively predicted high news coverage between 2007 and 2008. This suggests that the news focused on reporting about firms with negative stock market performance leading into the financial crisis. During and after the financial crisis, our estimates suggest that the news focused on reporting about firms with negative exposure to the momentum and size factors, high idiosyncratic volatilities, large analyst forecast dispersion, and many earnings analyst downgrades.

Our estimates indicate that different risk characteristics are important drivers of the editorial selection process over time. They indicate that news editors have time-varying preferences on what firms to report about. Motivated by these estimates, we define our measure of *editor*

[3]The four factors are market, size, value, and momentum. We use a 24-month rolling window to estimate each firm's factor loadings and use the residual volatility as our measure of idiosyncratic volatility. We tested longer rolling windows but only found negligible difference.

preference for Firm i in Month t as:

$$\text{EP}_{i,t} = \sum_k \beta_k^{(t)} \times \text{characteristic}_{k,i,t} \tag{3}$$

We intentionally exclude the influence of size, analyst coverage, and prior non-coverage because we want to focus on risk-driven rather than demand-driven editor preferences.

What do editor preferences tell us? To answer this question, consider the time series of editor preference for Citibank and Goldman Sachs displayed in Figure 6. Because the Regression (3) is log-linear, $\exp(\text{EP}_{i,t}) - 1$ roughly measures how much more often Firm i appeared in the news of Month t because of its characteristics than a baseline firm of similar size, analyst coverage, and prior news coverage history. With this interpretation in mind, Figure 6 implies that Citibank appeared up to 1.7-times more often in the news during the financial crisis than a corresponding baseline firm. In contrast, Goldman Sachs only appeared 0.3-times more often in the news during the financial crisis than a baseline firm. Considering that Citibank experienced significant financial distress during the financial crisis and was on the verge of default, while Goldman Sachs weathered the crisis relatively well, Figure 6 validate the interpretation that editor preferences capture risk-based coverage decisions by financial news editors.

3 The predictive power of editor preferences

The theories of Chahrour et al. (2021) and Nimark and Pitschner (2019) suggest that firms with high monthly editor preference are riskier and post higher returns in subsequent months than firms with low editor preference. We test this hypothesis by constructing quintile portfolios sorted by our estimated measures of editor preference each month. Our analyses differ from those in the extant literature because we consider the projection of coverage on risk characteristics when constructing our firm portfolios. The existing literature so far has considered raw coverage (the independent variable of Regression (3); Engelberg and Parsons (2011), Fang and Peress (2009), and others) or residual coverage (the residuals of the Regression (3); Hillert et al. (2014), Hillert and Ungeheuer (2021), and others). Because coverage can be viewed as the output of a news selection function that takes editor preferences for risks as input (see Eq. (1)), we argue that the predictive power of editor preferences may be different than the predictive power of coverage.

3.1 EP-sorted portfolios

We construct equally weighted decile portfolios based on the editor preference scores measured at the end of a month. We then keep track of the returns in the subsequent month of the top and bottom quintile portfolios. Figure 7 shows the cumulative returns of the highest and lowest EP decile portfolios. Consistent with Chahrour et al. (2021) and Nimark and Pitschner (2019), we see that high EP firms tend to post higher returns than low EP firms. Figure 7 suggests that high EP firms carry higher expected returns than low EP firms. They support the hypothesis that editor preferences highlight risky stocks with high risk premia to investors.

We formally analyze the performance of the EP-sorted portfolios by running factor regressions. We consider the 4-factor model of Carhart (1997). We add the betting-against-beta (BAB) factor of Frazzini and Pedersen (2014) because we include several factor betas as features in our definition of editor preference. As a result, our EP measure can be interpreted as a bet against beta. We also add a recession indicator because Figure 7 suggests that the portfolio may underperform during recessions.[4] Table 3 summarizes our findings.

We find that the high-low EP spread portfolio has a statistically significant alpha of 4.9% per year. The alpha holds with significant factor exposure and high adjusted R^2. These observations suggest that our editor preference scores contain information about risks that is not fully reflected in firms' exposure to the market, size, value, momentum, and betting-against-beta factors we consider. We do not find significant exposure to the recession indicator. This suggests that the underperformance of the strategy during recessionary periods highlighted in Figure 7 may be driven by an underperformance of risk factors during recessions.

The theories of Chahrour et al. (2021) and Nimark and Pitschner (2019) suggest that covered firms with high editor preference scores are the most risky and attention-grabbing stocks (see Figure 1 as well). In contrast, non-covered stocks with low editor preference scores are the least risky and least attention-grabbing stocks. A strategy that goes long on covered stocks with high editor preferences and short on non-covered stocks with low editor preferences should therefore post the highest possible alpha.

We assess this conjecture visually in Figure 8, in which we compare the cumulative returns of the covered high EP minus non-covered low EP strategy to that of the EP-spread strategy that neglects coverage effects (the high EP minus low EP strategy implied by Figure 7) and a coverage strategy that neglects editor preference effects (the high coverage minus low coverage strategy

[4] We provide an explanation of why this occurs in Section 4.1.

implied by Hillert et al. (2014) and Hillert and Ungeheuer (2021)). We find that the covered high EP minus non-covered low EP strategy outperform these alternative strategies that bet on the predictive power of coverage or editor preferences alone. Table 3 confirms these observations by running factor regressions. We find that the covered high EP minus non-covered low EP strategy has an annualized alpha of 8.3%. The alpha of this strategy is higher than the alpha of a strategy that bets only on predictability due to editor preferences (4.9%) or predictability due to coverage (3.9%). The high minus low coverage strategy loads to a weaker degree on the risk factors and also has a lower adjusted R^2. The high minus low coverage strategy also has a higher Sharpe ratio than the two editor preference strategies, suggesting that the coverage strategy is less risky than the EP-based strategies. All in one, these results of Table 3 indicate that coverage alone may not be as informative about risks as is the editor preference score that goes into the coverage selection function (see Eq. 1).

The performance of the covered high EP minus non-covered low EP strategy is economically significant given that its alpha represents more than 1.5-times the equity risk premium on the U.S. market after controlling for common risk factors. Going forward, we will call this strategy the *EP strategy*.

3.2 Sentiment effects

We evaluate the predictive power of editor preferences after accounting for sentiment. We begin by repeating the analysis of Table 3 after controlling for the investor sentiment index of Baker and Wurgler (2006). Table 4 summarizes our results. We find that the alphas of the editor preference strategies are slightly lower when we also control for sentiment. This suggest that some of the predictive power of editor preference may be due to investor sentiment effects. However, the alphas remain significant even after controlling for sentiment. The results of Table 4 show that our editor preference scores are predictive of future returns even after controlling for investor sentiment.

Garcia (2013), Niessner and So (2018), Tetlock (2007), and others show that the linguistic sentiment of the news from which we extract our editor preference scores may drive some of its predictive power. To assess to which degree this is the case, we repeat the experiment of Table 3 when measuring coverage from sentences with varying degrees of sentiment.

More precisely, we use the *coreNLP* package in R to evaluate the sentiment score of each sentence in which we identify a stock. This package assigns to each sentence an integer score

that measures sentiment ranging from 0 (very negative sentiment) to 4 (very positive sentiment). To understand how linguistic sentiment impacts the performance of our EP strategy, we measure coverage by counting the number of times a firm in mentioned in sentences with different sentiment scores and then estimate editor preferences as in Section 2.1.3 based on these sentiment-adjusted coverage measures. There are a total of 1,343,203 sentences in our news data that mention at least one firm. Out of these sentences, 5.3% have a sentiment score of 0, 83% have a sentiment score of 1, 5% have a sentiment score of 2, 6.5% have a sentiment score of 3, and 0.2% have a sentiment score of 4. The distribution of sentiment in our news sample suggests that editors tend to report about firms in negative tones, consistent with recent findings by García (2018) and Niessner and So (2018).

Table 5 repeats the experiment of Table 3 using editor preferences extracted from sentences with varying degrees of sentiment. We observe monotonically decreasing alphas as we move from very negative sentiment (sentiment score = 0) to positive sentiment (sentiment score = 3). Such decreasing predictive power as we move from negative to positive sentiment is consistent with Garcia (2013) and Tetlock (2007), who show that negative linguistic sentiment in financial news is more predictive of future stock returns than positive linguistic sentiment. The alpha of the EP strategy restricted to coverage extracted from sentences with very positive sentiment (sentiment score = 4) appears to be a small sample outlier given that only 0.2% of the sentences in our news data have very positive sentiment.[5] Table 5 documents statistically significant and positive alphas for the EP strategy restricted to all but one sentiment score (sentiment score = 3). These results suggest that the predictive power of editor preferences is not driven by linguistic sentiment effects.

