Wisdom of the Institutional Crowd: Implications for Anomaly Returns

AJ Chen Gerard Hoberg Miao Ben Zhang*

November, 2022

*All authors are from the University of Southern California Marshall School of Business. Chen, Hoberg and Zhang can be reached at chen663@usc.edu, hoberg@marshall.usc.edu, and miao.zhang@marshall.usc.edu, respectively. We thank Frederico Belo, Odilon Camara, Zhi Da, Nicholas Guest, Joel Perress, Mitch Waratchka and seminar participants at Chapman University and INSEAD for valuable comments. We also thank Miyon Sung, Joseph Hedary, Ed Tilihoi and everyone at the Wall Street Journal for providing the technical advice and support. Pratheek Athreya provided outstanding research assistance. We are grateful for the research support from the University of Southern California. Any errors are ours alone. Copyright ©2021 by AJ Chen, Gerard Hoberg, and Miao Ben Zhang. All rights reserved.

Electronic copy available at: https://ssrn.com/abstract=4207132

Wisdom of the Institutional Crowd: Implications for Anomaly Returns

ABSTRACT

We hypothesize that when price correction requires more capital than any one investor can provide, institutions coordinate trading via crowd-sourcing in the media. When the crowd reaches a consensus, synchronized trading occurs, prices are corrected, and anomaly returns result. We use over one million Wall Street Journal articles from 1980 to 2020 to develop a novel textual measure of institutional investors making predictions in the media (InstPred). We show that (i) both value and momentum anomaly returns are 34% to 63% larger when InstPred is higher, and (ii) institutional investors collectively trade the anomalies more aggressively when InstPred is higher. Our results are reinforced by tests using quasi-exogenous variation in temporal investor-WSJ connections and cannot be explained by existing measures such as document tone.

Electronic copy available at: https://ssrn.com/abstract=4207132

A large body of literature studies how the news media affects financial market outcomes. These studies develop several novel insights regarding how investors respond to the news. In practice, financial news articles frequently report the views and predictions of investment banks, fund managers, investment advisors, and their employees. Hence, these institutional investors might in fact influence the production of business and financial news. In this paper, we study how institutional investors' proactive engagement in news production affects asset prices. We consider two major asset pricing anomalies, whose pervasiveness and high visibility make them ideal for testing this thesis—momentum and value.

To illustrate our intuition, consider an institutional investor who receives a private signal that an entire sector's stocks with a given characteristic (e.g., past returns or book-to-market ratio) are mispriced. A major obstacle to arbitraging away the mispricing is that the liquidity needed to move prices for an entire sector might far exceed the investor's available capital (Abreu and Brunnermeier (2002)). Placing the order and waiting is not a desirable strategy, as that would risk the investor having to wait a long time before an adequate mass of other institutions trade and eventually correct prices. Instead, we propose that sharing the signal through a reputable media outlet can help attract more investors and reduce this waiting time. As more investors learn and share their signals through the media, the number of informed investors grows. When the size of this institutional crowd reaches a critical mass, arbitrage trades occur, prices correct, and anomaly returns result. Hence, the intensity of news articles involving institutional investors' predictions can itself predict anomaly returns and institutional investors' anomaly trading.

Momentum and value anomalies are ideal for testing this hypothesis because the two anomalies are known to have large systematic components requiring large amounts of liquidity. For example, momentum has a significant industry component (Moskowitz and Grinblatt (1999), Hoberg and Phillips (2018)) and significant factor-based components (Ehsani and Linainmaa (2021)). The value anomaly is also known to have a large systematic component (Davis, Fama, and French (2000)). Moreover,

Electronic copy available at: https://ssrn.com/abstract=4207132

value and momentum have been documented to be "everywhere" (Asness, Moskowitz, and Pedersen (2013)). The high visibility of these two anomalies makes potential coordination by institutional investors more attractive, and also makes them more difficult for any one investor to arbitrage in isolation.

We test our model predictions by constructing a novel measure of institutional investors' information sharing using over one million articles in the Wall Street Journal from 1979 to 2020. We take three steps to construct our measure. First, we use Google word2vec embedding to tag each article regarding the extent to which it relates to institutional investor content, and the extent to which the article text includes predictions about the future. We take the product of these content loadings to capture the intensity of institutional investors' predictive statements (*InstPred*) in each WSJ article. Second, we aggregate the article level measure to the Fama-French 48 industries using the fact that some articles are linked to specific stock tickers in our data. To do so, we train a neural network to assign industry tags to articles without ticker tags. Finally, we compute abnormal institutional prediction activity for each industry-month by comparing the InstPred intensity in the recent 3 months to its long-term average a year ago. We map the resulting signal to the standard firm-month return database used in the anomalies literature and assess whether abnormal institutional predictions amplify momentum and value anomaly returns.

Our main empirical finding is that abnormal institutional prediction activity in the news media crowd-sourced over 3 months strongly amplifies both momentum and value anomaly returns. The amplifications are economically large—a one-standard-deviation increase in institutional predictions boosts momentum by 51%-63% relative to the benchmark level and value by 34%-45% of the benchmark level. Results are strong both in cross-sectional Fama-MacBeth regressions and in stringent value-weighted portfolio tests. Additionally, while Asness, Moskowitz, and Pedersen (2013) find that momentum and value are negatively correlated in the time-series, we show that the impact of InstPred on amplifying *both* anomaly returns manifests throughout our sample period, even during the financial crisis. This finding is consistent with

Electronic copy available at: https://ssrn.com/abstract=4207132

our conceptual framework not relying on any link to the state of the economy.

Our results also confirm the model's prediction that a crowd-sourced consensus for correcting mispricing can only be reached over time. In particular, when aggregating our textual measure over different past month horizons, we find that 3 to 12 months of signal aggregation is needed to produce highly significant anomaly amplifications.[1] The amplification effect of InstPred on the anomalies is a robust and novel feature of the data. We show that only articles scoring highly on *both* institutional investor content and prediction content contribute to the crowd-sourced signal. Moreover, separating our findings from the existing literature, our results are fully robust to controls for positive tone, negative tone, and uncertain textual tenor that are widely used in existing studies.

Although our empirical design uses specific interpretable content in the WSJ to directly test model predictions, we acknowledge that this approach does not fully rule out endogeneity concerns, especially the possibility that unobserved industry state variables might be driving our results. We thus conduct two specialized tests directly examining our model's mechanism.

First, we test a central prediction that the effect of InstPred on anomaly returns in a sector should be strong specifically when the institutional investors who specialize in the sector are connected to the WSJ. We thus use institutional investors' historical interactions with the WSJ in unrelated industries to draw quasi-exogenous variation in institutional investors' WSJ connectedness for a focal industry. We find that the amplification effects of InstPred on the anomaly returns are indeed largest in sectors whose major institutional investors have strong connectedness to the WSJ. By using quasi-exogenous variation in investor-WSJ connectedness from unrelated industries, this test illustrates the central nature of crowd-sourcing connections in driving our results.

Our second test examines institutional investor trading behavior. Our model

[1]Our finding of a gradual accumulation of information is also consistent with the concept of gradual price accumulation noted in Da, Gurun, and Warachka (2014).

Electronic copy available at: https://ssrn.com/abstract=4207132

predicts that institutional investors will trade on the anomalies specifically when a consensus in the media is reached. We use the Thomson-Reuters Institutional Holdings (13F) database and examine changes in stock holdings by actively trading institutions. We find that institutions indeed trade more aggressively on the anomalies when our measure of institutional prediction activity is higher. These results are important given the growing interest in both price and quantity tests in asset pricing (Koijen and Yogo (2018)).

We end our analysis by examining the thematic content from the WSJ articles that institutional investors discuss when making predictions. In particular, we ask which content themes relating to institutional predictions specifically boost the momentum and the value anomaly returns. The answer to this question can shed light on the root of the anomalies themselves, especially in our framework where institutional investors have the strongest incentives to share signals through the media. To answer this question, we use the taxonomy of media content with 180 themes developed by Bybee et al. (2020). We identify 25 of these themes as being "economic themes" of potential relevance in predicting returns, and we compute article-level exposures to each of these 25 themes, and then aggregate the scores to our baseline firm-month return database.

We run Fama-MacBeth return regressions that interact each of the 25 themes with our baseline institutional investor predictions variable. Several new insights relevant to the roots of anomalies emerge. Especially regarding the value premium, we find that a wide array of economic themes are important in driving value anomaly returns, as 14 of the 25 economic themes are statistically significant with a t-statistic greater than 3.0. For momentum, 6 of the 25 themes are significant at this level. These findings suggest that multiple economic forces come together to create these anomaly returns, and hence more than one theory of anomaly returns might be valid. Moreover, we find that themes that boost momentum and value are quite distinct. Themes relating to economic growth and macro conditions are uniquely important for momentum returns. In contrast, issues in corporate finance such as corporate

Electronic copy available at: https://ssrn.com/abstract=4207132

earnings, governance, IPOs, and managerial change are uniquely related to the value premium. These results add new insights regarding anomaly returns, which can motivate future research exploring the roots of the anomalies themselves.

Our paper makes two novel contributions to the literature. First, our conceptual framework and empirical findings suggest a new way to view the interaction between institutional investors, media, and asset prices. Prior studies on media and asset pricing primarily view investors as responding to news while treating the news itself as essentially exogenous.[2] As investors' response to news are oftentimes short-lived, most studies typically focus on the short-term return response to news over a horizon of a couple days (Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Tetlock (2010), Huberman and Regev (2001), Engelberg, McLean, and Pontiff (2018), among others). We propose that institutional investors also contribute proactively to the production of financial news itself by sharing their predictions in order to contribute to institutional crowd sourcing that can improve their timing of anomaly returns. As predicted by our theoretical extension of Abreu and Brunnermeier (2002), we find that news not only has the short-term instantaneous impact on returns documented in the literature, but it also has longer-term amplification effects on major anomalies that are economically large.

Our second contribution is to offer a new approach to inspect the timing and the nature of momentum and value anomalies. Our empirical findings demonstrate that institutional investors' prediction activity in news significantly amplifies momentum and value anomaly returns. Hence, our work complements existing literature on time-varying momentum and value as we illustrate a new condition that drives their significance.[3] Our news-based setting is also flexible and allows researchers to in-

[2]For instance, Tetlock (2007) demonstrates that media coverage directly influences how investors process information. Media coverage attracts investor attention (Engelberg and Parsons (2011); Fang, Peress, and Zheng (2014); Solomon, Soltes, and Sosyura, (2014)) and reduces information asymmetry (Fang and Peress (2009); Tetlock (2010); Huberman and Regev (2001); Peress (2008)). News has also been extensively studied to lead to stock market reactions (Peress (2014), Engelberg, McLean, and Pontiff (2018), Jeon, McCurdy, and Zhao (2021), Guest (2021), among others).

[3]For instance, Barroso and Santa-Clara (2015) and Daniel and Moskowitz (2016) show that while momentum is profitable on average over a long period, its profitability can vary significantly, and even crash. Similarly, a recent study shows that value premium returns also vary over time,

Electronic copy available at: https://ssrn.com/abstract=4207132

spect potential drivers of momentum and value. Many theories regarding how these anomalies arise have been proposed.[4] If informed investors believe that the proposed drivers can plausibly lead to anomaly returns, our model suggests that these institutions will discuss the relevant value drivers with the news media when making their predictions in order to increase credibility and the quality of their signal.[5]

We end this section with a note on limitations and motivations for future research. First, we focus on just one (albeit important and highly reputable) news outlet, the Wall Street Journal (see, for example, Dougal et al. 2012 and Guest 2021). Future research might explore the role of less reputable news outlets (or other information sharing systems such as social media platforms) where cheap talk is more likely. A second limitation is that we lack instruments that might further address endogeneity concerns. This issue is mitigated in part as our paper is motivated by theory with many unique predictions, and our use of textual analysis facilitates tests that more directly examine our proposed mechanisms. Yet we believe that future work further exploring endogeneity in this setting should be fruitful.

1 Conceptual Framework

In this section, we describe a simple framework to characterize how institutional investors use news media to disseminate tradeable information and coordinate price

and the returns not been significant in past two decades (Eisfeldt, Kim, and Papanikolaou (2022)).

[4]Regarding momentum for example, see Conrad and Kaul (1998), Berk et al. (1999), Johnson (2002), and Sagi and Seasholes (2007) for risk-based explanations; Jegadeesh and Titman (1993, 2001), Daniel et al. (1998), Hong and Stein (1999), Frazzini et al. (2012), Cooper et al. (2004), Griffin et al. (2003), and Asness et al. (2013) for behavioral explanations; and Korajczyk and Sadka (2004), Lesmondet al. (2004), and Avramov et al. (2013) for trading friction-based explanations.

[5]A growing strand of literature has attempted to explore the role of media in unveiling the underlying causes of anomalies. Chan (2003) is one of the earlier papers studying the relationship between news and momentum, finding evidence of slow information diffusion using news headlines. Hillert et al. (2014) use articles from 45 newspapers between 1989 and 2010 and find that firms with high media coverage exhibit stronger momentum, suggesting that media attention can impact investor behavior, thus supporting an overreaction-based story for the momentum profits. Using a novel high-frequency decomposition of daily stock returns, Jiang et al. (2021) find evidence of pervasive stock market underreaction to firm news. Our study further complements this stream of literature by providing novel evidence on corporate finance themes. By connecting corporate themes to asset pricing anomalies, we also introduce new methods to the literature that can help future researchers examining the roots of anomalies.

Electronic copy available at: https://ssrn.com/abstract=4207132

correction. Our framework adopts the basic setup of Abreu and Brunnermeier (2002), *AB model* hereafter, but adds news media into their setting.

To start with, consider a market in which the prices of certain stocks deviate from their fundamental value.[6] When mispricing is corrected, an anomaly return results. There are two types of agents labeled as rational arbitrageurs and behavioral traders, following the terminology of the AB model. Arbitrageurs actively trade on information, while behavioral traders function as the liquidity providers who absorb trading orders and stabilize stock prices to a certain limit. Each arbitrageur is assumed to be infinitesimal, and the total mass of arbitrageurs is assumed to be 1.

At time 0, $\delta < 1$ faction of arbitrageurs are informed that certain stocks are mispriced, where δ is common knowledge among informed arbitrageurs. For instance, they learned that investors underreacted to past stock returns, leading to a profitable momentum trading strategy, or investors overvalued growth stocks, leading to a profitable value strategy.[7]

There are two key assumptions about arbitrageurs in the AB model. First, all arbitrageurs are risk-neutral but face capacity constraints. Hence, trading orders from one or few arbitrageurs cannot move the price. Instead, we specify that price correction occurs only when $\kappa \leq 1$ fraction of arbitrageurs trade in the same direction.[8] When it happens, the aggregate order imbalance of arbitrageurs exceeds the absorption threshold of behavioral traders, resulting in price correction. The price of a mispriced stock thus appears constant during the buildup period when increasing arbitrageurs' place their orders and corrects at the moment when κ fraction of arbitrageurs have placed their orders. This critical mass requirement introduces a coordination element among arbitrageurs.

[6]Following Abreu and Brunnermeier (2002) and Engelberg, McLean, and Pontiff (2018), we do not specify the exact frictions that have led to the mispricing. An example from Abreu and Brunnermeier (2003) is that investors are over-optimistic about the cash flow impact of new technologies in a sector such as railway, telecommunication, and electric vehicles, leading to overpricing in some or all stocks in the sector.

[7]We assume that there is no exogenous arrival of information to arbitrageurs after time 0. Hence, arbitrageurs do not face the sequential arrival of private information as in the AB model.

[8]For simplicity, we assume that all arbitrageurs face the same maximal amount of orders they can place. Because arbitrageurs are risk neutral, they will place their order to the maximum capacity.

Electronic copy available at: https://ssrn.com/abstract=4207132

The second key assumption in the AB model is that arbitrageurs incur holding costs c per unit of time between the time they place their orders and the time the mispricing is corrected. Such holdings costs can be motivated by explicit costs such as margin requirements and borrowing costs for short selling, or opportunity costs such as inability to deploy capital to other trading strategies once the arbitrageur places the buy orders. Another example is implicit costs such as relative performance evaluation of fund managers (see more examples in Abreu and Brunnermeier (2002)). The critical mass requirement along with the holdings costs provide incentive for informed arbitrageurs to push for the price correction as early as possible.

Our key new ingredient to the AB model is that we add a news media (e.g., WSJ) through which informed arbitrageurs can disseminate their private information to uninformed arbitrageurs.[9] Without loss of generality, we assume that one news article comes out with the private information shared by an informed arbitrageur each period. Importantly, we assume that one piece of news does not perfectly transfer the private information to all uninformed arbitrageurs at once. Otherwise, price correction occurs immediately after the first arbitrageur shares her private information with the news media. Instead, we assume only ψ fraction of the remaining uninformed arbitrageurs fully accept the private information and become informed. This imperfect diffusion of information can be motivated by many reasons in practice. For instance, not all uninformed arbitrageurs may pay attention to each piece of news, even if the news contain profitable trading information. Alternatively, some uninformed arbitrageurs who read the news may not infer the tradeable information the first time they see it. As a result, as more news about the private information comes out over time, uninformed arbitrageurs progressively become informed and trade on

[9]Following Abreu and Brunnermeier (2003) and Hong and Stein (1999), we assume that only arbitrageurs watch the news. However, our anomaly return results can be obtained even if we allow some behavioral traders to read public news and trade accordingly. Abreu and Brunnermeier (2002) state that arbitrageurs in the AB framework have a strong incentive to disclose their private information to shorten their holding periods. Yet, they question whether other investors perceive the disclosures as credible. Ljungqvist and Qian (2016) study 124 cases of arbitrageurs individually publicizing privately-gathered information and found that the disclosures indeed led to strong reactions from other investors, which further supports the foundations of our crowd-sourcing hypothesis. We argue that such disclosures are more likely to be credible when publicized in highly reputable media (such as the WSJ) and that such platforms can be used for crowd-sourcing.

Electronic copy available at: https://ssrn.com/abstract=4207132

the information accordingly. More precisely, at any time $t > 0$, we can compute the mass of informed arbitrageurs to be $1 - (1 - \psi)^t(1 - \delta)$.

Once an arbitrageur becomes informed on the tradeable information, she can look back at the news and back out the mass of informed arbitrageurs. As a result, all informed arbitrageurs can fully anticipate the timing regarding when price correction occurs, i.e., when $1 - (1 - \psi)^t(1 - \delta) = \kappa$. All informed arbitrageurs thus place their orders right before the mass of informed arbitrageurs reaches κ. We thus have the following proposition:

Proposition 1: *If $\delta < \kappa$, there exists a time $t^* > 0$ at which all informed arbitrageurs place their trades and anomaly returns realize, where*

$$t^* = \frac{log(1 - \kappa) - log(1 - \delta)}{log(1 - \psi)}. \tag{1}$$

Proposition 1 makes an important empirical prediction that anomaly returns are realized only when enough (i.e., t^*) arbitrageurs share their information via the news. Hence, the intensity of WSJ articles citing statements from institutional investors (our empirical analogy for arbitrageurs) *over a span of past periods* provides a condition for the realization of anomaly returns.

Our model mechanism for Proposition 1 also makes an empirical prediction on the trading behavioral of institutional investors. In particular, as more institutional investors learn the tradeable information from WSJ, they trade in a synchronized fashion that corrects mispricing and results in anomaly returns. This leads to the following two empirical predictions that we test in our empirical section.

Empirical Prediction 1: *Anomaly returns are greater when more WSJ articles mention predictive statements from institutional investors.*

Empirical Prediction 2: *Institutional investors trade more aggressively on the anomaly when more WSJ articles mention predictive statements from institutional investors.*

Electronic copy available at: https://ssrn.com/abstract=4207132

Discussion: For simplicity, our model assumes that informed arbitrageurs can perfectly foresee the time of price correction. As a result, they all trade synchronously right before the price correction. In practice, some informed arbitrageurs may place their orders before the price correction. For instance, they may face negligible holding costs or their expected timing for the price correction is observed with noise. In these cases, we expect that anomaly returns are still realized approximately when the mass of informed arbitrageurs reaches κ at t^*. Institutional investors' trading on anomalies will not all trade precisely at t^* but their trades will become more intense as time approaches t^*.

2 Data and Measures

2.1 Data

Our news data set consists of full text of all articles published in the Wall Street Journal from June 1979 to December 2020, provided by the Dow Jones Newswires. This data set has several desirable features for testing our hypotheses. First, WSJ is among the largest newspapers on business and financial news by circulation in the U.S., making it one of the most effective media to spread tradeable information among investors.[10] Second, WSJ is widely regarded as authoritative and independent, making the quoted information providers accountable for any spreading of fake news.[11] Hence, informed institutional investors can find it worthwhile to share their information with WSJ without worrying that their news are discredited by other uninformed traders cheap talking in the same media.[12] Third, the WSJ article data set represents the longest history of digitized news available from Dow Jones & Company, allowing us to study anomaly returns over a long span of 40 years.

[10] According to the SEC 10-Q filing of News Corp (WSJ's holding company), WSJ had average daily subscriptions of 3.22 million as of December 2020. See https://www.sec.gov/ix?doc=/Archives/edgar/data/0001564708/000156470821000004/nws-20201231.htm

[11] For instance, WSJ is one of four news medias and the only business-focused news media that reached the prestigious "newspapers of record by reputation" status. See https://en.wikipedia.org/wiki/Newspaper_of_record.

[12] Abreu and Brunnermeier (2002) argue that without an institution monitoring the credibility of news, there may exist an equilibrium in which no informed trader publicize the private information.

Electronic copy available at: https://ssrn.com/abstract=4207132

We start by transforming the raw article-level texts following standard data-processing procedures (e.g., see Bybee et al., 2020). We first process each article by setting all characters to lower case, removing common stop words and words with fewer than 4 letters, and separating texts into small units (i.e., tokenization). We next convert the inflected forms of each word (e.g., "find", "finds" and "found") to be the same (i.e., a light lemmatization). From the uni-grams generated by these procedures, we obtain bi-grams of all pairs of adjacent uni-grams in the article's original word ordering as our final level of units. Our processed vocabulary includes uni-grams and bi-grams occurring in the WSJ news space of the 40-year sample.

We are interested in industry-level economic news. Hence, we exclude WSJ articles with subject tags corresponding to non-economic content such as books, sports, entertainment, lifestyles, arts, and reviews. In addition, we use the journal section tags to further exclude WSJ sections pertaining to non-economic topics such as Books, Bookshelf, Off Duty, Life & Arts and Golf Journal.

We next classify articles into industries using their verbal content. For articles that are about publicly traded firms, our data from Dow Jones come with tickers of the firms. We match these articles with tagged firms' industry standard industrial classification (SIC) from CRSP and then group SIC to Fama-French 48 (FF48) classification. For articles that do not have tickers, we apply a machine learning algorithm that classifies articles into FF48 industries based on the narrative structure of the articles itself and its topical attributes. Internet Appendix A provides more details of our machine learning procedure. To be conservative, we exclude articles that are assigned to multiple FF48 industries with no industry dominating the assignment for the prediction process. Our final sample includes 1,018,718 WSJ articles.

We combine several other firm-level data for our analysis. We use Standard and Poor's Compustat database to obtain firm financial information. We use CRSP database to obtain stocks monthly returns. Finally, we use the Thomson-Reuters Institutional Holdings (13F) database to acquire information on each stock's institutional ownership. We restrict our final sample of stocks to be common shares

Electronic copy available at: https://ssrn.com/abstract=4207132

(shrcd = 10 or 11) that are traded on NYSE, Amex, or Nasdaq. We also require stocks to have positive book value of equity. Finally, we exclude stock with price less than one dollar to avoid drawing inferences from penny stocks. We construct log book-to-market ratio as the sorting variable for the value anomaly, and stocks' past returns from $t-12$ to $t-1$ as the sorting variable for the momentum anomaly. Internet Appendix B details the definition of these variables and also other firm control variables including size, investment, profitability, standardized unexpected earnings (SUE), etc.

2.2 Measuring Institutional Investors' Information Sharing

Leveraging on the rich textual information of WSJ articles, we identify articles that encompass institutional investors' predictive statements. Institutional investors' predictive statements naturally capture their sharing of tradeable information via WSJ. After we construct the article-level measure, we aggregate the measure to FF48 industries and merge the industry-level measure with stocks for analyses and further validation tests.

2.2.1 "Institutional Investor" and "Prediction" Contents in WSJ

We measure each article's relatedness to institutional investors in two steps. First, we use Google's word2vec embedding model to identify words that are strongly related to the bigram "institutional investor." We choose the Google open-source word-embedding model that is trained on 100 billion words using Google News corpus.[13] The use of Google News as input to the model ensures that the mapping of news-media content and related vocabularies is consistent with the contextual style of WSJ newspaper language. This procedure generates a list of vocabularies that are likely to co-appear in news articles relevant to our seed word "institutional investor."

[13]The word2vec technique uses a neural network to learn the contextual use of each word based on the distribution and ordering of the words in the news corpus (Mikolov et al., 2013ab). The embedding method has been applied in recent financial studies of systemic exposures and transmission (Hanley and Hoberg, 2019) and of corporate culture (Li et al., 2020).

Electronic copy available at: https://ssrn.com/abstract=4207132

Following Hanley and Hoberg (2019), we select the top 250 words with the highest similarity score to "institutional investor" and that also appear in our WSJ article sample. Table 1 lists the top 50 words for "institutional investor." These words intuitively capture many investment banks, hedge funds, mutual funds, and words that are likely to appear in articles relevant to "institutional investors". We provide the full list of the 250 words in the Internet Appendix C.

Our second step quantifies a WSJ news article's relatedness to institutional investors based on the keywords from the first step. We compute a cosine similarity score between each WSJ article in our sample and the 250 words that are associated with "institutional investor." Cosine similarity captures the similarity of two texts irrespective of their size and has has been widely used in finance, accounting and economics studies (e.g., Bhattacharya, 1946; Salton and McGill, 1983; Hoberg and Phillips, 2016). This step yields a score that is bounded between 0 and 1, for each WSJ article regarding its relatedness to institutional investors.

To illustrate the informativeness of our measure, we examine the extent to which our institutional investor measure correlates with mentions of institutional investors' names in the WSJ articles. We collect and clean up the names of all large institutional investors from the Thomson-Reuters Institutional Holdings (13F) database, names of hedge funds from the Thomson Lipper Hedge Fund database, names of mutual funds from the CRSP mutual fund database, and names of top 100 investment banks from Corporate Finance Institute. We then count the occurrence of all names from each list in each WSJ article.[14] Table 2 shows the results of regressing our "institutional investor" measure on each of the four name-based scores while controlling for year and FF48 industry fixed effects. We observe statistically significant correlations between an article's "institutional investor" measure and its mentioning the names of the various lists of institutional investors.

We use the Google News word2vec keywords for institutional investors (instead of searching for institutional investor names as above) to identify an article's institu-

[14]Internet Appendix D provides more details on processing the names of institutional investors.

Electronic copy available at: https://ssrn.com/abstract=4207132

tional investor focus for three reasons. First, some institutional investor names are common words that can create widespread measurement error in article searches.[15] Second, some articles might draw content from individuals who work with institutional investors but they might not reference the company's name. Third, articles refer to institutional investors in many different ways and the word2vec keyword approach is specifically designed to measure this content in a comprehensive way (see Mikolov et al., 2013ab). For example, this same technology is used in search engines. As an example, consider the following paragraph from a WSJ article in 2003: "*Also, the passage of time has eroded the stigma attached to the Internet sector, prompting some institutional investors to return for a fresh look. In recent weeks, Mr. Rashtchy has gotten phone calls from fund managers he hasn't heard from in a couple of years, asking about Web stocks. Covering of positions by short-sellers has also contributed to the rise.*"[16] This paragraph and many others in the article discuss the views of institutional investors, including Mr. Rashtchy (an Internet analyst at U.S. Bancorp Piper Jaffray, an investment bank), but the investment bank's name appeared only once in the whole article.

Similar to what we did for the bigram "institutional investor", we construct an analogous score for the unigram "prediction". We thus use the word "prediction" as the seed word for the Google News word2vec embedding model, and Table 1 lists the top 50 related terms for "prediction." These include many words that commonly appear in predictive statements such as "forecast," "projection," "estimate," and "assertion." We provide the full list of 250 words in the Internet Appendix C.

Finally, our main measure for capturing a WSJ article's relatedness to predictions made by institutional investors is the product of the cosine similarity score for "institutional investor" and the cosine similarity score for "prediction." This product is multiplied by 100 for ease of reporting. Table 3 provides examples of WSJ articles that score high on the resulting "institutional investor & prediction" (InstPred) measure. These examples, in which analysts or managers of institutions share

[15]Examples include Boston Co Inc, Trust Co, and Society Corp.
[16]The article can be found at https://www.wsj.com/articles/SB104948462248570600.

Electronic copy available at: https://ssrn.com/abstract=4207132

their views on an industry's trajectory, illustrate our intuition for the "institutional investor & prediction" measure.

Panel A of Table 4 provides the summary statistics for our three article-level measures. On average, roughly 1% of the content of the WSJ articles in our sample are strongly related to the institutional investor vocabulary, and 0.8% are strongly related to predictive statements. These statistics suggest that our corpus is informative regarding the two themes we consider. The "institutional investor & prediction" score has a mean of 0.01 and a standard deviation of 0.02. To compare our measure with other textual theme measures in the literature, we also construct cosine similarities for positive tone, negative tone, and uncertainty using keywords from Loughran and McDonald (2011). Panel B shows that our "institutional investor & prediction" score is only mildly (20%-34%) correlated with these tone and uncertainty measures.

2.2.2 Standardized Institutional Investor Prediction for Industries

We next aggregate the "institutional investor & prediction" score to FF48 industry and month level to facilitate our analyses of monthly anomaly returns. To mitigate the concern that some industries persistently have higher "institutional investor & prediction" score than others, we standardize our score so that a high score indicates abnormally high media coverage of the given theme relative to the industry's long-term average. This step is crucial as such a measure will identify periods that should be seen as salient and surprising to investors relative to what has occurred in the past.

We develop a simple 2-step procedure. First, for each industry i in each month t, we compute the average "institutional investor & prediction" score Q_{it} over all articles mapped to the industry in the month. Second, we standardize Q_{it} by computing the mean and standard deviation of the thirteen observations $Q_{i,t-24}, ..., Q_{i,t-13}$, and Q_{it} itself. The standardized measure $Z_{i,t} = \left[\frac{12}{13}Q_{i,t} - \frac{1}{13}(\sum_{k=13,...24} Q_{i,t-k})\right] / \sigma_{i,t}$, where $\sigma_{i,t}$ is the standard deviation of $Q_{i,t-24}, ..., Q_{i,t-13}$, and Q_{it}.[17] Our use of the ex-ante

[17] Our standardization that includes Q_{it} in the calculation of the standard deviation ensures that

Electronic copy available at: https://ssrn.com/abstract=4207132

window from months $(-24, -13)$ ensures that information is standardized relative to the level of media coverage of the theme from a "clean period" that was over one year in the past. A high value of $Z_{i,t}$ indicates that the "institutional investor & prediction" theme is highly present in WSJ articles that cover industry i in month t.

Finally, we note that our model suggests that price correction and institutional trading occur only when the media has been publicizing institutional investors' predictive statements a period of time (after t^* periods to be precise). Hence, we construct our final measure of "institutional investor & prediction" (*InstPred*) as the rolling average of $Z_{i,t-2}$, $Z_{i,t-1}$ and $Z_{i,t}$. Our choice of a 3-month window for computing the moving average is arbitrary. However, we believe that since fund managers have to file for quarterly reports for performance evaluation, a quarter can be a natural window that informed fund managers target to complete such a trading strategy. Moreover, we experiment with various lengths of windows from 1 month to 36 months, and we find that WSJ InstPred significantly boosts momentum and value anomaly returns when the window size is between 2 months and 13 months (t-statistics > 3), with 3 to 6 months being uniformly strong. See details in Figure 1.