4 What do editor preferences capture?

We evaluate the informational content of editor preferences. We begin by analyzing whether our measure of editor preference provides additional information beyond what is captured in the characteristics that go into its construction (see Eq. (3)). We construct quintile portfolios sorted by each of the 10 characteristics we use to compute editor preferences, and go long on the stocks in the top characteristic quintile and short on the stocks in the bottom characteristic quintile. We then regress the returns of these characteristic-based spread portfolios on the risk factors of

[5]The R^2 for this strategy is also lower.

Table 3. Table 6 reports the estimated alphas and Sharpe ratios.

We observe that none of the characteristic-based spread portfolios achieve statistically significant positive alphas.[6] The Sharpe ratios of the characteristic-based spread strategies are also much lower than those of the editor-preference-based strategies in Table 3. These results indicate that editor preferences contain information about future returns that exceeds the information contained in the individual characteristics that go into the construction of the signal.

4.1 Timeliness

We evaluate why the editor preference signal is more informative about future returns than its individual characteristic components. We envision that there are two reasons. First, editor preference is a comprehensive signal of an array of risk characteristics, while the individual characteristics only reflect one risk characteristic. Second, editor preferences weight the different risk characteristics based on their relevance, and these weights (i.e.; the loadings $\beta_k^{(t)}$ in Eq. (3)) change dynamically over time as highlighted in Figure 5. In constrast, the individual characteristics do not provide any information about how relevant the characteristic is for the overall riskiness of a firm. Based on this intuition, we evaluate the timeliness and comprehensiveness of our editor preference signals.

We begin by evaluating the timeliness of the editor preference signal. We construct a static measure of editor preference using the time series average of the $\beta_k^{(t)}$ estimates of Figure 5:

$$\text{SEP}_{i,t} = \sum_k \bar{\beta}_k \times \text{characteristic}_{k,i,t}. \tag{4}$$

We then construct sorted quintile portfolios based on this static editor preference measure (SEP) and run factor model regressions as in Section 3. Table 7 summarizes our findings.[7] We find that none of the static EP spread portfolios post statistically significant alphas. These results indicate that the time-varying weighting in the editor preference signal is informative on its own.

Next, we evaluate the persistence of the editor preference signal. For this, we consider holding the long and short positions implied by our editor preference measures for several months, and evaluate the performance of the implied portfolios. Table 8 summarizes our findings.

[6] Only the analyst-dispersion-based spread portfolio achieves a significant alpha, but the alpha is negative.

[7] Note that our approach is slightly biased because the $\bar{\beta}_k$ loadings in Eq. (4) contain forward-looking information given that these are averages across the whole sample period. However, this bias should lead to higher estimated alphas, suggesting that our findings may be overly optimistic.

Table 8 shows that the EP-related alphas die out quickly if the positions are held open for more than one month. For instance, if we keep the positions open for one month, then the annualized alpha of the EP strategy is 8.26% as indicated in Table 3. However, if we hold the position open for 3 months, then the annualized alpha of the EP strategy decreases to 4.39%. The annualized alpha of the EP strategy is only 3.44% if we keep the positions open for 12 months. These results suggest that editor preferences capture information about short-lived risks rather than long-lived risks.

4.2 Comprehensiveness

We now evaluate the comprehensive nature of the editor preference signal. We study how strongly our editor preference signal forecasts individual firm returns when controlling for firm characteristics. We develop a test similar to that of Fama and French (2020). We run the following pooled regression:

$$R_{i,t} - R_{M,t} = a_S \times \ln(size_{i,t-1}) + a_A \times \ln\left(\#\text{analysts}_{i,t-1}\right) + \sum_k b_k \times \text{characteristic}_{k,i,t-1} + c \times \text{EP}_{i,t-1} + e_{i,t}, \quad (5)$$

where $\text{characteristic}_{k,i,t-1}$ are the same as in Regression (3), $R_{i,t}$ is the monthly return of Firm i, and $R_{M,t}$ is the average return of largest 500 firms in Month t.[8] Because there is co-linearity between EP and $\text{characteristic}_{k,i,t-1}$, we use an LASSO regularization approach which pushes the coefficients of variables that have limited explanatory power toward zero. We again standardize all variables before running the regression.

Figure 9 displays the estimated coefficients for Regression (5) for different levels of the L_1-norm constrain imposed by the LASSO approach. We observe that EP is among the most significant forecasters of the return of a firm in the following month. The relationship between EP and future firm returns is positive, suggesting that EP is a measure of firm-level risk. We find that EP remains a significant predictor of firm-level returns, together with a firm's leverage and analyst forecast dispersion, even as the L_1-norm of the coefficient vector becomes more constrained by the LASSO approach and more irrelevant variables are filtered out. These results suggest that EP provides more information about future firm-level returns than most of

[8] Notice that we do not use risk-free rate or other orthogonal portfolio returns. This is because the focus of this test is not to estimate factor returns, but to assess the forecasting power of EP. Also, we do not include an intercept because the average of the independent variable in Eq. (5) is zero.

the individual risk characteristics that go into its construction, highlighting the comprehensive nature of our editor preference signal.

Figure 3 shows that the news covers only up to 35% of the firms in our sample. This means that many firms that are in our sample are not mentioned in the news and we are still able to generate predictive editor preference signals for these non-covered firms as highlighted by Figure 7. This observation suggests that our editor preference signal may be informative about the future returns of another set of firms that is likely never covered in the news: the smallest firms in the economy. To assess this hypothesis, we estimate the editor preference model in Eq. (3) using data for the largest 500 firms in a month and then extrapolate the model to compute editor preferences for firms that rank below the largest 500 firms by market capitalization. After this, we construct long-short portfolios by EP scores for wider cross sections than in Section 3 and analyze the performance of the resulting portfolios in factor model regressions.

Table 9 reports our findings. We find that the EP-implied spread portfolio alphas and Sharpe ratios are mostly increasing in the size of the cross section. These results suggest that our editor preference scores are informative about the future performance of a wide cross section of firms. They further highlight the comprehensive nature of our editor preference signal. In addition, we find that the alpha and the Sharpe ratio of the coverage spread portfolio is decreasing in the size of the cross section. These observations suggest that raw coverage may not be as predictive of a signal of future returns in the wide cross section of firms.

Figure 8 suggests that the EP strategy underperforms during recessions.[9] While the regressions of Table 3 do not confirm a negative relationship between the performance of the EP strategy and the recession indicator, we argue that there may still be a reason to expect lower performance during recessionary periods: editor preferences may not accurately separate risky from non-risky firms during recessions when all firms face elevated risks. If this were the case, then editors preference may not clearly identify risky firms during recessions. To assess whether this is the case, we repeat the regressions of Table 3 for long-short strategies adjusted as follows: whenever we observe two consecutive negative market returns, we do not enter any long or short positions in the next month. This occurs 37 times during our sample.

Table 10 reports the estimates of our factor model regressions. We achieve better performance when we avoid trading when the market declines over two consecutive months. The annualized alpha of the EP strategy is 9.65% with the trading exclusion rule and 8.26% without

[9] There are two NBER recessions in our data: the dot-com crash (2000-2002) and the financial crisis (2007-2009).

the trading exclusion rule. The Sharpe ratio is 0.43 with the trading exclusion rule and 0.31 without the trading exclusion rule. The results of Table 10 indicate that performance of an editor-preference implied strategy is worse during times when the market underperforms and all firms experience financial distress. They highlight our editor preference measure as a signal that disentangles risky from non-risky firms.

5 Coverage: is it risk or attention flow effects?

We evaluate the informational content of coverage in relation to editor preferences. We begin by decomposing the coverage spread return as follows:

$$\begin{aligned}\underbrace{\text{High Cov.} - \text{Low Cov.}}_{\text{coverage spread}} = &\underbrace{(\text{High Cov.} - \text{High EP})}_{\text{high coverage residual}} \\ &+ \underbrace{(\text{High EP} - \text{Low EP})}_{\text{EP spread}} \\ &+ \underbrace{(\text{Low EP} - \text{Low Cov.})}_{\text{low coverage residual}}.\end{aligned}$$

Here, the high and low portfolios are quintile portfolios sorted by either coverage ("Cov.") or editor preference ("EP"); see Section 3. Figure 10 visually represents this decomposition in the time series over our sample.