Using the above standardization and moving-average procedure, we also construct several other measures from WSJ articles at the industry-month level. These measures include the standardized "institutional investor" theme, standardized "prediction" theme, standardized number of WSJ articles, standardized themes regarding tone and uncertainty, and several permutations of our main standardized InstPred theme. We merge these industry-month level measures to firms for empirical analyses.

Table 5 reports summary statistics of our InstPred measure along with other asset pricing variables for our firm-month sample. The InstPred measure has a mean of 0.161 and is bounded beteen $-\frac{12}{\sqrt{13}}$ and $\frac{12}{\sqrt{13}}$. InstPred also has low correlations with

$Z_{i,t}$ is bounded in $\left[-\frac{12}{\sqrt{13}}, \frac{12}{\sqrt{13}}\right]$. Excluding Q_{it}, in contrast, would allow $Z_{i,t}$ to be unbounded and include outliers.

Electronic copy available at: https://ssrn.com/abstract=4207132

all popular cross-sectional return predictors including book-to-market ratio, past returns, size, investment, profitability and standardized unexpected earnings. InstPred has a mild (13%) correlation with the standardized number of WSJ articles. We thus include the standardizes number of WSJ articles in all of our main regressions.

3 Evidence on Momentum and Value Anomalies

This section derives our main results documenting the relationship between our media content themes and stock returns. We first focus on monthly cross-sectional Fama-MacBeth (1973) return prediction regressions. We then examine portfolio sorts.

3.1 Cross-Sectional Regressions

3.1.1 Baseline Results

We conduct Fama-MacBeth monthly regressions in which the dependent variable is stocks' monthly returns at $t+1$. To ease interpretation, we report annualized returns in percentage by multiplying the monthly returns by 1,200. Our first empirical prediction in Section 1 is that anomaly returns are stronger when institutional investors communicate more via WSJ, i.e., when InstPred is greater. Hence, we run the Fama-MacBeth regression with the following specification for each month:

$$ret_{i,t+1} = \beta_1 Anomaly_{i,t} \times InstPred_{i,t} + \beta_2 Anomaly_{i,t} + \beta_3 InstPred_{i,t} + X_{i,t} + \epsilon_{i,t+1},$$

where $Anomaly_{i,t}$ is either the stock's past cumulative return from $t-12$ to $t-1$ for the momentum anomaly or the natural logarithm of the stock's book-to-market ratio for the value anomaly, $InstPred_{i,t}$ is the WSJ institutional investor predict measure for the stock's FF48 industry, $X_{i,t}$ is an array of control variables that have been shown to predict returns, including the stocks' market capitalization (in logarithm), investment, profitability, standardized unexpected earnings (SUE), and the standardized number of WSJ articles for the stock's FF48 industry.

To ease interpretation, we standardize all non-interactive independent variables to have mean 0 and standard deviation of 1. The term $Anomaly_{i,t} \times InstPred_{i,t}$

Electronic copy available at: https://ssrn.com/abstract=4207132

is the product of the two standardized variables. In all tests, we report t-statistics based on Newey-West adjusted standard errors with two lags.

Table 6 presents the baseline Fama-MacBeth regression results. Column (1) shows that the momentum anomaly is significantly stronger when WSJ InstPred is higher. A one-standard-deviation increase in InstPred corresponds to an increase in the momentum anomaly by 1.94% per year with a t-statistic of 3.42.[18] Compared to the benchmark momentum anomaly of 3.81% when InstPred is at its mean, a one-standard-deviation increase in InstPred boosts momentum anomaly by 51% (= 1.94%/3.81%) of the benchmark level. Column (2) shows similar results after further controlling for other stock characteristics that are known to predict returns: A one-standard-deviation increase in InstPred corresponds to an increase in momentum anomaly by 1.57% per year (with a t-statistics of 3.06) or 63% of the benchmark momentum anomaly.

Columns (3) and (4) show that the value premium anomaly is also significantly stronger when WSJ InstPred is higher. A one-standard-deviation increase in InstPred corresponds to an increase in the value anomaly by 1.38% (with a t-statistic of 3.61) and 1.40% (with a t-statistic of 4.01) without and with controls, respectively. Compared to the benchmark value anomaly when InstPred is at its mean, a one-standard-deviation increase in InstPred boosts value anomaly by 34% and 45% of the benchmark value anomaly without and with controls, respectively.

Lastly, in Column (5), we inspect our control variables by running the regression without InstPred. Consistent with the literature, we observe that past returns, book-to-market ratio, profitability, earnings surprise (SUE) positively predict future returns, and investment negatively predicts future returns. Size and number of WSJ articles also show negative associations with future returns, but they are not statistically significant.

In summary, our baseline results in Table 6 show that InstPred has economically

[18]Given that our independent variables are standardized, a momentum anomaly is defined as the annual return difference between two stocks with a one-standard-deviation difference in past returns in the cross-section.

Electronic copy available at: https://ssrn.com/abstract=4207132

large impact on both momentum and value anomaly returns. These results support our model's prediction that institutional investors' predictive statements via news media are an important synchronization device that predict price corrections and anomaly returns.

3.1.2 Results using Permutations of WSJ InstPred

We next sharpen our understanding of the baseline return results by examining decompositions of the InstPred measure focusing on the importance of its parts. Note that InstPred is a product of an article's intensity on "institutional investor" and its intensity on "predict." In Panel A of Table 7, we consider measures based on the intensity of "institutional investor" alone and based on "prediction" alone. Both are constructed using the same standardization procedure we use for InstPred. Columns (3) and (5) show that both have lower t-statistics for boosting the momentum anomaly compared to InstPred in Column (1). Analogously, Columns (4) and (6) show that both measures have lower t-statistics for boosting value anomaly compared to InstPred in Column (2). Hence, news that incorporates both the "institutional investor" and "prediction" themes most effectively boosts anomaly returns. News that just mentions "institutional investors" or "prediction" alone are noisier in in predicting future outcomes. Consistent with our theoretical predictions, such news are less effective in forming a synchronization device among institutional investors that can coordinate price correction in the same direction.

To further stress-test the importance of having both the "institutional investor" and "prediction" themes mentioned in WSJ news when boosting anomalies, we construct four additional permutations in Panel B of Table 7. Specifically, we examine the efficacy of the "institutional investor" theme when measured over articles that specifically lack content from the "prediction" theme. If mentioning "institutional investor" alone is adequate for boosting anomalies, then interacting "institutional investor" intensity with either a high-"prediction" dummy or a low-"prediction" dummy should generate similar results. We construct two additional WSJ thematic

Electronic copy available at: https://ssrn.com/abstract=4207132

variables, Inst&HiPred and Inst&LoPred which are constructed by multiplying each article's "institutional investor" intensity with an above or below median "prediction" intensity, respectively. In sharp contrast to "institutional investor" content alone being adequate, we observe in Columns (1)-(4) of Panel B that although Inst&HiPred significantly boosts both the momentum and value anomalies, Inst&LoPred is not effective in boosting either anomaly. This result indicates that institutional investor content without the presence of prediction content is inadequate to predict returns, indicating that both types of content are necessary. In Columns (5)-(8) of Panel B, we report similar findings for Pred&HiInst and Pred&LoInst. Hence prediction content without the presence of institutional investor content also is inadequate, as both types of content are necessary to predict returns.

The results in this section tie our empirical measure to our conceptual framework, which predicts that price corrections follow when institutional investors communicate tradeable information through the news media, indicating the presence of actionable "wisdom from the institutional crowd".[19]

3.1.3 Controlling for Other News Measures

We next inspect whether our InstPred is an artifact of other known news measures that have been shown to predict returns. It is important to note that our research using InstPred has some major distinctions from most studies on media and stock returns. First, our goal is different from most studies as we do not seek to show that InstPred alone predicts future stock returns. Indeed most prior studies focus on unconditional measures of news such as positive tone, negative tone or uncertainty tenor to predict returns.[20] Instead, our goal is to show that InstPred specifically boosts anomalies like momentum and value. Second, our conceptual framework and empirical design do not aim to study short-term effects of news on daily returns,

[19] In the Internet Appendix Table IA.1, we address a potential concern that InstPred might capture news about institutional investors instead of news containing institutional investor' predictions of other industries. This test excludes the financial industries that contain institutional investors, and we find very similar results to our baseline findings in Table 6.

[20] See pioneering works such as Tetlock (2007), Tetlock et al. (2008), Garcia (2013), Hillert et al. (2014), Da et al. (2015), and Soo (2018) in the literature.

Electronic copy available at: https://ssrn.com/abstract=4207132

oftentimes the effects on returns on the news day (Engelberg, McLean, and Pontiff (2018)). Rather, we focus on longer-term monthly anomaly returns which is the basis for almost all of the anomalies literature. In fact, an interesting contribution of our study is that we find that lagged content from the media can predict monthly anomaly returns even when value-weighted. This is important as few studies in the media and asset pricing literature find long-term effects.

Within these distinctions in mind, we run horse race tests by controlling for four other news measures, one at a time, along with their anomaly interactions to examine their incremental effects. Columns (1) and (2) of Table 8 show that although InstPred robustly boosts momentum and value anomaly returns, the control for the number of articles does not boost either. This is consistent with Engelberg, McLean, and Pontiff (2018), who find that having news on a particular day only boosts anomalies on the same day itself. As we average content over longer periods of time (3 months) and lag our measures another month, and focus on industry-level signals, it is intuitive that our results are highly distinct from those in that study and are geared toward testing a different hypothesis rooted in difficult-to-coordinate anomaly trading.

Columns (3)-(8) of Table 8 add controls for the aforementioned widely-used features of news articles: positive tone, negative tone, and uncertainty. We observe that InstPred continues to robustly boost both the momentum and value anomaly returns even in the presence of these controls. We also find that neither measure of tone or uncertainty predicts future monthly anomaly returns. On the surface, the lack of results for tone might seem at odds with the literature. However, this is not the case as existing studies that find results for tone focus only on short-term return prediction, and they focus on measures of media content that are firm specific rather than the more liquidity-heavy industry-wide measures we focus on. Overall, the robustness of our findings to including controls for these measures further illustrates that our results are distinct from findings in the existing literature.

Electronic copy available at: https://ssrn.com/abstract=4207132

3.2 Portfolio Sorts

In this section, we use portfolio sorts to reinforce our findings using Fama-MacBeth regressions. We construct portfolios following Fama and French (1993; 2015). Specifically, we construct breakpoints for portfolios only using NYSE stocks. In each month, we sort stocks into 2 groups based on the NYSE median market capitalization. Independently, we sort stocks into 3 groups based on the 30% and 70% percentiles of the anomaly predictors among NYSE stocks (i.e., book-to-market for the value anomaly and past returns for the momentum anomaly). Also independently, we sort firms into 3 groups based on the 30% and 70% percentiles of InstPred among NYSE stocks. This procedure results in 18 2×3×3 size-anomaly-InstPred portfolios. We then construct value-weight excess returns for each of the 18 portfolios. Following Fama and French (1993; 2015), we compute the returns of the 3×3 anomaly-InstPred portfolios by averaging returns of the large-cap and small-cap portfolios within each anomaly-InstPred category.

Table 9 shows the results. Panel A shows the momentum anomaly conditional on InstPred. When the industry's InstPred is low, we do not observe a significant momentum anomaly. The long-short portfolio based on past returns generates an insignificant 3.70% return per year (t-statistics = 1.50). As we move from low-InstPred to high-InstPred, the long-short portfolio returns increase monotonically to a highly significant 6.84% per year (t-statistics = 2.70) for the high-InstPred group.

Panel B shows the portfolio sort results for the value premium conditional on WSJ InstPred. Similar to the momentum anomaly, the long-short book-to-market portfolio generates an insignificant 0.16% per year (t-statistics = 0.08) when the industry's InstPred is low. The long-short returns increase monotonically from low-InstPred to 4.31% per year (t-statistic = 2.21) for the high-InstPred group.[21]

[21]It is well known that value premium is not significant in recent 15 years (see Eisfeldt, Kim, and Papanikolaou (2022)). Such low performance of the value premium drags down the overall significance in each long-short portfolio return in Panel B of Table 9. In the Internet Appendix Table IA.2, we confirm that unconditional value anomaly returns in our sample, i.e., our replication of the HML factor of Fama and French (1993), and also the HML factor from Kenneth French's website, are insignificant in our sample period. Our unconditional value anomaly returns replicate

Electronic copy available at: https://ssrn.com/abstract=4207132

We next explore the distribution of our portfolio returns over our sample period. Our thesis based on institutional investors using crowd sourcing to correct mispricing does not predict our return results to be state dependent. Figure 2 plots ten-year smoothed portfolio returns (see Linnainmaa and Roberts (2018)) for the momentum anomaly (Panel A) and the value anomaly (Panel B) in High-InstPred and Low-InstPred portfolios. As the red dotted line (High-InstPred portfolio) is notably above the blue solid line (Low-InstPred portfolio) throughout our entire sample period with few exceptions for both anomalies, we conclude that the amplification effect of InstPred is not strongly linked to the state of the economy. We also highlight that the gap between the two lines even holds up during the financial crisis of 2008, indicating our results are not exposed to the momentum crash noted by Daniel and Moskowitz (2016). These findings differentiate our channel from alternative explanations, for example, the notion that media attention to extreme events such as crisis periods might generate our results.

In summary, our portfolio results (in value-weighted calendar time) reinforce our baseline findings established using Fama-MacBeth regressions. These results suggest that these returns are potentially tradeable in practice. This conclusion supports our model mechanism in which uninformed arbitrageurs not only recognize the tradeable information from informed arbitrageurs via news media, but they also trade on the information so as to push for a price correction. We show evidence of institutional investors' trading behavior in the next section.

4 Tests of Mechanism

In this section, we conduct two tests of our theoretical mechanism. The first considers plausibly exogenous variation in the extent to which institutional investors are connected to the WSJ. This test directly examines our model's proposed mechanism, where direct connections between informed institutions and the WSJ specifically drive our asset pricing results. The second test examines our model prediction re-

the HML factor with a correlation of 98%.

Electronic copy available at: https://ssrn.com/abstract=4207132

garding institutional investors' trading patterns, as we expect institutions to trade more aggressively on anomalies when InstPred is high.

4.1 Institutional Investor Connections to the WSJ

Our empirical design using interpretable content in the WSJ directly tests our model's predictions about institutional investors' crowd-sourcing and reduces alternative interpretations. Yet, this approach does not fully rule out endogeneity concerns, especially the possibility that unobserved industry state variables might be driving both InstPred and future anomaly returns. For example, our industry-specific InstPred measure may correlate with a hidden signal regarding industry performance that drives the industry's future anomaly returns. If this was the case, InstPred would appear to boost anomaly returns regardless of whether institutions communicate trading signals to the WSJ or not. To address this concern, we now consider plausibly exogenous variation in investors' connectedness to the WSJ. Our tests use this variation to examine whether institutional investors' interactions with the WSJ (instead of the industry's state variable in the given month) are central to our anomaly findings.

Our theory's most direct prediction is that observable interactions between informed institutions and the WSJ should drive our predictable anomaly returns. Our model requires that informed institutions have direct connections to reputable media, such as the WSJ. In practice, any given institutional investor will have strong or weak connections to the WSJ. This variation motivates a rather direct test of our proposed mechanism. When the set of institutions most focused on a given sector in a given month has strong connections to the WSJ, the crowd-sourcing mechanism we propose is likely to be stronger relative to when the focal institutions are not connected. InstPred should thus amplify anomaly returns more in these sector-months. In contrast, when the investor-WSJ connection is weak, InstPred should be less informative.

To measure ex-ante connectedness using only plausibly exogenous variation, we

Electronic copy available at: https://ssrn.com/abstract=4207132

measure each Fama-French 48 industry's average *investor-WSJ connectedness* in each month using three steps. First, We identify an industry i's major institutional investors in the Thompson Reuters Institutional Holdings (13F) Database as those having over 20% of their portfolio allocated to industry i's stocks in the previous quarter, or those institutions whose percentage allocation to the given industry i ranks among the top 10 of all institutional investors in the previous quarter.[22]

In the second step, we measure each major institutional investor's connectedness to the WSJ based on the occurrence of the given institutional investor's name appearing in WSJ articles during the three years prior to the previous quarter.[23] In order to ensure the variation we use is plausibly exogenous to the state of industry i in month t, we measure each major institutional investor's WSJ connectedness using only WSJ articles that covered industries other than i. Because this measure excludes all content from industry i itself, any state variable relevant to industry i cannot drive this variable's impact on anomalies returns.

In our final step, we compute industry i's weighted average investor-WSJ connectedness in a month by averaging its major institutional investors' WSJ connectedness (as of the previous quarter based on past articles excluding industry i's as noted). We value-weight this average by each major institutional investor's dollar holdings of the industry's stocks in the previous quarter.

To test our core asset pricing predictions, we sort FF48 industries into above and below-median groups based on each industry's average investor-WSJ connectedness defined above. The two groups are *Industries with High Investor-WSJ Connectedness* and *Industries with Low Investor-WSJ Connectedness*. Our crowd sourcing thesis predicts that InstPred is more likely to amplify anomaly returns in industries with high investor-WSJ connectedness.[24]

[22] Using our approach, each industry-month has, on average, 26 major institutional investors in a given quarter.

[23] Lagging the measures by one quarter ensures that the investor-WSJ connectedness measure is not directly affected by our policy variable InstPred, which is constructed based the three months prior to the return realization month. Internet Appendix D provides details on counting an institutional investor's occurrence in WSJ articles.

[24] Note that high or low investor-WSJ connectedness is not a permanent feature of an industry

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 10 displays the results and reports our baseline Fama-MacBeth regressions for the high and low connectedness subsamples. We observe that our key cross terms between InstPred and each anomaly are economically large and statistically significant in industries with high investor-WSJ connectedness, but not in industries with low connectedness. These results obtain for both momentum and the value premium. Difference tests of the coefficients between the two subsamples are reported in Column (9), and show that the cross-term coefficients in the high WSJ connectedness subsample are significantly greater than those in the low connectedness sample (see also Internet Appendix Table IA.3) at either the 5% or the 1% level.

Overall, this quasi-natural experiment using variation in connectedness measured from unrelated industries shows that WSJ articles containing institutional predictions significantly amplify anomalies when an industry's major institutional investors are ex-ante connected to the WSJ. Results are essentially insignificant otherwise. These findings support our model's mechanism that institutional investors use reputable media such as the WSJ to coordinate their trades when they have access to the media, accelerating price adjustments associated with anomaly returns.

4.2 Evidence on Institutional Investor Trading

After showing the price effects (i.e., returns) of the WSJ InstPred, we now explore its effects on the quantity of institutional investors' holdings to further test our model mechanism. The second empirical prediction of our model states that institutional investors will trade more aggressively on anomalies when InstPred is high, which then leads to the price correction. We thus examine the important prediction that institutional investors will change their holdings precisely when InstPred is high and anomaly returns are expected to be large.

We obtain institutional common stock holdings from the Thomson-Reuters Institutional Holdings (13F) Database, which are compiled from the quarterly filings of

as it varies over time. For example, the high investor-WSJ connectedness dummy has a quarterly autocorrelation of 0.70.

Electronic copy available at: https://ssrn.com/abstract=4207132

SEC Form 13F. All institutional investment managers that exercise investment discretion on accounts holding Section 13(f) securities, exceeding $100 million in total market value, must file the form. These institutions collectively manage 68 percent of the US stock market, with the cross-term32 percent attributed to holdings directly held by households and non-13F institutions (Koijen and Yogo (2019)). Form 13F reports only long positions and not short positions. A stock's institutional ownership is defined as the ratio between shares held by the institutional investors (from the 13F database) and the stock's total shares outstanding (from the CRSP database).

We next identify high-activity and low-activity institutional investors following the widely-used categorization from Brian Bushee's website.[25] In particular, Bushee (2000, 2001) categorizes institutional investors with high portfolio turnover and highly diversified portfolio holdings as "transient." We regard the institutional investors labeled as transient as being the actively-trading institutional investors seeking to coordinate trades as modeled by our theory.[26] Bushee (2001) categorizes institutional investors investing in certain portfolio firms with low turnover as "dedicated." We regard these more passive institutional investors as being less likely to be those in our model that are seeking coordination regarding anomaly profits. Finally, as there are no predicitons regarding funds that simply track indices, following the convention in the literature, we remove the funds classified as quasi-indexers from our sample.[27] We note that the highly active institutional investors that are most relevant to our theory are also institutionally important. In particular, they account for 80 percent of the institutional ownership in our sample for the average stock from 1980 to 2020.

We compute our test variable as the quarterly change in institutional ownership for each institutional investor. We then compare results for high-activity institutional

[25] https://accounting-faculty.wharton.upenn.edu/bushee/

[26] Bushee's website provides time-varying labels and permanent labels of transient for each institutional investors. We choose the permanent labels to mitigate the concern that institutional investors who traded on anomalies in the month are mechanically labeled as active.

[27] It is common practice to exclude index funds from studies exploring how funds create alpha as our framework models. See for example Kacperczyk, Sialm, and Zheng (2005) and Hoberg, Kumar, and Prabhala (2018).

Electronic copy available at: https://ssrn.com/abstract=4207132

investors and low-activity institutional investors to further test if our prediction that results should be stronger for the most active institutional investors.

To test our model's predictions, we merge our database of anomalies and monthly stock return predictors (including our key variable WSJ InstPred) with this database of quarterly changes in institutional ownership for stocks. We then run the following cross-sectional regression:

$$\begin{aligned}\Delta InstOwn_{i,t+1} =& \beta_1 Anomaly_{i,t} \times InstPred_{i,t} \\ &+ \beta_2 Anomaly_{i,t} + \beta_3 InstPred_{i,t} + X_{i,t} + FE_t + \epsilon_{i,t+1},\end{aligned}$$

where $\Delta InstOwn_{i,t+1}$ is the change in institutional ownership from month $t-2$ to $t+1$, $Anomaly_{i,t}$ is the stock's past returns or natural logarithm of book-to-market ratio, $InstPred_{i,t}$ is the WSJ institutional investor predict measure for the stock's FF48 industry, $X_{i,t}$ is an array of control variables that have been shown to predict returns. Importantly, our regressions indicate a high bar as our dependent variable is a change in holdings (not a level), and we include FE_t, the monthly fixed effect. We standardize all independent variables to have a mean of zero and a standard deviation of one for the ease of interpretation.

Table 11 presents the results. Columns (1) shows the overall non-index institutional investors' trades on momentum. We observe that when WSJ InstPred is high, institutional investors buy more past winners and sell more past losers. Moreover, confirming our characterization that active institutional investors trade consistent with the arbitrageurs in our model, we observe that the effect mainly derives from the high-activity institutional investors (in Column (2)), but not from low-activity institutional investors (in Column (3)). The economic magnitude is moderate in nominal terms but quite meaningful in relative terms. Column (2), for example, shows that a one-standard-deviation increase in InstPred boosts high-activity institutional investors' momentum trading by 2.96 basis points. When compared to high-activity institutional investors' unconditional momentum trading benchmark of 37.28 basis points, a one-standard-deviation increase in InstPred boosts momentum

Electronic copy available at: https://ssrn.com/abstract=4207132

trading by a rather substantial 8% (=2.96/37.28) of the benchmark level.

Similarly, in Column (4), we observe that our institutional investors buy more value stocks and sell more growth stocks when WSJ InstPred is high. In Columns (5) and (6), we observe a similar pattern that the effects in Column (4) are entirely driven by high-activity institutional investors rather than low-activity institutional investors. The economic magnitude of InstPred on value trading is larger than on momentum trading. Compared to the benchmark value trading by high-activity institutional investors, a one-standard-deviation increase in InstPred boosts value trading by 24% (=1.67/6.97) of their benchmark trading level. The stronger results for the value premium echo the fact that we also find a somewhat statistically sharper link for the value premium when predicting returns. Moreover, as the value premium is a slow moving anomaly, it is intuitive that coordination of the timing of trades would be more important for this anomaly, as better timing for a slow-moving anomaly is likely to boost Sharpe ratios more dramatically.

Overall, our results for active institutional investors' trading activity further support our proposed hypotheses regarding coordination of trading and anomaly returns. Our unified findings on both quantity and price changes depict a fuller picture of our proposed mechanism. These results are consistent with these active institutional investors coordinating their trading and signals via the news media to synchronize their trading such that they can time price corrections related to value and momentum.

5 What Content Indicates Momentum and Value?

In this section, we further explore the rich content in our over 1 million WSJ articles to shed light on two questions. First, what economic topics do the articles with high InstPred discuss? The answer to this question not only provides a potential validation regarding the plausibility and substance of our InstPred measure, but it also offers a rare opportunity to show descriptively which content themes institutional investors use to form predictions, and which in turn actually predict returns. Second, we

Electronic copy available at: https://ssrn.com/abstract=4207132

ask which content themes relating to InstPred specifically boost the momentum and the value anomaly returns. In our conceptual framework, arbitrageurs communicate their private trading signals regarding momentum or value strategies via news media. Hence, the specific economic content that drives InstPred's ability to boost anomalies can further inform the nature of what drives the anomalies' mispricing.

We select WSJ content themes from those provided by Bybee, Kelly, Manela, and Xiu (2020).[28] We use the following decision criteria regarding which of the 180 topic themes to include in our analysis. We wish to focus on themes that are plausibly important to predicting stock returns, and themes that are plausibly relevant to valuing corporate assets in general. We thus include all themes in the following two categories: corporate earnings and economic growth. These two themes include discussions of earnings, financial reports, macroeconomic data, recessions, and the Federal Reserve. In addition, we scanned other themes outside of these two areas for additional topics of interest, which led us to additionally include: mergers, corporate governance, control stakes, takeovers, payouts, IPOs, competition, venture capital, executive pay, and management change. We believe the resulting set of 25 topic themes provides a relevant and interesting snapshot of many relevant economic issues. At the same time, the list is not too large to preclude us from adding all of them into one regression while avoiding multicollinearity concerns.

To explore the first question (what economic themes are covered by articles with high-InstPred), we regress each article's intensity of mentioning institutional investor prediction (InstPred) on the article's intensities regarding the 25 topic themes.[29] With these intensity measures, we run the following regression:

$$InstPred_{j,t} = \sum_{k=1,..25} \beta_k Theme_{j,k,t} + FE_t + FE_{ind} + \epsilon_{j,t},$$

where $InstPred_{j,t}$ is the InstPred intensity of article j at month t, $Theme_{j,k,t}$ is the

[28]The topics and the keywords for each topic of Bybee, Kelly, Manela, and Xiu (2020) can be downloaded at http://structureofnews.com/#. We thank the authors for making these data publicly available.

[29]Following our procedure in Section 2.2.1, we define each article's intensity regarding a topic theme to be the cosine similarity of the article's content with the keywords for the topic theme provided by Bybee, Kelly, Manela, and Xiu (2020).

Electronic copy available at: https://ssrn.com/abstract=4207132

intensity of topic theme k for article j at month t, and FE_t and FE_{ind} are year fixed effects and FF48-industry fixed effects, respectively.

Table 12 shows that articles that score high on InstPred tend to be most related to fundamentals such as share payouts and earnings forecasts; corporate finance and innovation issues such as IPOs and venture capital activity; and also to macro variables such as recessions. Interpreting this finding through our conceptual framework, the somewhat wide-ranging set of topics suggest that the roots of various anomaly mispricing are likely not uni-dimensional, as their predictable returns likely arise from a basket of economic shocks and state variables. This is further consistent with the multiplicity of rather distinct anomaly theories in the literature (ranging from risk-based to behavioral to macro-based), all declaring success in various tests each study focuses on.

To explore the second question regarding which of these content themes specifically boost the momentum and value anomaly returns, we construct 25 standardized measures corresponding to each of the 25 topic themes for each of the FF48 industries in each month. We then map the industry values to our firm-month return database. For each topic theme k, we then run the following Fama-MacBeth regression that includes all control variables in our baseline regressions:

$$\begin{aligned} ret_{i,t+1} =& \beta_1 Anomaly_{i,t} \times InstPred_{i,t} \times Theme_{i,k,t} + \beta_2 Anomaly_{i,t} \times InstPred_{i,t} \\ &+ \beta_3 Anomaly_{i,t} \times Theme_{i,k,t} + \beta_4 InstPred_{i,t} \times Theme_{i,k,t} \\ &+ \beta_5 Anomaly_{i,t} + \beta_6 InstPred_{i,t} + \beta_7 Theme_{i,k,t} + X_{i,t} + \epsilon_{i,t+1}. \end{aligned}$$

For each topic theme k, we run the above regression for the momentum anomaly and also for the value premium anomaly, and for each regression, we record the t-statistic of the key triple interaction term interacting the anomaly variable, InstPred, and the given economic theme. Figure 3 plots the ordered-pairs of two t-statistics for each of the 25 topic themes (one for the value anomaly and one for the momentum anomaly), with x-axis representing the t-statistics for value and the y-axis representing the t-statistics for momentum. This provides an intuitive visualization regarding

Electronic copy available at: https://ssrn.com/abstract=4207132

which economic themes are most important for each anomaly, which are important to both anomalies, and which are not important at all.

Two observations stand out from Figure 3. First, there are many themes that boost value anomaly returns when InstPred is high. For instance, 14 of the 25 topic themes interacted with InstPred and book-to-market have t-statistics for β_1 above 3. In contrast, the topics that boost momentum when InstPred is high are fewer: Only 6 of the 25 topic themes have an analogous t-statistic for β_1 greater than 3. Second, the topics for which a higher InstPred boosts value and momentum returns are visibly different. Those that facilitate the effects of InstPred on the value anomaly but not the momentum anomaly are more related to corporate governance and corporate earnings, such as management change, executive pay, various measures of earnings, and corporate control, etc. Topics that uniquely facilitate InstPred effects on momentum, in contrast, are more related to issues that relate to economic growth including the economic growth theme, macro data, optimism, and European sovereign debt etc.

Interpreting these results through the lens of our model, we conclude that the momentum-related signals relate most to economically important *changes* in fundamentals. This is consistent with the view that momentum is an underreaction to major shifts in prices (e.g., Hong and Stein (1999), Daniel, Hirshleifer, and Subrahmanyam (1998), Jegadeesh and Titman (2011), and Hoberg and Phillips (2018)). It follows that private signals that indicate major revisions to fundamentals such as the economic growth themes are valuable for identifying the underreaction in prices and in timing when to trade, especially when large amounts of liquidity are required to move stock prices (a key foundation for how we construct our tests). Our results also suggest that signals that do not facilitate InstPred to boost the momentum anomaly are typically related to corporate variables and issues that are more passive in nature.

In contrast, value-premium related signals are more numerous overall. The economic themes that facilitate the value anomaly but not the momentum anomaly are arguably more related to longer-term issues such as managerial incentives and effort,

Electronic copy available at: https://ssrn.com/abstract=4207132

value creation through innovation, and the value each manager brings to the firm. Intuitively, this accords well with the conventional wisdom that the value anomaly is a slower moving anomaly than is momentum. While these results are ultimately supportive of the conclusion that the value anomaly likely has multiple economic roots, they also are consistent with a strong link between corporate finance and asset pricing in the context of the value anomaly.

Overall, our results in this section add intuition that corroborates the mechanisms and plausibility behind the return predictability that drives our key InstPred variable. Institutional investors have information sets that are widely varied, and these findings suggest that they are willing to share many different types of signals in order to potentially induce crowd sourcing of investment timing by other institutions as our model suggests. We also believe these results motivate future researchers to use these methods and results to further assess the predictions of specific theories regarding the roots of various anomaly returns, an exercise that is outside the scope of our current study (which is focused on institutional investor coordination and crowd-sourcing).

6 Conclusion

We propose that a coordination game among limited liquidity institutions plays out on a regular basis, as many traders must act in unison to generate anomaly returns when each is too small to move prices. Our central thesis is that these institutions coordinate, at least in part, through highly reputable and visible news sources such as the Wall Street Journal. Our theoretical extension to Abreu and Brunnermeier (2002) predicts that informed institutions will share signals through the media, and price correction occurs only after enough such articles accumulate and enough uninformed investors become informed.