We observe that the EP spread makes up a large fraction of the coverage spread. Our estimates suggest that the EP spread makes up 148% of the coverage spread on average, while the high coverage residual only makes up 9% and the low coverage residual makes up -56% of the coverage spread. Our results suggests that the coverage spread is driven to a large part by risk. They suggest that high coverage stocks post slightly higher returns than the riskiest stocks in the market, likely due to behavioral attention flow effects as in Barber and Odean (2008), Engelberg and Parsons (2011), Peress (2014), Hillert et al. (2014), and Hillert and Ungeheuer (2021). Our results also suggest that the coverage spread strategy performs worse than the EP spread strategy primarily because low coverage stocks post relatively high returns. This observation is consistent with Fang and Peress (2009), who document that no-coverage stocks trade at high expected returns in order to appear more attractive to attention-constraint investors.[10]

[10] One concern in the analysis of Figure 10 is that we know that news coverage is biased towards larger stocks. In unreported experiments, we repeat the analysis of Figure 10 for residual coverage rather than raw coverage

We push this analysis further by running factor regressions similar to those of Section 3.1 for coverage spread strategy when excluding firms with high or low EP. A coverage strategy that is restricted to the subset of firms that do not fall in the top or bottom quintiles by editor preferences loads on firms with average risk levels. As a result, the performance of such a strategy is primarily driven by attention flow effects. We compare the alphas of the coverage strategy when including or excluding the extreme EP firms in order to measure how strongly attention flow effects contribute to the overall performance of the coverage strategy. We consider both a raw coverage strategy as in Figure 10, as well as a residual coverage strategy that goes long (short) on high (low) residual coverage firms. Here, we define residual coverage as the residual after regressing normalized coverage on normalized firm size (see Section 2.1.3). This strategy accounts for the fact that we know that news reporting is biased towards larger firms (Mullainathan and Shleifer (2005), Solomon and Soltes (2012), and others).

Table 11 summarizes our findings. We find that the alpha of the coverage spread strategy falls by 68% after removing the extreme EP firms each month from the sample of firms. The alpha of the residual coverage strategy falls by 46% after we remove the extreme EP firms from the monthly samples. These results suggest that behavioral attention flow effects contribute around 50% of the total alpha of a coverage strategy.

All in one, the results of this section show that attention flow effects drive around half of the predictive power of news coverage. They suggest risk is a major contributor to the predictive power of news coverage.

6 Conclusion

Using articles from the New York Times over 20 years, we show that editors have time-varying preferences for different types of firms. We construct an editor preference score to measure how much more frequently a firm is chosen to appear in the news because of its risk characteristics. Our editor preference scores provide timely and comprehensive signals of firm-level risks that are not captured by individual firm characteristics. We show that firms with high editor preference earn higher future returns than firms with low editor preference, and construct long-short portfolios with economically and statistically significant alphas that cannot be explained by

and find similar results. The residual coverage spread can be decomposed as follows: EP spread: 128%. High residual coverage residual: 9%. Low residual coverage residual: -36%. For us, residual coverage is the residual after regressing normalized coverage on normalized firm size as in Section 2.1.3.

standard risk factors. We use our measure of editor preferences to disentangle the informational content of news coverage. We show that risk and attention flow effects are equal contributors to the predictive power of news coverage that has been documented in the existing literature.

Our findings validate recent theories that argue that financial news provide a larger service to their audience than just reporting about current events. The motivation behind the editorial choice to cover some firms over others is also of importance. In an equilibrium between news editors and attention-constrained investors who delegate their information collection to news outlets, editors help investors by monitoring and reporting firms that face elevated risks. Our paper empirically validates precisely this mechanism.

To obtain our results, we use a natural language processing toolkit to identify companies from news. Our research lies among a new genre of machine learning applications in asset pricing that considers alternative data to gain insights about future economic outcomes.

A Data

We obtain monthly pricing data from CRSP. For each stock series, we take total returns including dividends and cash payouts (CRSP item "trt1m"). We compute market capitalizations for a series as the product of the closing price in a month (CRSP item "prccm") and the number of common shares outstanding from the previous quarter (CRSP item "cshoq"). We then computer a total market capitalization of a firm as the sum of the market capitalizations of each series, and the total monthly return of a firm as the market-cap-weighted average of the returns of each series. In any regression in which residuals are assumed to be Gaussian, we take log returns computed as the logarithm of one plus the total return of a firm. We compute the equity volatility of a firm as the standard deviation of log returns over rolling 60-month windows.

We obtain Fama-French factor data from Ken French's website. We obtain Betting-Against-Beta factor data from AQR's website. For each firm, we compute factor loadings as a regression of excess returns on factor returns over rolling 60-month windows. We compute idiosyncratic volatilities as the standard deviation of the residuals of these rolling factor model regressions.

We obtain fundamentals data from the CRSP / Compustat merged database and match all firms through their GVKEY. All data are taken from quarterly reports. We take the sector of a firm to be identified by the first two digits of the associated NAICS code. In Figure 2, we cluster sectors as follow for ease of exposition:

- Construction: NAICS code 23.
- Finance, insurance, & real estate: NAICS codes 52 and 53.
- Health Care: NAICS code 62.
- Information: NAICS code 51.
- Manufacturing: NAICS codes 31–33.
- Natural resources: NAICS codes 11 and 21.
- Services: NAICS codes 54, 55, 56, 61, 71, and 72.
- Transportation: NAICS codes 48–49.
- Utilities: NAICS code 22.
- Wholesale & retail trade: NAICS codes 42, 44, and 45.
- Other: All remaining NAICS codes.

We define book equity as the sum of common equity (Compustat item "CEQQ") and deferred taxes ("TXDITCQ"). Book debt is the difference between total assets ("ATQ") and book equity. We define leverage as the ratio of book debt over total assets. We only include firms in our sample if the exchange code on Compustat ("EXCHGCD") is 11, 12, or 14.

We collect earnings analyst data from I/B/E/S. Each month, we match firms by their CUSIP codes and, when unavailable, by their tickers. For each firm, we collect all EPS analyst forecasts for the current and next three fiscal quarters. We compute the number of analysts that track a firm in a month as the average number of earnings analysts that reported forecasts for each of the considered fiscal quarters in one month. Analyst coverage is defined as the logarithm of 1 plus the number of analyst tracking the firm in a month. Forecast dispersion is defined as the logarithm of one plus the I/B/E/S-reported standard deviation of EPS forecasts in one month. We compute the number of earnings analysts upgrades (downgrades) as the average number of forecast upgrades (downgrades) across any of the considered quarters. For all data points, we assign a value of zero when unavailable.

References

Ahern, Kenneth R. and Denis Sosyura (2015), 'Rumor has it: Sensationalism in financial media', *The Review of Financial Studies* **28**(7), 2050–2093.

Arnold, Taylor (2017), 'A tidy data model for natural language processing using cleannlp', *The R Journal* **9**(2), 1–20.

Baker, Malcolm and Jeffrey Wurgler (2006), 'Investor sentiment and the cross-section of stock returns', *The Journal of Finance* **61**(4), 1645–1680.

Barber, Brad M. and Terrance Odean (2008), 'All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors', *The Review of Financial Studies* **21**(2), 785–818.

Bybee, Leland, Bryan T. Kelly, Asaf Manela and Dacheng Xiu (2020), The structure of economic news, Technical report, National Bureau of Economic Research.

Carhart, Mark M. (1997), 'On persistence in mutual fund performance', *The Journal of Finance* **52**(1), 57–82.

Chahrour, Ryan, Kristoffer Nimark and Stefan Pitschner (2021), 'Sectoral media focus and aggregate fluctuations', *American Economic Review* **111**(12).

Cong, Lin William, Tengyuan Liang and Xiao Zhang (2019), Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. Working Paper.

Engelberg, Joseph E. and Christopher A. Parsons (2011), 'The causal impact of media in financial markets', *The Journal of Finance* **66**(1), 67–97.

Fama, Eugene F and Kenneth R French (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* .

Fama, Eugene F. and Kenneth R. French (2020), 'Comparing cross-section and time-series factor models', *The Review of Financial Studies* **33**(5), 1891–1926.

Fang, Lily and Joel Peress (2009), 'Media coverage and the cross-section of stock returns', *The Journal of Finance* **64**(5), 2023–2052.

Frazzini, Andrea and Lasse Heje Pedersen (2014), 'Betting against beta', *Journal of Financial Economics* **111**(1), 1–25.

Garcia, Diego (2013), 'Sentiment during recessions', *The Journal of Finance* **68**(3), 1267–1300.

García, Diego (2018), The kinks of financial journalism. Working Paper, CU Boulder.

Hillert, Alexander, Heiko Jacobs and Sebastian Müller (2014), 'Media makes momentum', *The Review of Financial Studies* **27**(12), 3467–3501.

Hillert, Alexander and Michael Ungeheuer (2021), The value of visibility. Working paper.

Kahneman, D. (1973), *Attention and Effort*, Prentice-Hall series in experimental psychology, Prentice-Hall.

Kelly, Bryan, Asaf Manela and Alan Moreira (2021), 'Text selection', *Journal of Business & Economic Statistics* **39**(4), 859–879.

Larsen, Vegard H., Leif Anders Thorsrud and Julia Zhulanova (2021), 'News-driven inflation expectations and information rigidities', *Journal of Monetary Economics* **117**, 507–520.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard and David McClosky (2014), The stanford corenlp natural language processing toolkit, *in* 'Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations', pp. 55–60.

Mullainathan, S. and A. Shleifer (2005), 'The market for news', *American Economic Review* **95**(1), 1031–1053.

Newey, Whitney K. and Kenneth D. West (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation', *Econometrica* **55**(3), 703–708.