To test the model predictions, we construct a novel measure using the full text of over one million WSJ articles from 1979 to 2020 to capture institutional investors' sharing of their predictions (InstPred) in the news. Using the measure, we present

Electronic copy available at: https://ssrn.com/abstract=4207132

new evidence of economically large amplifications of two important anomalies in the literature known to have broad-sectoral and systematic components that are likely too large for any individual fund to influence: value and momentum. Consistent with the theory, these returns are indeed largest when news covering institutional investors' predictions are accumulated. Accumulations of roughly 3 months to 12 months produce consistent large amplifications. These features fundamentally distinguish our work from most of the prior studies on media and asset pricing, which focus on short-term (i.e., daily) effects of news, and that focus on the tone of articles (rather than interpretable content as we do).

Our conclusions are reinforced by two customized experiments targeting our proposed mechanism. (i) We find using quasi-exogenous variation in investor connections to the WSJ (based on unrelated industries) that our return prediction results are strongest when major institutional investors in an industry are most connected to the WSJ. (ii) We confirm that active institutional investors indeed trade more aggressively on the anomalies when InstPred is higher.

Further analysis of economic topics in WSJ articles indicates that institutions build signals using information that spans a wide-array of economic forces. Value anomaly returns are most likely to become amplified when institutional investors discuss corporate finance themes including corporate earnings, venture capital, and governance. Momentum returns are most likely to become amplified when institutional investors discuss macro themes such as economic growth, product prices, and macro data. We believe that technologies enabling sharp thematic content analysis will be invaluable to researchers exploring the roots of asset pricing anomalies.

Electronic copy available at: https://ssrn.com/abstract=4207132

References

Abreu, D. and Brunnermeier, M.K. "Synchronization risk and delayed arbitrage." *Journal of Financial Economics*, 66.2-3 (2002), 341-360.

Abreu, D. and Brunnermeier, M.K. "Bubbles and crashes." *Econometrica*, 71.1 (2003), 173-204.

Asness, C.S., Moskowitz, T.J. and Pedersen, L.H. "Value and momentum everywhere." *The Journal of Finance*, 68.3 (2013), 929-985.

Avramov, D., Cheng, S. and Hameed, A. "Time-varying momentum payoffs and illiquidity." (2013) Available at SSRN.

Avramov, D., Chordia, T., Jostova, G. and Philipov, A. "Anomalies and financial distress." *Journal of Financial Economics*, 108.1 (2013), 139-159.

Barberis, N.; A. Shleifer; and R. Vishny. "A model of investor sentiment." *Journal of Financial Economics* 49 (1998), 307–343.

Barroso, P. and Santa-Clara, P. "Momentum has its moments." *Journal of Financial Economics*, 116.1 (2015), 111-120.

Berk, J.B., Green, R.C. and Naik, V. "Optimal investment, growth options, and security returns." *The Journal of Finance*, 54.5 (1999), 1553-1607.

Bhattacharya, A. "On a measure of divergence between two multinomial populations." (1946).

Blei, D.M., Ng, A.Y. and Jordan, M.I. "Latent dirichlet allocation." *Journal of machine Learning research*, 3.Jan (2003), 993-1022.

Bybee, L., Kelly, B.T., Manela, A. and Xiu, D. "The structure of economic news." (No. w26648) National Bureau of Economic Research (2020).

Bushee, B.J. "Do institutional investors prefer near-term earnings over long-run value?." *Contemporary Accounting Research*, 18.2 (2001), 207-246.

Bushee, B.J. and Noe, C.F. "Corporate disclosure practices, institutional investors, and stock return volatility." *Journal of Accounting Research*, (2000), 171-202.

Chan, W.S. "Stock price reaction to news and no-news: drift and reversal after headlines." *Journal of Financial Economics*, 70.2 (2003), 223-260.

Cohen, L., and A. Frazzini. "Economic links and predictable returns." *Journal of Finance* 63 (2008), 1977–2011.

Conrad, J. and Kaul, G. "An anatomy of trading strategies." *The Review of Financial Studies*, 11.3 (1998), 489-519.

Cooper, M.J., Gutierrez Jr, R.C. and Hameed, A. "Market states and momentum." *The Journal of Finance*, 59.3 (2004), 1345-1365.

Da, Z., Engelberg, J. and Gao, P. "The sum of all FEARS investor sentiment and asset prices." *The Review of Financial Studies*, 28.1 (2015), 1-32.

Da, Z., Gurun, U. and Warachka, M. "Frog in the Pan: Continuous Information and Momentum." *The Review of Financial Studies*, 27.7 (2014), 2171-2218.

Daniel, K., Hirshleifer, D. and Subrahmanyam, A. "Investor psychology and security market under- and overreactions." *Journal of Finance*, 53.6 (1998), 1839-1885.

Electronic copy available at: https://ssrn.com/abstract=4207132

Daniel, K. and Moskowitz, T.J. "Momentum crashes." *Journal of Financial Economics*, 122.2 (2016), 221-247.

Davis, J.; E. Fama; and K. French. "Characteristics, covariances, and average returns: 1929-1997." *Journal of Finance* 55 (2000), 389–406.

Dougal, C., Engelberg, J., Garcia, D. and Parsons, C.A., 2012. "Journalists and the stock market." *The Review of Financial Studies*, 25(3), pp.639-679.

Ehsani, S. and Linnainmaa, J.T. "Factor momentum and the momentum factor." *Journal of Finance* (2021), Forthcoming.

Eisfeldt, A.L., Kim, E. and Papanikolaou, D. "Intangible Value." (2022) NBER Working Paper, (w28056).

Engelberg, J.E. and Parsons, C.A. "The causal impact of media in financial markets." *Journal of Finance*, 66.1 (2011), 67-97.

Engelberg, J., McLean, R.D. and Pontiff, J. "Anomalies and news." The *Journal of Finance*, 73.5 (2018), pp.1971-2001.

Fama, E., and K. French. "The cross section of expected stock returns." *Journal of Finance* 47 (1992), 427–465.

Fama, E.F. and French, K.R. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, 33.1 (1993), 3-56.

Fama, E.F. and French, K.R. "A five-factor asset pricing model." *Journal of Financial Economics*, 116.1 (2015), 1-22.

Fama, E.F. and MacBeth, J.D. "Risk, return, and equilibrium: Empirical tests." *Journal of Political Economy*, 81.3 (1973), 607-636.

Fang, L. and Peress, J. "Media coverage and the cross-section of stock returns." *The Journal of Finance*, 64.5 (2009), 2023-2052.

Fang, L.H., Peress, J. and Zheng, L. "Does media coverage of stocks affect mutual funds' trading and performance?." *The Review of Financial Studies*, 27.12 (2014), 3441-3466.

Frazzini, A., Israel, R. and Moskowitz, T.J. "Trading costs of asset pricing anomalies." (2012) Fama-Miller Working Paper, Chicago Booth Research Paper, (14-05).

Garcia, D. "Sentiment during recessions." *The Journal of Finance*, 68.3 (2013), 1267-1300.

Gebhardt, W.; S. Hvidkjaer; and B. Swaminathan. "Stock and bond market interaction: Does momentum spill over?" *Journal of Financial Economics* 75 (2005), 651–690.

Gervais, S.; R. Kaniel; and D. Mingelgrin. "The high volume return premium." *Journal of Finance* 56 (2001), 877–919.

Griffin, J.; X. Ji; and J. Martin. "Momentum investing and business cycle risk: Evidence from pole to pole." *Journal of Finance* 58 (2003), 2515–2547.

Grundy, B., and J. Martin. "Understanding the nature of the risks and the source of the rewards to momentum investing." *Review of Financial Studies* 14 (2001), 29–78.

Guest, N.M. "The information role of the media in earnings news." *Journal of Accounting Research*, 59.3 (2021), 1021-1076.

Hanley, K.W. and Hoberg, G. "Dynamic interpretation of emerging risks in the financial sector". *The Review of Financial Studies*, 32.12 (2019), 4543-4603.

Electronic copy available at: https://ssrn.com/abstract=4207132

Hillert, A., Jacobs, H. and Müller, S. "Media makes momentum." *The Review of Financial Studies*, 27.12 (2014), 3467-3501.

Hoberg, G., and G. Phillips. "Product market synergies and competition in mergers and acquisitions: A text-based analysis." *Review of Financial Studies* 23 (2010), 3773–3811.

Hoberg, G., and G. Phillips. "Text-based network industry classifications and endogenous product differentiation." *Journal of Political Economy* 124 (2016), 1423–1465.

Hoberg, G., Kumar, N., and N. Prabhala. "Mutual Fund Competition, Managerial Skill, and Alpha Persistence." *Review of Financial Studies* 31 (2018), 1896–1929.

Hong, H., and J. Stein. "A unified theory of underreaction, momentum trading, and overreaction in asset markets." *Journal of Finance* 54 (1999), 2134–2184.

Hou, K. "Industry information diffusion and the lead-lag effect in stock returns." *Review of Financial Studies* 20 (2007), 1113–1138.

Hou, K., Xue, C. and Zhang, L. "Digesting anomalies: An investment approach." *The Review of Financial Studies*, 28.3 (2015), 650-705.

Huberman, G. and Regev, T. "Contagious speculation and a cure for cancer: A nonevent that made stock prices soar." *The Journal of Finance*, 56.1 (2001), 387-396.

Jegadeesh, N., and S. Titman. "Returns to buying winners and selling losers: Implications for stock market efficiency." *Journal of Finance* 48 (1993), 65–91.

Jegadeesh, N., and S. Titman. "Profitability of momentum strategies: an evaluation of alternative explanations." *Journal of Finance* 56 (2001), 699–720.

Jegadeesh, N., and S. Titman. "Momentum." *Annual Review of Financial Economics* 3 (2011), 493–509.

Jeon, Y., McCurdy, T.H. and Zhao, X. "News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies." *Journal of Financial Economics*.

Jiang, H., Li, S.Z. and Wang, H. "Pervasive underreaction: Evidence from high-frequency data." *Journal of Financial Economics*, 141.2 (2021), 573-599.

Johnson, T.C. "Rational momentum effects." *The Journal of Finance*, 57.2 (2002), 585-608.

Korajczyk, R.A. and Sadka, R. "Are momentum profits robust to trading costs?." *The Journal of Finance*, 59.3 (2004), 1039-1082.

Kacperczyk, M., Sialm, C., and L. Zheng. "On the Industry Concentration of Actively Managed Equity Mutual Funds." *Journal of Finance* 60 (2005), 1983–2011.

Koijen, R.S. and Yogo, M. "A demand system approach to asset pricing." *Journal of Political Economy*, 127.4 (2019), 1475-1515.

Latane, H.A. and Jones, C.P. "Standardized unexpected earnings–1971-77." *The journal of Finance*, 34.3 (1979), 717-724.

Lesmond, D.A., Schill, M.J. and Zhou, C. "The illusory nature of momentum profits." *Journal of Financial Economics*, 71.2 (2004), 349-380.

Li, K., Mai, F., Shen, R. and Yan, X. "Measuring corporate culture using machine learning." *The Review of Financial Studies*, 34.7 (2021), 3265-3315.

Linnainmaa, J., and Michael R. "The history of the cross-section of stock returns." *Review of Financial Studies* 31.7 (2018): 2606-2649.

Electronic copy available at: https://ssrn.com/abstract=4207132

Ljungqvist, A. and Qian, W., 2016. "How constraining are limits to arbitrage?." *Review of Financial Studies*, 29(8), pp.1975-2028.

Loughran, T. and McDonald, B. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." *Journal of Finance*, 66.1 (2011), pp.35-65.

Mclean, D., and J. Pontiff. "Does academic research destroy stock return predictability." *Journal of Finance* 71 (2016), 5–32.

Menzley, L., and O. Ozbas. "Market segmentation and cross-predictability of returns." *Journal of Finance* 65 (2010), 1555–1580.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. "Efficient estimation of word representations in vector space." (2013) arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. "Distributed representations of words and phrases and their compositionality." (2013) Advances in neural information processing systems, 26.

Moskowitz, T., and M. Grinblatt. "Do industries explain momentum." *Journal of Finance* 54 (1999), 1249–1290.

Peress, J. "Media coverage and investors' attention to earnings announcements." (2008) Available at SSRN 2723916.

Peress, J. "The media and the diffusion of information in financial markets: Evidence from newspaper strikes." *Journal of Finance*, 69.5 (2014), 2007-2043.

Rouwenhorst, G. "International momentum strategies." *Journal of Finance* 53 (1998), 267–284.

Rouwenhorst, G. "Local return factors and turnover in emerging stock markets." *Journal of Finance* 54 (1999), 1439–1464.

Salton, G. and McGill, M.J. "Introduction to modern information retrieval." mcgraw-hill (1983).

Sagi, J.S. and Seasholes, M.S. "Firm-specific attributes and the cross-section of momentum." *Journal of Financial Economics*, 84.2 (2007), 389-434.

Solomon, D.H., Soltes, E. and Sosyura, D. "Winners in the spotlight: Media coverage of fund holdings as a driver of flows." *Journal of Financial Economics*, 113.1 (2014), 53-72.

Soo, C.K. "Quantifying sentiment with news media across local housing markets." *The Review of Financial Studies*, 31.10 (2018), 3689-3719.

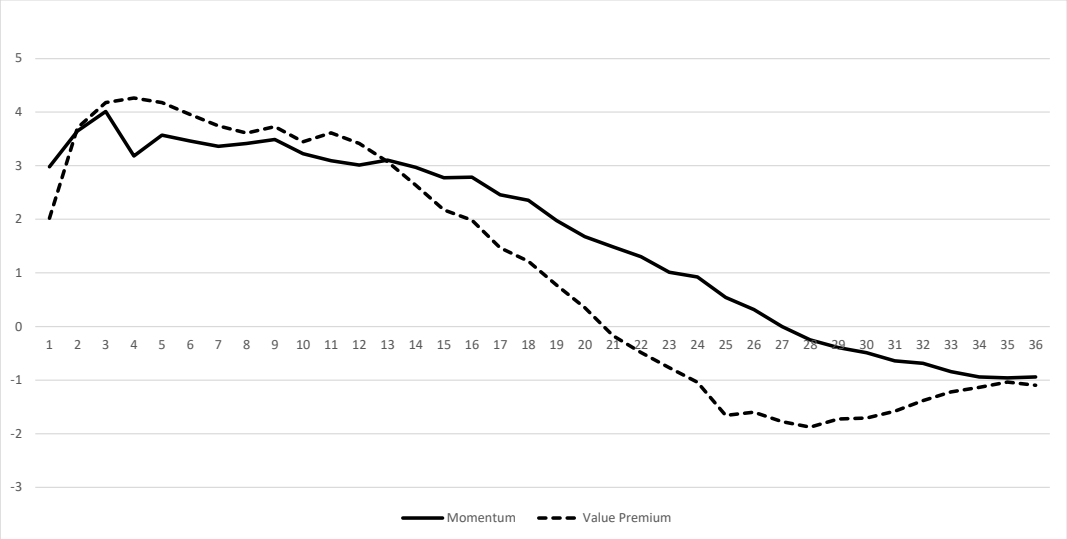
Tetlock, P.C. "Giving content to investor sentiment: The role of media in the stock market." *Journal of finance*, 62.3 (2007), 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. "More than words: Quantifying language to measure firms' fundamentals." *Journal of Finance*, 63.3 (2008), 1437-1467.

Tetlock, P.C. "Does public financial news resolve asymmetric information?." *The Review of Financial Studies*, 23.9 (2010), 3520-3557.

Electronic copy available at: https://ssrn.com/abstract=4207132

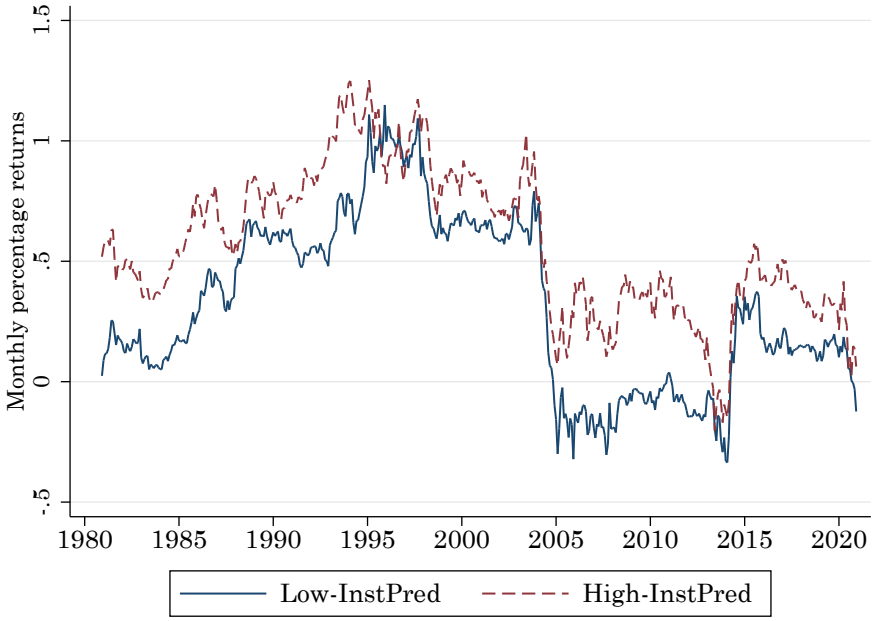
Figure 1: **Effect of WSJ InstPred on Anomalies by Measurement Window.** The figure displays the economic strength of the signal from the WSJ "institutional investor prediction" (InstPred) as we change the window for constructing the InstPred measure. See Section 2.2.2 for more details on constructing InstPred and measurement window. The solid black line below reports the coefficient of the t-statistic of the interaction term $PastRet \times InstPred$ in Fama-MacBeth regressions in Section 3.1.1 as we increase its measurement window for $InstPred$ from 1 month to 36 months. The dotted line reports analogous t-statistics for the interaction term $BM \times InstPred$.



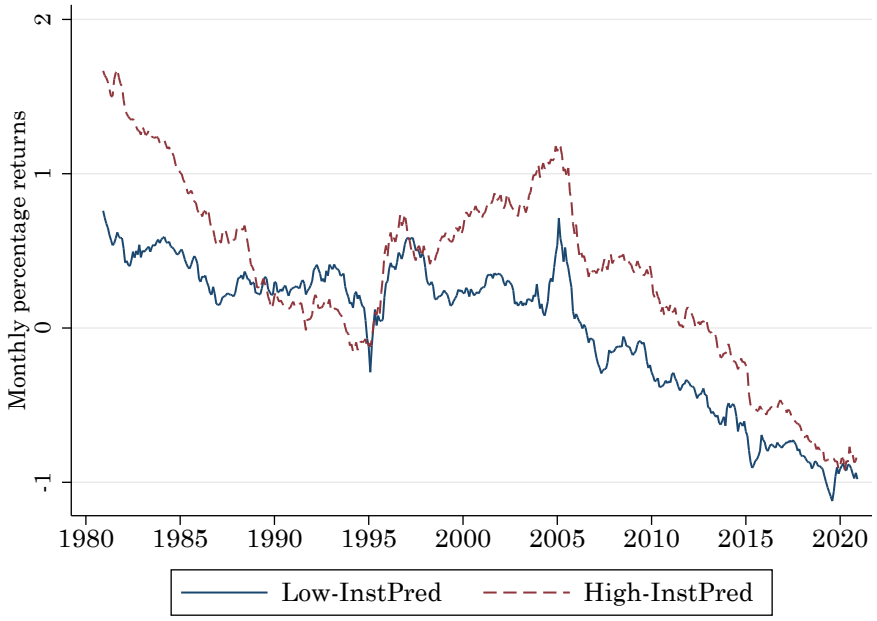
Electronic copy available at: https://ssrn.com/abstract=4207132

Figure 2: **Anomaly Returns in Low-InstPred and High-InstPred Portfolios.** The figure plots the smoothed monthly percentage returns of the momentum anomaly (in Panel A) and the value anomaly (in Panel B) in the Low-InstPred and High-InstPred portfolios. See Table 9 for details of the portfolio formation. Each point in the line represents the average quantity for a ten-year window centered around date indicated by the x-axis (Linnainmaa and Roberts (2018)). For example, the point on June 1990 represents the average return from July 1985 through June 1995.

Panel A: Momentum

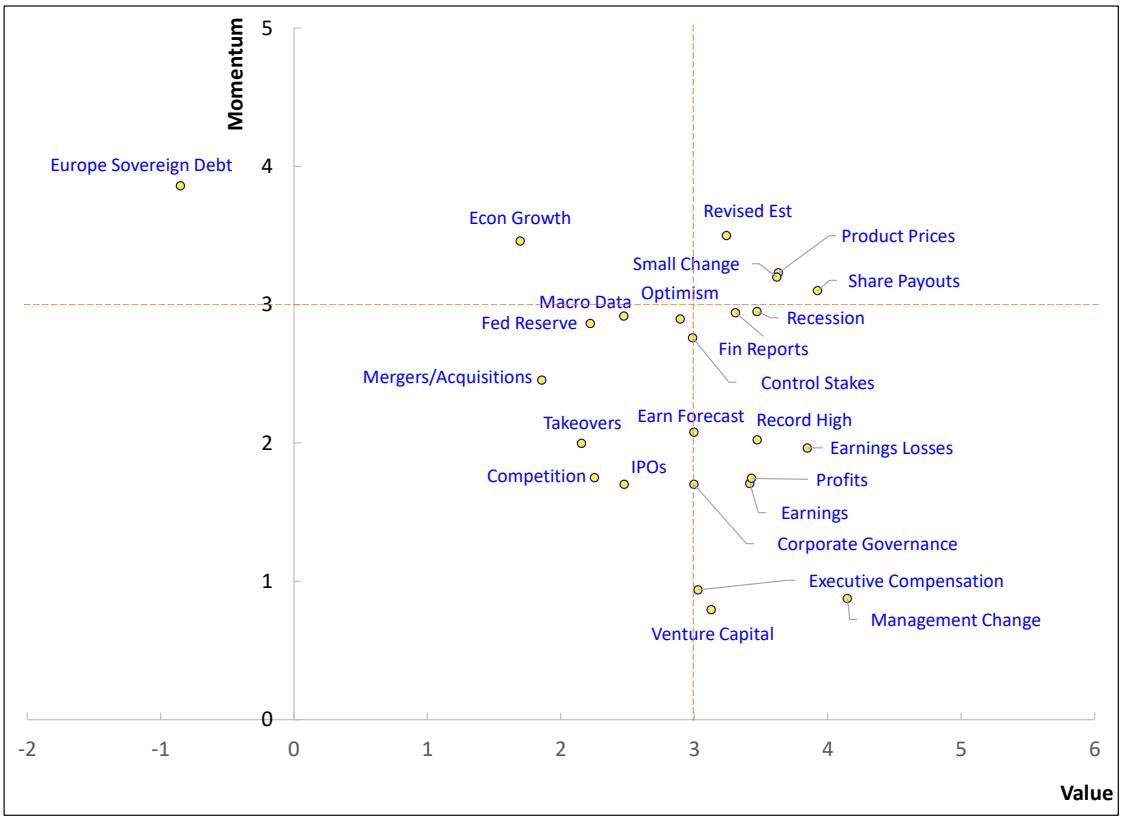


Panel B: Value



Electronic copy available at: https://ssrn.com/abstract=4207132

Figure 3: **WSJ Content Themes and the InstPred Effect on Anomalies.** The figure displays pairs of t-statistics of the triple interaction term between anomaly, our main measure of institutional investor prediction (InstPred), and a WSJ content theme in Fama-MacBeth regressions. We construct 25 WSJ content theme measures for each FF-48 industry based on word lists from Bybee, Kelly, Manela, and Xiu (2020). Then, we interact each content theme, one at a time, with variables in our main Fama-MacBeth regression in Table 6. x-axis represents the t-statistics of the triple interaction term between book-to-market ratio, InstPred, and the content theme, while y-axis represents the t-statistics of the triple interaction term between past returns, InstPred, and the content theme.



Electronic copy available at: https://ssrn.com/abstract=4207132

Table 1: **Top 50 Keywords for "Institutional Investor" and "Prediction"**

This table lists the top 50 keywords with the highest similarity to our seed words "institutional investor" and "prediction", respectively, from Google word2vec and also appear in the Wall Street Journal articles. See Section 2.2.1 for more details. Internet Appendix C provides all the 250 keywords that we use to compute a WSJ article's relevance to "institutional investor" and "prediction."

Rank	Keywords for "institutional investor"	Keywords for "prediction"
1	institutional_investor	prediction
2	fixed_income	predictions
3	morningstar	predicting
4	morgan_stanley	forecast
5	lipper	forecasts
6	portfolio_manager	forecasting
7	fortune_magazine	projections
8	brokerage_firms	projection
9	merrill_lynch	predicted
10	private_equity	estimation
11	hedge_fund	estimate
12	investment_banking	guesses
13	emerging_markets	assertion
14	credit_suisse	predict
15	hedge_funds	estimates
16	jpmorgan	assumption
17	zacks	expectation
18	institutional_investors	prophecy
19	investor	hunch
20	gabelli	predicts
21	brokerage	prognosis
22	brokerages	assertions
23	goldman_sachs	calculations
24	equities	forecasted
25	clsa	assessment
26	blackrock	probability
27	asset_allocation	belief
28	factset	outlook
29	barclays_capital	forecaster
30	capital_markets	estimating
31	piper_jaffray	pronouncement
32	banc	expectations
33	mutual_fund	conventional_wisdom
34	analyst	theory
35	smith_barney	hypothesis
36	banker	forecasters
37	mutual_funds	observations
38	oppenheimer	scenario
39	quantitative	suggestion
40	dealogic	conjecture
41	nomura	conclusions
42	high_yield	overly_optimistic
43	global	assumptions
44	magazine	pessimistic
45	bear_stearns	recommendation
46	deutsche_bank	observation
47	legg_mason	notion
48	forbes_magazine	analogy
49	csfb	calculation
50	citigroup	projecting

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 2: **WSJ Institutional Investor Measure and Name Mentions**

This table reports regressions of our WSJ institutional investor variable (*Inst*) on an array of measures of an article's mentioning of institutional investors' names. See the definition of Inst in Section 2.2.1. The name measures are constructed using the Thomson-Reuters Institutional (13F) Holdings database (for large institutional investors), the CRSP Mutual Fund database (for mutual funds), the Thomson/Refinitiv Lipper Hedge Fund database (for hedge funds), and Corporate Finance Institute (for investment banks). For each name list, we count the occurrence of the names from the list in an article. All name measures are standardized to mean 0 and standard deviation of 1. We include year and FF-48 industry fixed effects. t-statistics are clustered at the industry level and reported in parentheses. Our sample spans January 1981 to December 2020.

	"Institutional investor" theme				
	(1)	(2)	(3)	(4)	(5)
13F Institutional names	0.437***				
	(15.53)				
Mutual fund names		0.423***			0.256***
		(10.51)			(7.94)
Hedge fund names			0.375***		0.156***
			(7.02)		(2.69)
Investment bank names				0.384***	0.139***
				(11.21)	(2.82)
R^2	0.353	0.340	0.308	0.318	0.380
Observations	1,018,718	1,018,718	1,018,718	1,018,718	1,018,718

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 3: **Sample WSJ Articles for Institutional Investor Prediction**

This table shows a sample of WSJ articles that have high "Institutional Investor & Prediction" (InstPred) score. See details in Section 2.2.1. We highlight words related to institutional investors, prediction, and industry sectors.

Example 1: The Momentum Game Has Returned to the Stock Market, 2018-01-16

"Forget fundamentals: Momentum is back in the stock market. ... The bullish explanation is that it takes time for investors to price in a new environment. ... **Goldman Sachs's chief U.S. equity strategist**, David Kostin, said profit **forecasts** for the entire S&P 500 produced by strategists such as himself are, unusually, higher than the sum of individual company **forecasts** partly because analysts haven't yet included tax cuts. ... the current momentum portfolio perfectly captures today's consensus: **heavily overweight banks** (for interest-rate rises and deregulation) and **technology companies** (for low-inflationary growth); heavily underweight **real estate** (hurt by higher rates) and **consumer staples** (who needs downside protection?) ..."

Example 2: Einhorn Hits Fracking Stocks, 2015-05-05

"David Einhorn, an outspoken **hedge-fund manager**, took aim at the hard-hit **hydraulic-fracturing industry** Monday, when he **unveiled** bearish views on companies such as Pioneer Natural Resources Co. and Concho Resources Inc., which are under pressure from falling oil prices and environmental concerns ... Investment in shale fracking companies will "contaminate" investment returns, said Mr. Einhorn, founder of $12 billion **Greenlight Capital Inc.** ... "

Example 3: Try 'Made in the U.S.A.' as a 2014 Investing Theme, 2013-12-04

"The U.S. economy is recovering slowly but steadily. While productivity and earnings growth aren't **expected** to soar next year, some surprising recent developments have shifted U.S. industry into higher gear—especially in the **factories** and **oil fields** of America's heartland. 'Middle America,' says Liz Ann Sonders, chief investment strategist at **Charles Schwab Co.**, 'is our favorite emerging market.' ... "

Example 4: REITs Seek to Diversify Business to Sustain Their Growth, 1997-12-31

"Continuing to grow may no longer make as much sense, some analysts **warn**. 'As acquisition prices get higher, when are we going to see companies buy back their stock or take their dividends up to 100% of their cash flow?' asks Gregory Whyte, an analyst with **Morgan Stanley** Dean Witter. 'That's what they should do if **properties** cost shareholders more than they add to cash flow.' ... "

Example 5: Chip Analyst Turns Heads With Flip-Flops, 1997-08-14

"Early this month, **Merrill Lynch's** semiconductor analyst, Thomas Kurlak, reported in two notes to clients that the 'best of all hoped-for scenarios' was developing for **semiconductor** stocks. He **forecast** a 20% move up for the group ..."

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 4: **Summary Statistics of WSJ Articles**

Panel A reports the summary statistics of variables in 1,018,718 Wall Street Journal articles from June 1979 to December 2020. Panel B reports the Pearson correlation coefficients of the variables. Each variable is constructed based on the cosine similarity between the full content of a WSJ article and the keywords for the variable. For *Institutional Investor* and *Prediction*, we use the top 250 synonyms "institutional_investor" and the top 250 synonyms of "prediction" from Google word2vec model, respectively. For *Positive Tone, Negative Tone*, and *Uncertainty*, we use the keywords provided by Loughran and McDonald (2011). *Institutional Investor & Prediction (InstPred)* is the product of *Institutional Investor* and *Predict* multiplied by 100 for the ease of reading. See Section 2.2.1 for more details.

Panel A: Summary Statistics

Variable	Mean	Std.Dev.	Minimum	Median	Maximum	Obs.
Institutional Investor & Prediction (InstPred)	0.010	0.020	0.000	0.000	0.487	1,018,718
Institutional Investor (Inst)	0.010	0.012	0.000	0.006	0.127	1,018,718
Prediction (Pred)	0.008	0.008	0.000	0.006	0.076	1,018,718
Positive Tone	0.008	0.008	0.000	0.007	0.076	1,018,718
Negative Tone	0.009	0.008	0.000	0.007	0.070	1,018,718
Uncertainty	0.006	0.007	0.000	0.005	0.059	1,018,718

Panel B: Pearson Correlation Coefficients

	Institutional Investor & Prediction	Institutional Investor	Prediction	Positive Tone	Negative Tone
Institutional Investor	0.675				
Prediction	0.687	0.287			
Positive Tone	0.333	0.234	0.379		
Negative Tone	0.202	0.102	0.335	0.259	
Uncertainty	0.339	0.210	0.467	0.369	0.423

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 5: **Summary Statistics of Firms**

This table reports the summary statistics of our key variables at the stock-month level from January 1981 to December 2020. *Monthly Return* is current month t's stock return. Our main variable of interest *InstPred* is based on the text of newspaper articles from the Wall Street Journal. Wall Street Journal theme variables are first computed at the article-level, and are based on cosine similarities between each article's text and a word list corresponding to each theme. Thematic word lists are obtained from the Google word2vec embeddings database, and tone word lists are from Loughran and McDonald (2011). *InstPred* is the intensity of WSJ articles mentioning institutional investors and predict from month t-4 to t-1 for the stock's FF-48 industry, standardized relative to months t-24 to t-13 (see Section 2 for more details). *Article* is the number of WSJ articles about the stocks' industry over the past 3 months, also standardized relative to the industry's article counts from months t-24 to t-13. *BM* is the natural logarithm of book-to-market ratio. *PastRet* is the return from $t-12$ to $t-1$. *Size* is the natural logarithm of market capitalization as of June. *Investment* is the growth rate of total assets. *Profitability* is the operating profitability defined following Fama and French (2015). *SUE* is earnings surprise multiplied by 100 for the ease of reading.