Niessner, Marina and Eric C. So (2018), Bad news bearers: The negative tilt of financial press. Working Paper.

Nimark, Kristoffer P. and Stefan Pitschner (2019), 'News media and delegated information choice', *Journal of Economic Theory* **181**, 160–196.

Peng, Lin (2005), 'Learning with information capacity constraints', *Journal of Financial and Quantitative Analysis* **40**(2), 307–329.

Peng, Lin and Wei Xiong (2006), 'Investor attention, overconfidence and category learning', *Journal of Financial Economics* **80**(3), 563–602.

Peress, Joel (2014), 'The media and the diffusion of information in financial markets: Evidence from newspaper strikes', *The Journal of Finance* **69**(5), 2007–2043.

Scherbina, Anna and Bernd Schlusche (2015), Economic linkages inferred from news stories and the predictability of stock returns. Working Paper.

Schwenkler, Gustavo and Hannan Zheng (2019), The network of firms implied by the news. Working Paper.

Schwenkler, Gustavo and Hannan Zheng (2021), News-driven peer co-movement in crypto markets. Working Paper.

Sims, Christopher A. (2003), 'Implications of rational inattention', *Journal of Monetary Economics* **50**(3), 665–690.

Solomon, David H. and Eugene F. Soltes (2012), Managerial control of business press coverage. Working Paper.

Tetlock, Paul C. (2007), 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of Finance* **62**(3), 1139–1168.

Variables	Mean	Std Dev.	Min.	Max.
Number of articles per month	558	139	327	980
Number of articles per year	6677	1417	4782	9845
Number of unique firms per article	3	2	1	62
Number of mentions per article	7	8	1	167
Total news mentions of a firm	344	1510	5	33440

Table 1: Summary statistics of news articles in our news data. All numbers are rounded to be integers. We download 140,227 news articles between January 1, 1995, and December 30, 2015, from the "Business" and "Business Day" sections of The New York Times. In the last row, we only consider firms that are mentioned at least once in our data.

Panel (a): All firms					
Variables	Mean	Median	Std Dev.	Min.	Max.
Monthly mentions	3	0	13	0	647
Monthly market capitalization (million USD)	26,512	11,968	43,798	1,327	1,201,845
Analyst coverage	1.75	2.08	1.11	0.00	3.87
Forecast dispersion	0.04	0.02	0.14	0.00	5.53
Forecast upgrades	1.20	0.25	2.43	0.00	37.50
Forecast downgrades	1.40	0.25	2.71	0.00	40.75
Leverage	0.56	0.55	0.21	0.00	1.29
Monthly return	1.58%	1.39%	10.23%	-83.48%	640.74%
3-month cumulative return	4.77%	4.06%	18.77%	-92.18%	1,328.57%
Monthly idiosyncratic volatility	7.46%	6.66%	3.31%	0.21%	52.11%
Market factor beta	1.06	0.99	0.55	-0.89	5.14
Size factor (SMB) beta	0.09	0.05	0.58	-4.72	4.36
Value factor (HML) beta	0.08	0.12	0.78	-4.99	6.53
Momentum factor (MOM) beta	-0.06	-0.04	0.40	-3.40	2.71
Prior month non-coverage indicator	0.75	1.00	0.43	0.00	1.00
Number of firms			1,391 (100%)		
Panel (b): Covered firms					
Variables	Mean	Median	Std Dev.	Min.	Max.
Monthly mentions	4	0	15	0	647
Monthly market capitalization (million USD)	32,272	14,853	49,783	1,337	1,201,845
Analyst coverage	1.83	2.17	1.09	0.00	3.87
Forecast dispersion	0.04	0.02	0.16	0.00	5.53
Forecast upgrades	1.29	0.25	2.52	0.00	37.50
Forecast downgrades	1.51	0.25	2.83	0.00	40.75
Leverage	0.57	0.57	0.20	0.03	1.29
Monthly return	1.46%	1.36%	9.77%	-83.48%	262.66%
3-month cumulative return	4.44%	3.94%	17.52%	-92.18%	850.22%
Monthly idiosyncratic volatility	7.29%	6.50%	3.23%	0.67%	31.14%
Market factor beta	1.04	0.99	0.54	-0.89	5.14
Size factor (SMB) beta	0.05	0.02	0.58	-4.72	4.20
Value factor (HML) beta	0.10	0.14	0.76	-4.99	5.54
Momentum factor (MOM) beta	-0.06	-0.04	0.40	-3.26	2.71
Prior month non-coverage indicator	0.65	1.00	0.48	0.00	1.00
Number of firms			681 (48.96%)		
Panel (b): Non-covered firms					
Variables	Mean	Median	Std Dev.	Min.	Max.
Monthly mentions	0	0	0	0	0
Monthly market capitalization (million USD)	12,276	7,906	15,790	1,327	310,734
Analyst coverage	1.56	1.85	1.12	0.00	3.67
Forecast dispersion	0.03	0.01	0.08	0.00	4.06
Forecast upgrades	0.99	0.00	2.16	0.00	30.75
Forecast downgrades	1.13	0.00	2.36	0.00	31.00
Leverage	0.51	0.51	0.21	0.00	1.29
Net profit margin	0.09	0.09	0.31	-12.19	0.75
Monthly return	1.87%	1.45%	11.27%	-58.51%	640.74%
3-month cumulative return	5.60%	4.31%	21.53%	-77.51%	1,328.57%
Monthly idiosyncratic volatility	7.88%	7.08%	3.47%	0.21%	52.11%
Market factor beta	1.09	1.00	0.58	-0.66	5.05
Size factor (SMB) beta	0.18	0.13	0.57	-2.53	4.36
Value factor (HML) beta	0.03	0.08	0.83	-4.76	6.53
Momentum factor (MOM) beta	-0.06	-0.04	0.41	-3.40	2.07
Prior month non-coverage indicator	1.00	1.00	0.02	0.00	1.00
Number of firms			710 (51.04%)		

Table 2: Summary statistics of our data variables. The statistics are sample moments during our sample period, which begin in January 1995 and ends in December 2015. Panel (a) reports summary statistics for firms from the Compustat/CRSP database that at least once are counted as one of 500 largest firms by monthly market capitalization. Panel (b) restricts the sample summary statistics to those firms from Panel (a) that were at least mentioned once in the news. Panel (c) restricts the sample to all remaining firms that were never covered by the news. Appendix A defines each of the variables.

	High EP - Low EP	Covered High EP - Non-covered Low EP	High coverage - Low coverage
Mkt-RF	** 0.2751	** 0.2751	0.0357
	(2.8615)	(2.6537)	(0.8791)
SMB	−0.1473	** −0.2704	** −0.1415
	(−1.2869)	(−2.8358)	(−3.0883)
HML	*** 0.4805	*** 0.4115	0.0472
	(4.6569)	(4.1776)	(0.8248)
MOM	*** −0.3094	*** −0.3559	* −0.0593
	(−3.5960)	(−4.6342)	(−2.1687)
BAB	−0.2060	* −0.2402	* −0.0957
	(−1.8651)	(−2.0816)	(−2.4808)
Recession	−0.0067	−0.0125	−0.0027
	(−1.0606)	(−1.8193)	(−0.9611)
Alpha	* 0.0041	** 0.0069	** 0.0033
	(2.0680)	(2.6959)	(2.8065)
Observations	251	251	251
Adjusted R^2	0.524	0.507	0.170
Annualized Sharpe ratio	0.21	0.31	0.43

Table 3: *Factor model regressions.* The dependent variables are the returns of different high minus low quintile portfolios. Independent variables are the three factors of Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor of Carhart (1997) (Mom), the betting-against-beta factor of Frazzini and Pedersen (2014) (BAB), and an NBER recession indicator. We take the risk-free rate to be the one in the Fama-French data. Standard errors are adjusted for serial autocorrelation as in Newey and West (1987) with a lag of 12 months. The parentheses report t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	High EP - Low EP	Covered High EP - Non-covered Low EP	High coverage - Low coverage
Mkt-RF	** 0.2782	** 0.2787	0.0354
	(2.8971)	(2.7239)	(0.8587)
SMB	−0.1522	** −0.2761	** −0.1409
	(−1.3454)	(−3.0001)	(−3.0168)
HML	*** 0.4497	*** 0.3754	0.0509
	(4.2974)	(3.6955)	(0.8531)
MOM	*** −0.3207	*** −0.3692	* −0.0580
	(−3.8244)	(−4.9862)	(−2.1922)
BAB	· −0.2112	* −0.2463	* −0.0951
	(−1.8585)	(−2.0779)	(−2.5018)
Recession	−0.0099	* −0.0163	−0.0023
	(−1.2318)	(−2.1309)	(−0.8152)
Investor Sentiment	* 0.0061	** 0.0071	−0.0007
	(2.1729)	(2.7101)	(−0.6330)
Alpha	· 0.0038	** 0.0066	** 0.0033
	(1.8735)	(2.6037)	(2.8041)
Observations	251	251	251
Adjusted R^2	0.531	0.514	0.168

Table 4: *Factor model regressions accounting for investor sentiment.* The dependent variables are the returns of different high minus low quintile portfolios. Independent variables are the three factors of Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor of Carhart (1997) (Mom), the betting-against-beta factor of Frazzini and Pedersen (2014) (BAB), a NBER recession indicator, and the sentiment index of Baker and Wurgler (2006). We take the risk-free rate to be the one in the Fama-French data. Standard errors are adjusted for serial autocorrelation as in Newey and West (1987) with a lag of 12 months. The parentheses report t-statistics. ***, **, *, and · denote significance on the 99.9%, 99%, 95%, and 90% confidence levels, respectively.