Panel A: Summary Statistics

Variable	Mean	Std.Dev.	Minimum	Median	Maximum	Obs.
Monthly return	0.011	0.162	-0.981	0.001	19.884	1,936,537
InstPred	0.161	0.811	-3.092	0.132	3.276	1,936,537
Articles	0.073	1.124	-2.874	-0.045	3.284	1,936,537
BM	-0.598	0.936	-11.308	-0.506	5.685	1,936,537
PastRet	0.161	0.729	-0.996	0.064	98.571	1,936,537
Size	12.236	2.151	4.676	12.093	21.170	1,936,537
Investment	0.148	0.383	-0.587	0.059	5.307	1,936,537
Profitability	0.143	0.428	-5.370	0.199	2.705	1,936,537
SUE	-0.123	1.634	-48.858	0.000	20.568	1,936,537

Panel B: Pearson Correlation Coefficients

	Return	InstPred	Articles	BM	PastRet	Size	Investment	Profitability
InstPred	-0.007							
Articles	-0.007	0.126						
BM	0.027	-0.025	-0.047					
PastRet	0.004	-0.060	0.002	0.024				
Size	-0.006	-0.038	-0.088	-0.271	-0.002			
Investment	-0.021	0.046	0.029	-0.158	-0.051	0.067		
Profitability	0.015	0.024	-0.002	0.092	-0.021	0.236	0.065	
SUE	0.010	-0.028	-0.001	0.026	0.113	-0.019	-0.053	-0.060

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 6: **Fama-MacBeth Regressions of Anomalies and WSJ InstPred**

This table reports Fama-MacBeth regression of next-period monthly stock returns on the interaction between anomaly predictors (*PastRet* and *BM*) and WSJ institutional investor prediction measure (*InstPred*). Monthly stock returns are annualized by multiplying 1,200 for ease of interpretation. See Table 5 for variable definitions. All non-interactive independent variables are standardized to have mean 0 and standard deviation of 1. *t*-statistics are adjusted using Newey-West with two lags and reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from January 1981 to December 2020.

	(1)	(2)	(3)	(4)	(5)
PastRet×InstPred	1.94***	1.57***			
	(3.42)	(3.06)			
BM×InstPred			1.38***	1.40***	
			(3.61)	(4.01)	
InstPred	0.16	0.42	-0.18	0.03	
	(0.31)	(0.95)	(-0.32)	(0.07)	
PastRet	3.81***	2.51*		2.40*	2.55*
	(2.64)	(1.89)		(1.83)	(1.93)
BM		2.79***	4.08***	3.10***	2.78***
		(4.17)	(5.22)	(4.45)	(4.11)
Size		-0.64		-0.63	-0.68
		(-0.75)		(-0.73)	(-0.79)
Investment		-2.55***		-2.56***	-2.59***
		(-7.76)		(-7.76)	(-7.78)
Profitability		3.56***		3.60***	3.58***
		(5.08)		(5.12)	(5.06)
SUE		3.26***		3.27***	3.27***
		(14.17)		(14.11)	(14.02)
Articles		-0.72		-0.72	-0.62
		(-1.35)		(-1.35)	(-1.12)
Observations	1,936,537	1,936,537	1,936,537	1,936,537	1,936,537

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 7: **Fama-MacBeth Regressions Using Permutations of WSJ InstPred**

This table reports Fama-MacBeth regression of next-period monthly stock returns on the interaction between anomaly predictors (*PastRet* and *BM*) and permutations of the WSJ institutional investor prediction measures. Monthly stock returns are annualized by multiplying 1200 for ease of interpretation. Panel A reports results when we use the institutional investor theme (*Inst*) alone or we us the prediction theme (*Pred*) alone. Panel B reports results when we develop WSJ themes aimed at isolating separate effects from the institutional investor theme and the prediction theme. The first two columns are based on Inst&HiPred, which is the institutional investor theme loading for the given article multiplied by a dummy regarding if the article has an above median value for the predict theme relative to other artcles from the same month. The third and fourth columns are analogously defined as the institutional investor theme multiplied by the below median predict theme dummy. The fifth and sixth columns are analogously defined as the predict theme multiplied by the above median institutional investor theme dummy. The final two columns are analogously defined as the predict theme multiplied by the below median institutional investor theme dummy. All regressions control for stock characteristics including size, investment, profitability, SUE, and articles. See Table 5 for variable definitions. t-statistics are adjusted using Newey-West with two lags and reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from January 1981 to December 2020. There are 1,936,537 observations in each column.

Panel A: WSJ Institutional Investor and Prediction Alone

WSJ:	InstPred		Inst		Pred	
	(1)	(2)	(3)	(4)	(5)	(6)
PastRet× *WSJ*	1.57***		0.99**		1.08**	
	(3.06)		(2.07)		(2.46)	
BM× *WSJ*		1.40***		1.15***		1.17***
		(4.01)		(3.59)		(3.57)
WSJ	0.42	0.03	0.41	0.05	0.78	0.48
	(0.95)	(0.07)	(0.89)	(0.11)	(1.48)	(0.87)
BM	2.79***	3.10***	2.79***	3.32***	2.81***	2.89***
	(4.17)	(4.45)	(4.16)	(4.76)	(4.25)	(4.37)
PastRet	2.51*	2.40*	2.75**	2.43*	2.65**	2.42*
	(1.89)	(1.83)	(2.06)	(1.85)	(2.01)	(1.83)
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: WSJ Institutional Investor and Prediction Permutations

WSJ:	Inst&HiPred		Inst&LoPred		Pred&HiInst		Pred&LoInst	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PastRet× *WSJ*	1.48***		0.08		1.43***		-0.29	
	(2.89)		(0.16)		(3.18)		(-0.65)	
BM× *WSJ*		1.47***		-0.37		1.23***		-0.57**
		(4.58)		(-1.24)		(3.81)		(-2.00)
WSJ	0.60	0.17	0.28	0.17	0.72*	0.33	0.11	0.27
	(1.33)	(0.36)	(0.68)	(0.39)	(1.65)	(0.70)	(0.18)	(0.43)
BM	2.78***	3.01***	2.79***	3.04***	2.77***	2.81***	2.82***	2.89***
	(4.15)	(4.31)	(4.18)	(4.53)	(4.14)	(4.18)	(4.29)	(4.41)
PastRet	2.78**	2.40*	2.64**	2.50*	2.75**	2.42*	2.87**	2.44*
	(2.08)	(1.83)	(1.98)	(1.90)	(2.09)	(1.84)	(2.11)	(1.85)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 8: **Fama-MacBeth Regressions Controlling for Other News Themes**

This table reports Fama-MacBeth regression of next-period monthly stock returns on the interaction between anomaly predictors (*PastRet* and *BM*) and WSJ institutional investor prediction measure (*InstPred*) controlling for other news themes. Monthly stock returns are annualized by multiplying 1,200 for ease of interpretation. Articles is the number of WSJ articles for the stocks' industry during months t-3 to t-1 (standardized relative to months t-24 to t-13). *Positive Tone*, *Negative Tone*, and *Uncertainty* are all based on the Loughran and McDonald (2011) dictionaries, and are based on months t-3 to t-1, and are also standardized relative to months t-24 to t-13. All regressions control for stock characteristics including size, investment, profitability, SUE, and articles. See Table 5 for variable definitions. *t*-statistics are adjusted using Newey-West with two lags and reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from January 1981 to December 2020. There are 1,936,537 observations in each column.

OtherTheme:	Articles		Positive Tone		Negative Tone		Uncertainty	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PastRet×InstPred	1.41***		1.44***		1.39***		1.48***	
	(2.73)		(2.66)		(2.72)		(2.68)	
PastRet×*OtherTheme*	0.33		0.40		0.88		-0.03	
	(0.57)		(0.56)		(1.40)		(-0.05)	
BM×InstPred		1.49***		1.34***		1.21***		1.37***
		(4.17)		(3.67)		(3.51)		(3.67)
BM×*OtherTheme*		-0.33		0.09		0.27		0.32
		(-0.86)		(0.20)		(0.69)		(0.76)
InstPred	0.45	-0.00	0.02	-0.36	0.46	0.06	0.09	-0.23
	(1.04)	(-0.01)	(0.03)	(-0.71)	(1.04)	(0.14)	(0.19)	(-0.44)
OtherTheme	-0.96*	-0.60	1.19*	1.05	-0.05	0.01	0.90	0.89
	(-1.79)	(-1.09)	(1.81)	(1.50)	(-0.08)	(0.02)	(1.36)	(1.30)
PastRet	2.63*	2.39*	2.44*	2.36*	2.79**	2.37*	2.35*	2.33*
	(1.80)	(1.82)	(1.80)	(1.81)	(2.34)	(1.82)	(1.75)	(1.77)
BM	2.77***	3.33***	2.86***	3.18***	2.85***	3.13***	2.84***	3.18***
	(4.14)	(4.68)	(4.35)	(4.48)	(4.43)	(4.49)	(4.33)	(4.65)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 9: **Portfolio Sorts on Anomalies and WSJ InstPred**

This table reports excess returns of portfolios sorted on anomaly predictors and our main variable WSJ InstPred. Panel A reports the results of using past returns as the predictor of the momentum anomaly, and Panel B reports the results using book-to-market as the predictor of the value anomaly. Each month, we sort stocks into two size groups based on NYSE median market capitalization. Independently, we sort stocks into three groups by NYSE anomaly predictors (past returns from $t-12$ to $t-1$ in Panel A and book-to-market in Panel B). Also independently, we sort firms into three WSJ InstPred groups based on their NYSE breakpoints. We next compute value-weighted excess returns within each of the 18 portfolios and then take simple averages of the returns between large- and small-cap portfolios within each of the 3×3 anomaly-InstPred portfolios. Monthly excess returns are annualized by multiplying 1,200. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from January 1981 to December 2020.

	Panel A: Momentum Anomaly			
	Low PastRet	Med PastRet	High PastRet	H−L
Low InstPred	6.59*	8.27***	10.29***	3.70
	(1.88)	(3.27)	(3.57)	(1.50)
Med InstPred	7.16**	9.39***	12.22***	5.06**
	(2.02)	(3.73)	(4.22)	(2.16)
High InstPred	3.52	8.48***	10.35***	6.84***
	(0.97)	(3.38)	(3.65)	(2.70)
	Panel B: Value Anomaly			
	Low BM	Med BM	High BM	H−L
Low InstPred	7.76**	9.38***	7.91***	0.16
	(2.58)	(3.60)	(2.79)	(0.08)
Med InstPred	8.72***	10.01***	10.71***	1.99
	(2.90)	(3.70)	(3.67)	(1.15)
High InstPred	5.58*	9.14***	9.89***	4.31**
	(1.83)	(3.44)	(3.48)	(2.21)

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 10: **The Role of Institutional Investors' WSJ Connectedness**

This table reports our baseline Fama-MacBeth regression (in Table 6) for two subsamples based on industries' major institutional investors' average connectedness with the WSJ, i.e., *Investor-WSJ Connectedness*. See Section 4.1 for details regarding this variable's definition. Crucially, connectedness for each industry's major investors is computed using past interactions with the WSJ occurring in unrelated industries, ensuring that this variable is plausibly exogenous relative to the focal industry's state in the given month. In each month, we divide the sample into two groups based on the median of the industries' major investor WSJ-connectedness, resulting in *Industries with Low Investor-WSJ Connectedness* and *Industries with High Investor-WSJ Connectedness*. All non-interactive independent variables are standardized to have mean 0 and standard deviation of 1. *t*-statistics are adjusted using Newey-West with two lags and are reported in parentheses. Column (9) displays the statistical differences in the coefficients of the *Anomaly*×*InstPred* term between the two samples for each regression specification, i.e., between Columns (1) and (5), and between (2) and (6) in the first row, and between (3) and (7), and (4) and (8) in the second row. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from July 1981 to June 2019.

	Industries with Low Investor-WSJ Connectedness				Industries with High Investor-WSJ Connectedness				Coefficient Differences
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
PastRet×InstPred	0.03	-0.12			2.87***	2.18**			**/**
	(0.04)	(-0.20)			(2.76)	(2.35)			
BM×InstPred			0.64	0.70*			2.44***	2.42***	**/***
			(1.53)	(1.80)			(4.32)	(4.69)	
InstPred	0.01	0.06	0.46	0.32	0.49	0.83	-0.43	-0.11	
	(0.02)	(0.12)	(0.81)	(0.65)	(0.45)	(0.93)	(-0.41)	(-0.12)	
PastRet	4.04***	2.81**		2.43*	3.76**	2.08		2.61*	
	(2.64)	(2.00)		(1.74)	(2.29)	(1.35)		(1.82)	
BM		2.78***	4.30***	3.35***		3.02***	3.93***	3.31***	
		(4.35)	(5.74)	(4.96)		(4.21)	(4.37)	(4.28)	
Size		-0.66		-0.65		0.03		0.05	
		(-0.73)		(-0.72)		(0.04)		(0.06)	
Investment		-2.82***		-2.83***		-2.60***		-2.61***	
		(-7.37)		(-7.41)		(-6.43)		(-6.38)	
Profitability		4.33***		4.41***		3.00***		3.03***	
		(5.60)		(5.69)		(4.30)		(4.30)	
SUE		3.28***		3.31***		3.30***		3.30***	
		(11.81)		(11.88)		(13.69)		(13.71)	
Articles		-0.58		-0.58		-1.13		-1.09	
		(-1.03)		(-1.04)		(-1.19)		(-1.15)	
Observations	989,722	989,722	989,722	989,722	868,266	868,266	868,266	868,266	

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 11: **Changes in Institutional Holdings**

This table reports panel regression results of quarterly changes in institutional holdings from month $t-2$ to $t+1$ on the interaction between current anomaly predictors (*PastRet* and *BM*) and current WSJ institutional investor predict measure (*InstPred*). Institutional ownership of a stock is the ratio between shares held by institutional investors from the Thomson-Reuters Institutional Holdings (13F) database and the total shares outstanding from the CRSP database in basis points. *All* represents changes in ownership from all non-index institutional investors. *HiActive* and *LoActive* represent changes in ownership only from high-activity institutional investors and only from low-activity institutional investors, respectively. See Section 4 for more details. All non-interactive independent variables are standardized to have mean zero and standard deviation of one for each regression. All regressions control for quarter fixed effects. t-statistics clustered at stock level are presented in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from March 1981 to December 2018.