	Sentiment score = 0 (Very negative sentiment)		Sentiment score = 1 (Negative sentiment)		Sentiment score = 2 (Neutral sentiment)		Sentiment score = 3 (Positive sentiment)		Sentiment score = 4 (Very positive sentiment)	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Mkt-RF	** 0.2593	** 0.2637	** 0.2769	** 0.2793	*** 0.2773	*** 0.2794	*** 0.3101	*** 0.3129	0.0982	0.1021
	(2.7810)	(2.9998)	(2.7400)	(2.7826)	(4.1318)	(4.1327)	(6.2642)	(6.6187)	(1.3008)	(1.4389)
SMB	−0.1610	· −0.1677	** −0.2980	** −0.3017	· −0.1458	* −0.1490	*** −0.2470	*** −0.2512	−0.0801	−0.0860
	(−1.6492)	(−1.8516)	(−3.1192)	(−3.2117)	(−1.9158)	(−1.9924)	(−3.4672)	(−3.6476)	(−1.0382)	(−1.1836)
HML	** 0.3247	* 0.2794	*** 0.4049	*** 0.3804	*** 0.4114	*** 0.3897	*** 0.4209	*** 0.3926	** 0.2808	** 0.2412
	(2.8006)	(2.4541)	(4.4571)	(3.9860)	(5.9469)	(5.4034)	(5.8421)	(5.4438)	(3.2087)	(2.8400)
MOM	*** −0.3053	*** −0.3219	*** −0.3612	*** −0.3702	*** −0.3542	*** −0.3622	*** −0.3496	*** −0.3600	*** −0.2170	*** −0.2315
	(−4.9587)	(−5.5671)	(−5.1921)	(−5.4514)	(−7.3034)	(−7.8372)	(−9.5531)	(−10.3374)	(−3.3689)	(−3.9345)
BAB	** −0.2847	** −0.2924	* −0.2456	* −0.2498	** −0.2681	** −0.2718	· −0.1575	· −0.1623	* −0.1801	* −0.1869
	(−3.1037)	(−3.0934)	(−2.1396)	(−2.1374)	(−3.2488)	(−3.2158)	(−1.8940)	(−1.9179)	(−2.3925)	(−2.4959)
Recession	* −0.0152	* −0.0198	· −0.0119	· −0.0144	· −0.0091	* −0.0113	−0.0021	−0.0050	−0.0123	* −0.0163
	(−2.4047)	(−2.1889)	(−1.8531)	(−1.8997)	(−1.7109)	(−2.2215)	(−0.3558)	(−1.1139)	(−1.4116)	(−2.0724)
Investor Sentiment		** 0.0090		· 0.0048		* 0.0043		* 0.0056		* 0.0078
		(2.9890)		(1.9425)		(2.2175)		(2.3126)		(2.2195)
Alpha	** 0.0074	** 0.0071	* 0.0060	* 0.0059	* 0.0053	* 0.0051	0.0028	0.0026	** 0.0087	** 0.0084
	(2.6416)	(2.6399)	(2.4480)	(2.3537)	(2.0861)	(2.0320)	(1.3490)	(1.2138)	(2.7748)	(3.1009)
Observations	251	251	251	251	251	251	251	251	251	251
Adjusted R^2	0.429	0.442	0.527	0.529	0.526	0.528	0.526	0.530	0.221	0.234

Table 5: *Factor model regressions accounting for linguistic sentiment.* The dependent variables are the returns of the EP strategy when editor preferences are estimated from firm mentions in sentences with varying degrees of linguistic sentiment. Independent variables are the three factors of Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor of Carhart (1997) (Mom), the betting-against-beta factor of Frazzini and Pedersen (2014) (BAB), a NBER recession indicator, and the sentiment index of Baker and Wurgler (2006). We take the risk-free rate to be the one in the Fama-French data. Standard errors are adjusted for serial autocorrelation as in Newey and West (1987) with a lag of 12 months. The parentheses report t-statistics. ***, **, *, and · denote significance on the 99.9%, 99%, 95%, and 90% confidence levels, respectively.

Characteristic	Alpha	Sharpe ratio
Cumulative return	−0.0020	−0.17
Idiosyncratic volatility	−0.0006	−0.08
Leverage	−0.0020	0.07
Analyst forecast dispersion	* −0.0021	−0.47
Analyst upgrades	0.0003	0.11
Analyst downgrades	0.0011	0.04
Market factor beta	−0.0019	−0.03
Size factor beta	0.0000	0.00
Value factor beta	−0.0054	0.00
Momentum factor beta	−0.0015	0.04

Table 6: *Characteristic-based spread portfolios.* We construct quintile portfolios sorted by the individual standardized characteristics that go into the construction of our editor preference measure; see Section 2.1.3. For each characteristic and any given month, we go long on the firms in the top characteristic quintile that are covered by the news, and short on the firms that are in the bottom characteristic quintile that are not covered by the news. We then track the return of this long-short portfolio in the next month. We regress the returns of the characteristic spread portfolios on the risk factors of Table 3, and report the corresponding estimates of the alphas and annualized Sharpe ratios. Standard errors are adjusted for serial autocorrelation as in Newey and West (1987) with a lag of 12 months. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	High SEP - Low SEP	Covered High SEP - Non-covered Low SEP
Mkt-RF	*** 0.2876	*** 0.2992
	(7.3049)	(5.6956)
SMB	−0.0896	** −0.1852
	(−1.3133)	(−3.1459)
HML	*** 0.5811	*** 0.5618
	(14.9407)	(11.8534)
MOM	*** −0.2677	*** −0.3248
	(−5.9589)	(−7.6721)
BAB	0.0633	0.0044
	(1.2305)	(0.0767)
Recession	0.0014	−0.0031
	(0.3648)	(−0.6099)
Alpha	−0.0017	0.0011
	(−1.2793)	(0.5798)
Observations	251	251
Adjusted R^2	0.695	0.6254
Annualized Sharpe ratio	0.02	0.15

Table 7: *Factor model regressions for static EP spread portfolios.* The dependent variables are the returns of different static EP (SEP) spread portfolios. Independent variables are the three factors of Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor of Carhart (1997) (Mom), the betting-against-beta factor of Frazzini and Pedersen (2014) (BAB), and an NBER recession indicator. We take the risk-free rate to be the one in the Fama-French data. Standard errors are adjusted for serial autocorrelation as in Newey and West (1987) with a lag of 12 months. The parentheses report t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

Holding period	High EP - Low EP	Covered High EP - Non-covered Low EP
1 month	* 4.87%	** 8.26%
3 months	2.86%	· 4.39%
6 months	2.43%	3.32%
12 months	2.73%	3.44%

Table 8: *Longer holding periods.* This table reports the annualized alphas of spread portfolios that hold the positions open for a varying number of months. For portfolios held more than one month, the risk factor returns are also extended to a longer horizon accordingly. Standard errors are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 12 months. ***, **, *, and · denote significance on the 99.9%, 99%, 95%, and 90% confidence levels, respectively.

$N =$	High EP - Low EP	Covered High EP - Non-covered Low EP	High coverage - Low coverage
500	* 4.87% SR = 0.21	** 8.26% SR = 0.31	** 3.95% SR = 0.43
1,000	* 5.29% SR = 0.30	*** 5.77% SR = 0.58	* 2.89% SR = 0.25
3,000	** 8.12% SR = 0.60	*** 13.88% SR = 0.88	-2.58% SR = -0.43
5,000	** 8.77% SR = 0.63	*** 15.37% SR = 0.93	* -3.43% SR = -0.51
Observations	251	251	251

Table 9: *Wider cross section.* This table shows the annualized alphas and Sharpe ratios of different spread portfolios constructed from the set of the N largest firms in the economy. Each month, we estimate the Model (2) using data for the largest 500 firms in that month. We take the estimates of $\beta_k^{(t)}$ and construct editor preference scores as in Eq. (3) for the largest N firms in the economy, where N varies from 1,000 to 5,000. We then construct long-short portfolios and run factor model regressions as in Table 3. Standard errors are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 12 months. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

	High EP - Low EP	Covered High EP - Non-covered Low EP
Mkt-RF	** 0.2136	* 0.2240
	(2.6021)	(2.3487)
SMB	* −0.2143	** −0.3072
	(−2.0135)	(−3.2705)
HML	*** 0.4290	*** 0.3605
	(4.7770)	(3.7857)
MOM	*** −0.2271	*** −0.2757
	(−3.5647)	(−4.3592)
BAB	** −0.2451	* −0.2687
	(−2.6107)	(−2.5749)
Recession	−0.0043	−0.0093
	(−0.5217)	(−0.9462)
Alpha	* 0.0052	** 0.0080
	(2.3208)	(2.8280)
Observations	251	251
Adjusted R^2	0.457	0.438
Annualized Sharpe ratio	0.32	0.43

Table 10: *Factor model regressions when we avoid trading after two consecutive market declines.* The dependent variables are the returns of different spread portfolios with the following trading exclusion: we do not trade in a month if the market return in each of the two prior months was negative. We take the market to be the sum of the market factor and the risk-free rate in the Fama-French data. There are 37 months in which we do not trade based on our trading exclusion rule. Independent variables are the three factors of Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor of Carhart (1997) (Mom), and the betting-against-beta factor of Frazzini and Pedersen (2014) (BAB), and an NBER recession indicator. We take the risk-free rate to be the one in the Fama-French data. Standard errors are adjusted for serial autocorrelation as in Newey and West (1987) with a lag of 12 months. The parentheses report *t*-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

High & Low EP firms	Coverage spread Included	Coverage spread Excluded	Residual coverage spread Included	Residual coverage spread Excluded
Mkt-RF	0.0357	** −0.0822	0.0152	* −0.0584
	(0.8791)	(−2.9114)	(0.3676)	(−2.3186)
SMB	** −0.1415	−0.0739	* 0.0757	*** 0.1164
	(−3.0883)	(−1.2310)	(2.0828)	(3.9023)
HML	0.0472	−0.1050	** 0.1777	0.0355
	(0.8248)	(−1.7544)	(2.8277)	(0.7052)
MOM	* −0.0593	0.0533	*** −0.1374	−0.0029
	(−2.1687)	(1.5099)	(−4.3131)	(−0.1251)
BAB	* −0.0957	−0.0264	−0.0343	0.0169
	(−2.4808)	(−0.7741)	(−0.6366)	(0.4632)
Recession	−0.0027	0.0005	−0.0056	−0.0038
	(−0.9611)	(0.1451)	(−1.7830)	(−1.2928)
Alpha	** 0.0033	0.0010	*** 0.0042	* 0.0023
	(2.8065)	(0.9328)	(3.4312)	(2.0093)
Observations	251	251	251	251
Adjusted R^2	0.170	0.131	0.238	0.045

Table 11: *Factor model regressions of coverage strategies.* The dependent variables are the returns of coverage and residual coverage spread portfolios that include or exclude firms identified to have high or low EP in a given month. Independent variables are the three factors of Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor of Carhart (1997) (Mom), the betting-against-beta factor of Frazzini and Pedersen (2014) (BAB), and an NBER recession indicator. We take the risk-free rate to be the one in the Fama-French data. We define residual coverage as the residual after regressing normalized coverage on normalized firm size as in Section 2.1.3. Standard errors are adjusted for serial autocorrelation as in Newey and West (1987) with a lag of 12 months. The parentheses report t-statistics. ***, **, and * denote significance on the 99.9%, 99%, and 95% confidence levels, respectively.

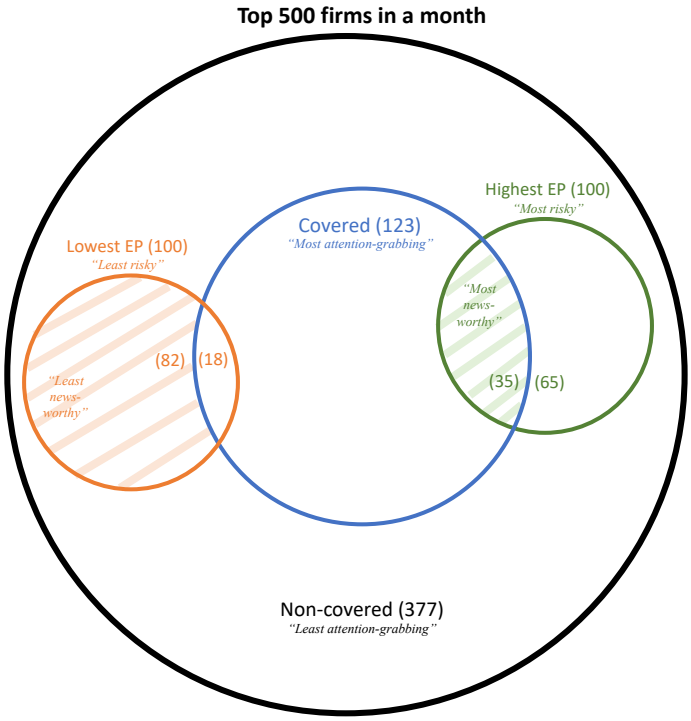


Figure 1: Comparison of covered and non-covered firms versus firms with high and low editor preference (EP). The numbers in parentheses indicate the number of firms in each group on average across all months in our sample.

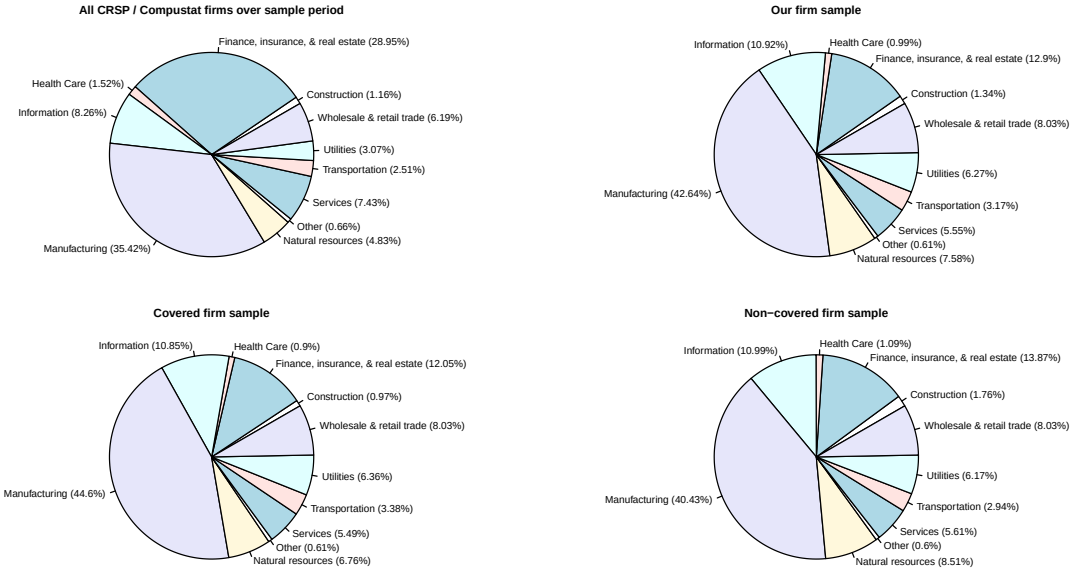


Figure 2: Sector distributions of 4 different firm groups in our sample.

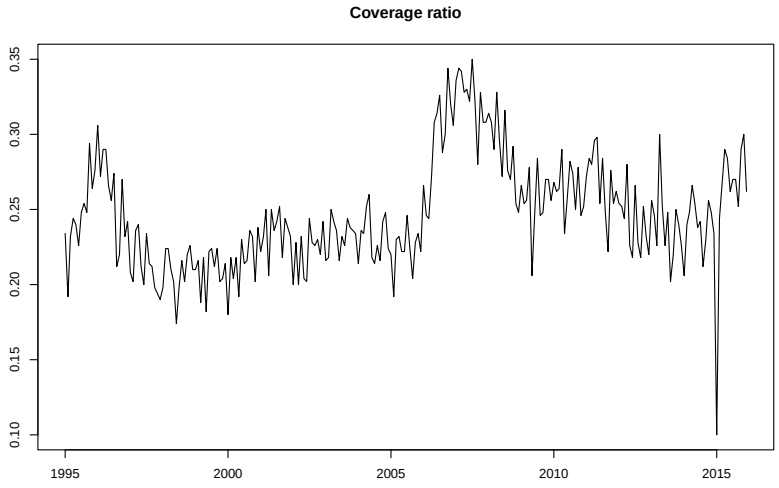
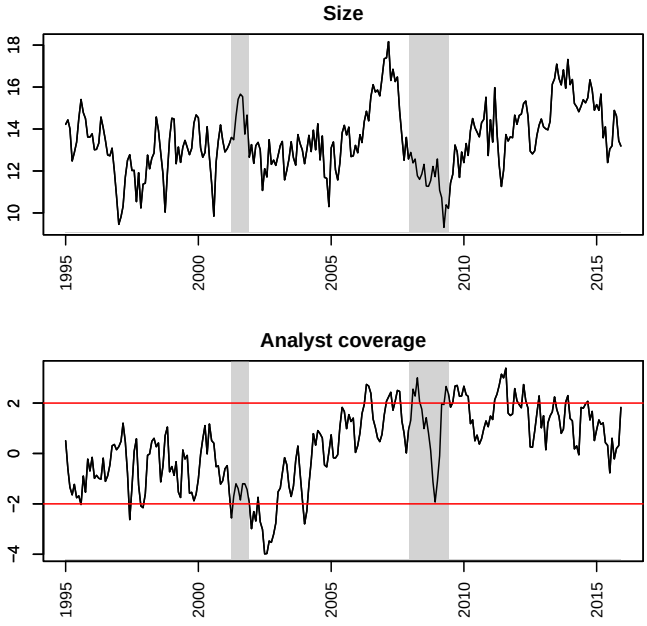
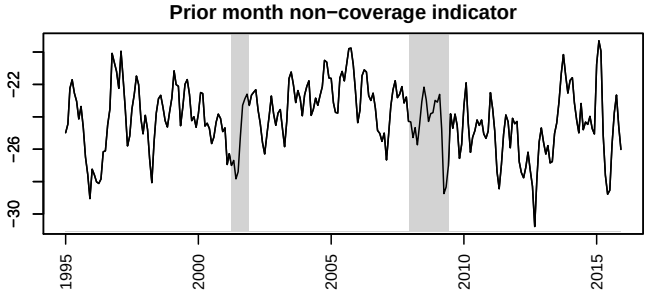


Figure 3: Fraction of the largest 500 firms in a month that are covered by the news in our sample.



(a) t-statistics of $\gamma_A^{(t)}$ and $\gamma_S^{(t)}$.



(b) t-statistics of $\delta^{(t)}$.

Figure 4: Monthly time series of the t-statistic of the coefficients $\gamma_A^{(t)}$, $\gamma_S^{(t)}$, and $\delta^{(t)}$ in the Regression (3). The red horizontal lines mark the critical values of ± 2. The gray shaded areas denote NBER recessions.

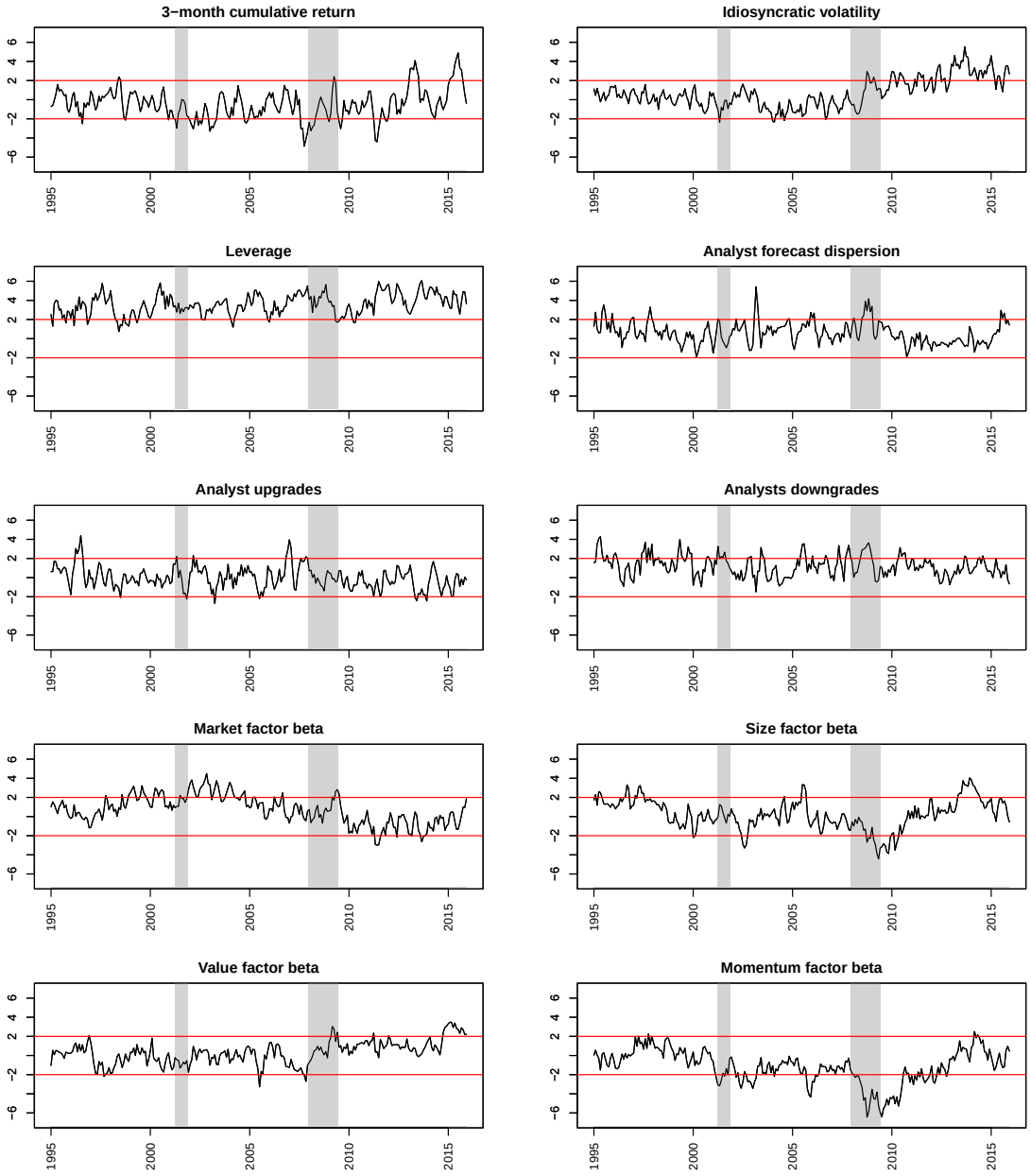
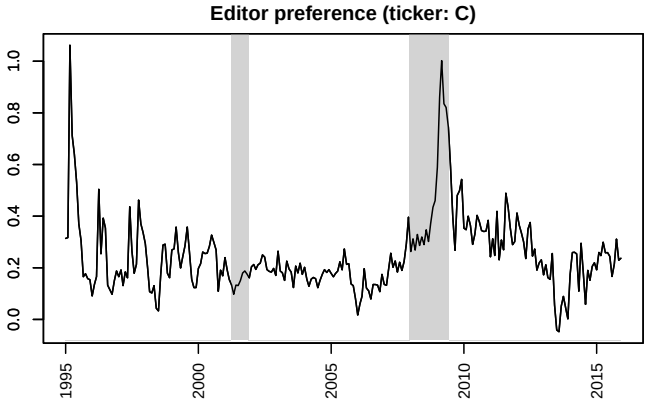
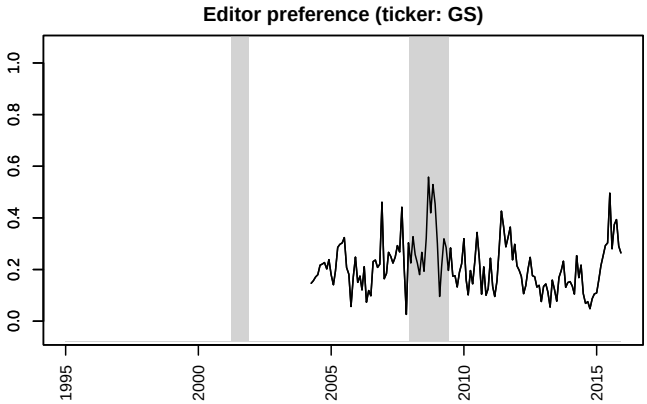


Figure 5: Monthly time series of the t-statistics of the coefficient $\beta_k^{(t)}$ in the Regression (3). The red horizontal lines mark the critical values of ± 2. The gray shaded areas denote NBER recessions.



(a) Editor preference for Citibank.



(b) Editor preference for Goldman Sachs. Goldman went public in May of 1999.

Figure 6: Monthly time series of editor preference for different firms. The gray shaded areas mark NBER recessions.

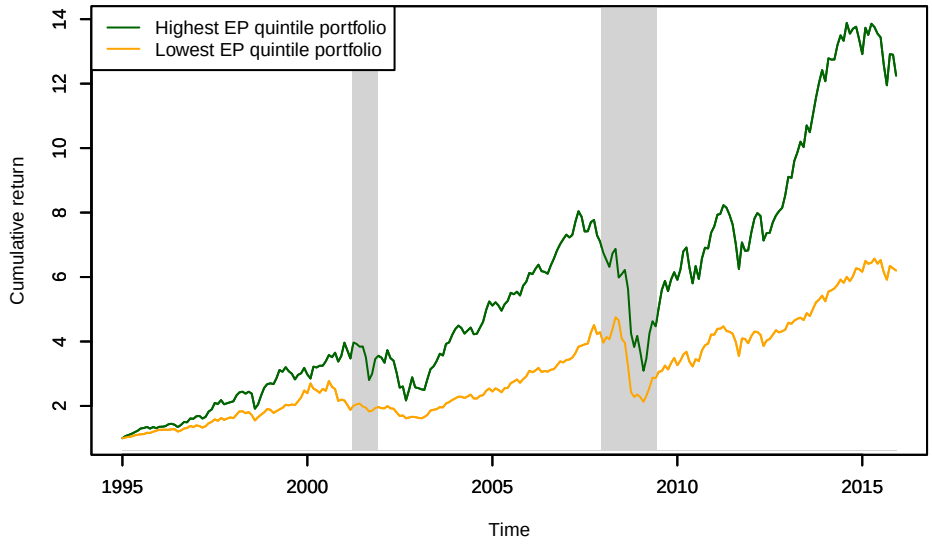


Figure 7: Cumulative returns of the highest and lowest EP quintile portfolios. The grey shaded bars indicate NBER recessions.

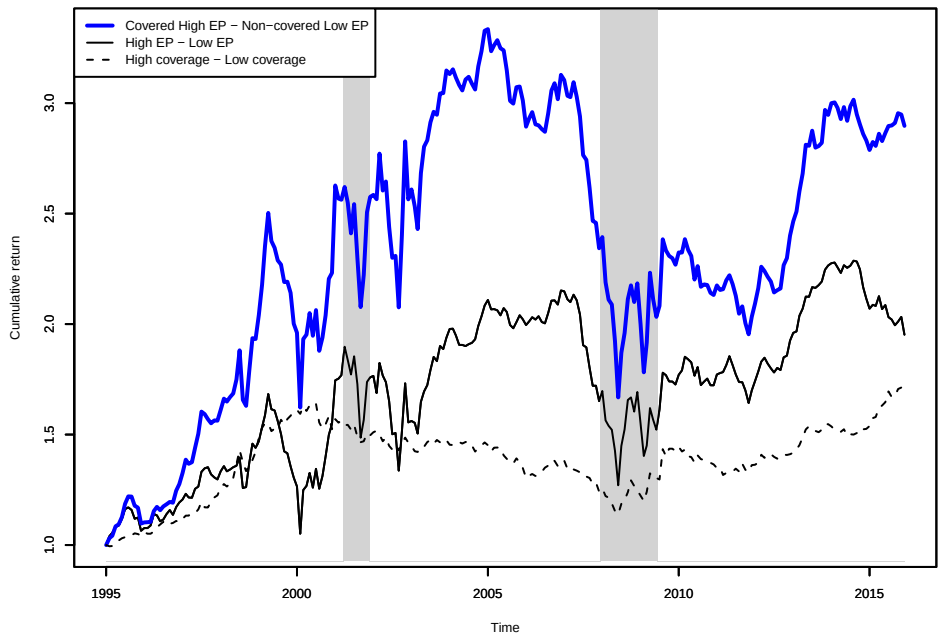


Figure 8: Cumulative returns of different quintile spread portfolios. The grey shaded bars indicate NBER recessions.

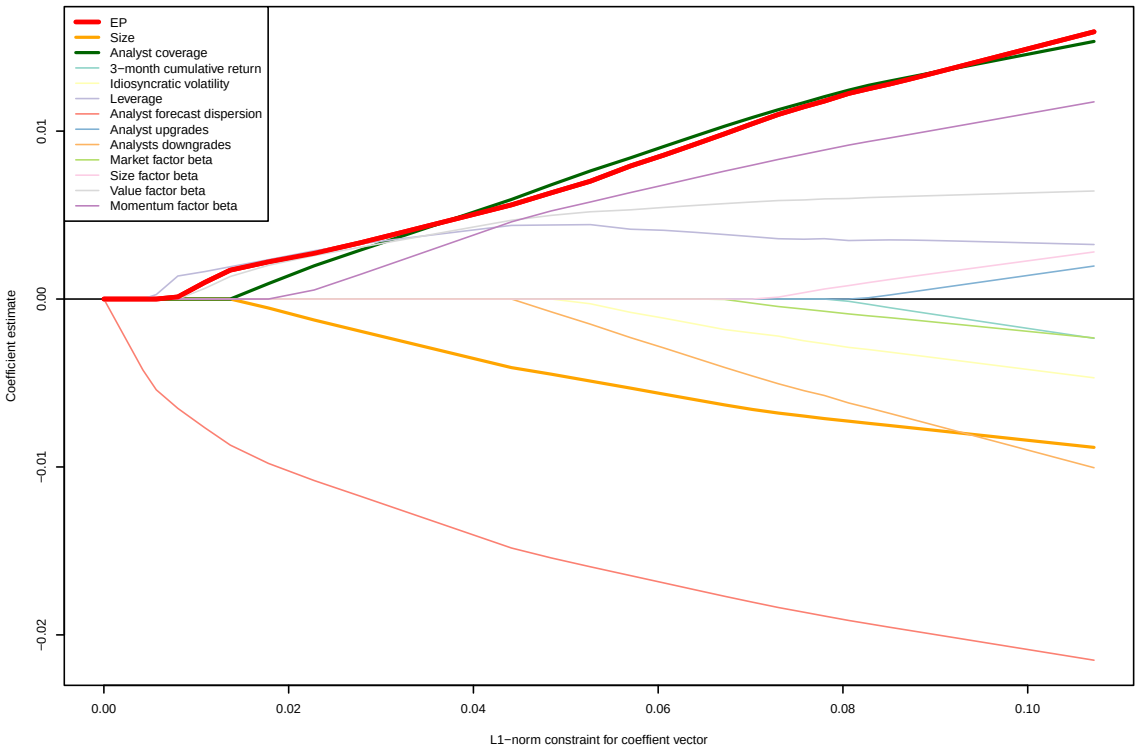


Figure 9: Estimates of b_k in the LASSO model of Eq. (5) against the constraint that LASSO imposes against the L_1 norm of the vector $(b_k)_k$.

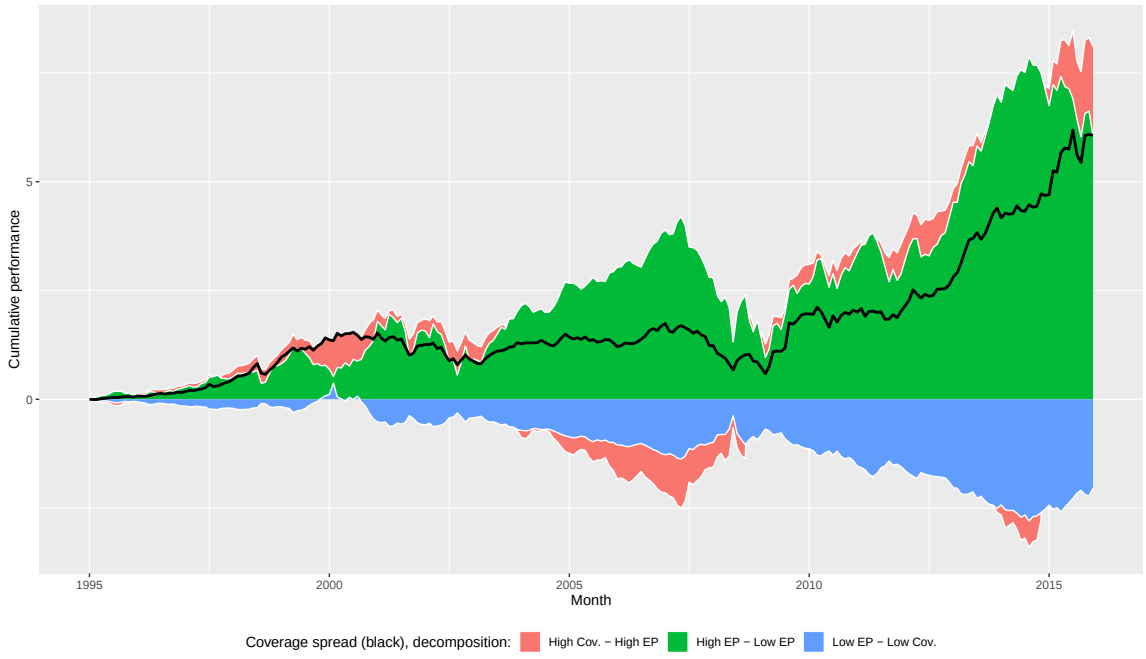


Figure 10: *Decomposition of the coverage spread.* We break down the coverage spread (High Cov. - Low Cov.) into three additive parts: The high coverage residual (High. Cov. - High EP), the EP spread (High EP - Low EP), and the low coverage residual (Low EP - Low Cov.).