	All	HiActive	LoActive	All	HiActive	LoActive
	(1)	(2)	(3)	(4)	(5)	(6)
PastRet×InstPred	2.67**	2.93***	-0.26			
	(2.32)	(2.62)	(-1.12)			
BM × InstPred				1.18**	1.66***	-0.47
				(2.02)	(3.20)	(-1.63)
PastRet	35.61***	36.97***	-1.35***	35.22***	36.54***	-1.32***
	(21.56)	(21.19)	(-4.51)	(20.30)	(20.00)	(-4.43)
BM	5.96***	6.85***	-0.88***	5.95***	6.85***	-0.89***
	(9.19)	(12.95)	(-2.85)	(9.15)	(12.92)	(-2.87)
InstPred	-1.34**	-1.46***	0.11	-1.36**	-1.47***	0.11
	(-2.29)	(-2.68)	(0.43)	(-2.33)	(-2.72)	(0.44)
Size	-7.07***	-6.14***	-0.93***	-7.06***	-6.13***	-0.93***
	(-17.73)	(-17.81)	(-4.93)	(-17.72)	(-17.79)	(-4.93)
Investment	-6.08***	-6.73***	0.65**	-6.07***	-6.72***	0.65**
	(-9.07)	(-11.25)	(2.55)	(-9.05)	(-11.22)	(2.53)
Profitability	-1.69***	-1.53***	-0.16	-1.63***	-1.47***	-0.16
	(-3.11)	(-3.39)	(-0.53)	(-3.00)	(-3.26)	(-0.53)
SUE	9.79***	11.44***	-1.65***	9.85***	11.51***	-1.66***
	(16.89)	(20.92)	(-6.80)	(16.91)	(20.92)	(-6.84)
Articles	-1.31**	-1.20**	-0.10	-1.41**	-1.33**	-0.08
	(-2.13)	(-2.21)	(-0.35)	(-2.30)	(-2.45)	(-0.27)
Observations	592,072	592,072	592,072	592,072	592,072	592,072

Electronic copy available at: https://ssrn.com/abstract=4207132

Table 12: **WSJ Content Themes and InstPred**

This table reports regressions of our WSJ InstPred variable and its two components institutional investor variable (*Inst*) and prediction variable (*Pred*) on an array of text-based content themes at the article level. See Table 1 for definitions of our WSJ variables. The content themes are derived using the word lists from Bybee, Kelly, Manela, and Xiu (2020). For each content theme, we compute the cosine similarity of the article's text and the word lists associated with the theme. We include year and FF48 industry fixed effects. t-statistics are clustered at the industry level and reported in parentheses. Our sample spans January 1981 to December 2020.

Theme	InstPred	Inst	Pred
Share payouts	0.435 (38.9)	0.256 (31.0)	0.064 (16.0)
IPOs	0.187 (23.4)	0.402 (11.9)	-0.015 (-4.10)
Earnings forecast	0.710 (17.8)	0.198 (24.3)	0.320 (71.5)
Record high	0.134 (15.1)	0.093 (13.3)	0.014 (3.53)
Recession	0.257 (14.1)	0.120 (14.9)	0.066 (27.3)
Optimism	0.294 (11.6)	0.029 (4.20)	0.171 (44.1)
Corporate governance	0.039 (9.38)	0.035 (8.13)	-0.001 (-0.40)
Venture capital	0.067 (6.20)	0.087 (6.61)	-0.001 (-0.26)
Revised estimate	0.068 (4.23)	-0.024 (-4.01)	0.134 (36.8)
Financial reports	0.019 (3.82)	-0.001 (-1.00)	0.029 (11.5)
Federal Reserve	0.132 (3.55)	0.032 (1.11)	0.038 (4.72)
Macroeconomic data	0.051 (2.72)	-0.027 (-2.84)	0.080 (14.1)
Competition	0.051 (2.26)	0.037 (2.60)	0.061 (7.83)
Takeovers	0.009 (2.10)	-0.047 (-4.53)	0.037 (19.8)
Mergers & Acquisitions	0.018 (2.05)	0.029 (2.85)	-0.012 (-4.35)
Management changes	0.014 (1.64)	0.020 (2.75)	0.003 (1.47)
European sovereign debt	0.012 (0.91)	0.081 (4.10)	0.023 (3.84)
Control stakes	0.001 (0.03)	0.050 (2.46)	-0.021 (-7.90)
Earnings losses	-0.000 (-0.02)	0.009 (3.15)	-0.016 (-4.83)
Small changes	-0.003 (-0.42)	0.037 (3.74)	-0.021 (-4.55)
Executive compensation	-0.013 (-1.08)	-0.005 (-1.08)	-0.003 (-0.95)
Economic growth	-0.101 (-7.45)	-0.049 (-4.39)	-0.012 (-2.25)
Product prices	-0.062 (-7.29)	-0.045 (-4.93)	0.006 (1.75)
Earnings	-0.121 (-9.19)	-0.057 (-8.47)	-0.032 (-10.1)
Profits	-0.264 (-16.7)	-0.093 (-16.2)	-0.124 (-23.8)
R^2	0.464	0.441	0.526
Observations	1,018,672	1,018,672	1,018,672

Electronic copy available at: https://ssrn.com/abstract=4207132

Internet Appendix

Wisdom of the Institutional Crowd: Implications for Anomaly Returns

AJ Chen, Gerard Hoberg, and Miao Ben Zhang

Electronic copy available at: https://ssrn.com/abstract=4207132

A Details on Constructing the InstPred Measure

A.1 Technical Details on Measuring Institutional Investor and Prediction Content

To measure each article's relatedness to institutional investors, we introduce a measurement method implemented using Google open-source word-embedding model trained on 100 billion words using Google News corpus. The Google model contains 300-dimensional vectors for 3 million words and phrases with the goal of representing the meanings of words using numeric vectors. As a breakthrough in computational linguistics, word-embedding method (Mikolov et al., 2013ab) uses a neural network to learn the contextual use of each word based on the distribution and ordering of the words in the news corpus. The use of Google news as input to the model ensures that the mapping of news-media concept and related vocabularies is consistent with the contextual style of WSJ newspaper language. The embedding method has been applied in recent financial studies of systemic exposures and transmission (Hanley and Hoberg, 2019) and of corporate culture (Li et al., 2021). Our goal is to use Google word2vec model to generate a list of vocabularies that are likely to co-appear in news articles relevant to the institutional investors. Using Google-news-based model allows us to generate words that are trained based on a larger scale of news articles and therefore improve the quality and relevance of the word list.

Specifically, we follow the methodology in the previous literature (Hanley and Hoberg, 2019; Li et al., 2021) and use "institutional investor" as the seed word that is fed into the pre-trained Google model. Next, we select the top 250 words with the highest similarity scores (i.e., the highest cosine similarity between their word vectors) from the Google word2vec model. In this process, we also map vocabularies from the Google word2vec to the WSJ corpus to ensure the top 250 words we select are in the WSJ corpus.

To quantify the extent of a WSJ news article's discussion of institutional crowd, we need to compare the WSJ news text with the related-word vector from Google

Electronic copy available at: https://ssrn.com/abstract=4207132

word2vec. We do this by computing the cosine similarity between the vocabulary list associated with institutional investor, and the raw text of each WSJ news article. This procedure has been widely used in finance, accounting and economics studies (Bhattacharya, 1946; Salton and McGill, 1983; Hoberg and Phillips, 2016). Specifically, two binary vector of 0's and 1's are separately created with the length of the WSJ dictionary: (1) Vector 1 is for the words present in each news article and (2) Vector 2 is the 250 related words from Google word2vec. Cosine Similarity is then calculated based on the two vectors. In general, the resultant cosine similarity score is a thematic score for every article that are bounded in [0,1] and each one indicates the intensity of media attention to the theme of "institutional crowd" that is specific to the given WSJ article.

A.2 Technical Details on Classifying WSJ Articles by Industry

For the WSJ news articles, Dow Jones has collected structured metadata that includes timestamps to the millisecond, categories, tickers pertaining to the news articles. For articles that have firm ticker tags, we are able to match each article with tagged firm's industry SIC classification from CRSP. We then match SIC to Fama-French 48 industry classifications.[30] For WSJ articles that do not have company tags, we apply a machine learning algorithm that classifies articles into industries based on the narrative structure of the articles itself and its topical attributes, in order to systematically score industry relevance for all articles in the WSJ data. To achieve this goal, we adopt a feed-forward neural network, which is used extensively in pattern recognition, combined with text-based topical modeling.

Specifically, we use topical modelling to reduce the dimensions of all the WSJ texts. We run latent dirichlet allocation (LDA) for our sample to identify the 1,000 topics of the WSJ corpus, which is a dimensionality reduction algorithm used extensively in computational linguistics (see Blei, Ng, and Jordan 2003). Similar to

[30] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Electronic copy available at: https://ssrn.com/abstract=4207132

principal components analysis for numerical data, LDA identifies verbal themes that best explain the variation in text across our sample. This step allow us to score each WSJ article with 1000 topic loadings, which represent the latent thematic structure of the document. To map un-tagged news articles to industry classification, we then train a simple multiple layer perception (MLP) feed-forward neural network that is widely used for pattern recognition. We use the 1000 LDA topic loadings and the industry classification of the articles that are tagged by Dow Jones as training and test sets. Our trained model outputs the industry classification with a prediction probability for the un-tagged articles. We only use articles with a higher than 50% probability of the industry assignment. Our final article count amounts to 1,018,718.

B Definitions of Financial Variables

The variable used in this study are defined as follows:

- **Past Return** is defined as a stock's past cumulative return from month $t-12$ to $t-1$. We avoid the returns in month t to mitigate the impact of microstructure effects such as short-term reversal effect.
- **Book-to-Market Ratio** is the natural logarithm of a firm's ratio of book equity and market value, defined following Davis, Fama, and French (2000). We exclude firms with negative book equity.
- **Investment** is defined as the growth rate of a firm's total assets (Fama and French (2015), and Hou, Xue, and Zhang (2015)).
- **Profitability** is defined as revenue minus cost of goods sold, SG&A, and interest expenses all normalized by the book value of equity (Fama and French (2015)). We exclude firms with negative book equity.
- **Size** is firms' market capitalization as of December of the fiscal year.
- **SUE** is the standardized unexpected earnings defined following Latane and Jones (1979).

Electronic copy available at: https://ssrn.com/abstract=4207132

C Keywords from Google News Word2Vec

We prepare this documentation to show the dictionary of the wordlist using the Google semantics model trained using Google News.

1. The list of 250 Expert-related words based on Google semantics model trained using Google News (ordered by similarity score)

{institutional_investor, fixed_income, morningstar, morgan_stanley, lipper, portfolio_manager, fortune_magazine, brokerage_firms, merrill_lynch, private_equity, hedge_fund, investment_banking, emerging_markets, credit_suisse, hedge_funds, jpmorgan, zacks, institutional_investors, investor, gabelli, brokerage, brokerages, goldman_sachs, equities, clsa, blackrock, asset_allocation, factset, barclays_capital, capital_markets, piper_jaffray, banc, mutual_fund, analyst, smith_barney, banker, mutual_funds, oppenheimer, quantitative, dealogic, nomura, high_yield, global, magazine, bear_stearns, deutsche_bank, legg_mason, forbes_magazine, csfb, securities, citigroup, wachovia_securities, cnbc, internet_retailer, broker_dealers, best, ranked, emerging, fortune, decade, academic, neuberger_berman, thomson_reuters, lazard, pimco, citi, schroders, broker, institutional, global_markets, derivatives, hottest, investment, reits, daiwa_securities, corporate_counsel, portfolio, consumer_goods, advisors, issuers, nomura_securities, recognized, putnam_investments, eaton_vance, morgan_keegan, methodology, weightings, outperform, analysts, consumer_staples, rankings, investment_management, cibc_world, stocks, instinet, underwriter, lehman_brothers, renaissance_capital, equity, publicly_traded, technical_analysis, alternative_investment, esquire, etfs, deutsche, msci, fool, factset_research, invesco, corporate_governance, newsweek, janus_capital, strategist, public_opinion, advisor, scholar, dresdner_kleinwort, prudential_financial, outstanding, nasd, semiconductor, structured_finance, comscore, responsive_politics, asia_pacific, finra, forbes, mergers, forrester_research, investor_relations, firms, financings, stifel_nicolaus, barron, thomson_financial, benchmark, senior_analyst, asset, mckinsey, medical_device,

Electronic copy available at: https://ssrn.com/abstract=4207132

janus, provider, vendor, gartner, casualty_insurers, innovator, ishares, outperformance, industrials, sectors, distinguished, portfolios, supplier, total_return, broker_dealer, prudential, dimon, ipos, morningstar_analyst, investing, thomson_first, supply_chain, investors, merrill, natixis, warburg_pincus, maxim, indices, private_banking, jefferies, tiaa_cref, contrarian, insight, credit_ratings, hewitt_associates, consumer_reports, babson, advance, icap, cbre, wasserstein, goldman, midcap, calpers, adrs, billboard, reit, service_provider, entrepreneur, strategic, product, fair_value, downgrades, redemptions, firm, fastest_growing, unicredit, brokers, vogue, economic_forum, jpmorgan_chase, client, calyon, issuer, commodities, indexes, publications, underperformed, funds, insider_trading, societe_generale, deloitte, preseason, julius_baer, customer_satisfaction, cdos, foolish, prestigious, caps, citic, outsourcing, fidelity_investments, value, daiwa, mellon_financial, altria, precious_metals, cnet, greatest, diversified_portfolio, intermediaries, stock_picks, franchise, pharma, miller_tabak, annualized_return, weighting, standard_chartered, convertible_bonds, innovative, rising_star, performer, stock_market, real_estate, markets, medco, managed, associate, needham, underperform, marketwatch}

2. 1. The list of 250 Prediction-related words based on Google semantics model trained using Google News (ordered by similarity score)

{prediction, predictions, predicting, forecast, forecasts, forecasting, projections, projection, predicted, estimation, estimate, guesses, assertion, predict, estimates, assumption, expectation, prophecy, hunch, predicts, prognosis, assertions, calculations, forecasted, assessment, probability, belief, outlook, forecaster, estimating, pronouncement, expectations, conventional_wisdom, theory, hypothesis, forecasters, observations, scenario, suggestion, conjecture, conclusions, overly_optimistic, assumptions, pessimistic, recommendation, observation, notion, analogy, calculation, pessimists, projecting, reasoning, analysis, optimists, consensus_estimate, pronouncements, likelihood, prescient, statistic, caveat, promise, predictive, probabilities, find-

Electronic copy available at: https://ssrn.com/abstract=4207132

ings, statistical_analysis, statistician, hypothetical, projected, guess, foresaw, wishful_thinking, theories, mathematical, diagnosis, conclusion, statistical, maxim, bets, optimistic, pledge, odds, guidance, quote, statistics, recollection, consensus, recommendations, doomsday, wager, warnings, skeptics, interpretation, certainty, analysts, predictor, guarantee, logic, opinion, outlooks, simulations, suggestions, mantra, meteorologists, remark, pundits, quip, guessed, economists, downward_revision, assuming, omen, figure, scenarios, skeptic, theoretical, probably, promises, comparisons, announcement, declaration, figures, warning, optimist, statisticians, suggest, analyst, view, explanation, expecting, reckon, declarations, revised_upward, decision, unscientific, revised_downward, naysayers, implying, assessments, harbinger, landfall, optimism, suggesting, thesis, plausible, methodology, hyperbole, stance, probable, adage, median_forecast, retort, reports, alarmist, proposition, speculation, implausible, meteorological, stats, inference, credo, proclamation, foresee, doubters, expect, reckoning, intuition, foregone_conclusion, upward_revision, almanac, presume, report, rumor, betting, speculating, worth_remembering, meteorologist, blueprint, eerily, simulation, premise, reckons, argument, saying, remarks, soundly, overestimated, advice, picks, however, target, believe, magnitude, gloomy, diagnoses, likely, optimistically, contrarian, pledges, description, tally, comparison, guessing, implication, explanations, bullish, oracle, timetable, judgment, rosy, pegged, sobering, calculates, hope, measurements, aberration, results, possibility, portends, portend, speculate, believing, prospects, watcher, foreseen, flatly, notions, worse, fluke, statement, cautiously_optimistic, comments, valuation, pollsters, thought, happen, bullishness, thoughts, almost, sanguine, pessimism, admonition, correlation, claim, assurances, outlier, camping, barometer, approach, miscalculation, mathematician, prospect}

D Technical Details on Counting Institutional Investor Names in WSJ Articles

We search WSJ text across all articles for each institution's name. We first collect the complete list of fund names and identifiers from the Thomson-Reuters Institu-

Electronic copy available at: https://ssrn.com/abstract=4207132

tional Holdings (13F) database. We then clean and pre-process fund names using the following steps. First, we drop "S & CO., INC" because they are uniquely identified. Next, we replace symbols with space. We then drop common fund suffixes at the end of the fund names. We review the intermediate outcomes and run this step multiple times to replace all occurrences of these suffixes. In addition, to better align the fund names with media references, we replace "MGMT" with "MANAGEMENT", "MGT" with "MANAGEMENT", "INVT" with "INVESTMENT", "INVMT" with "INVESTMENT", "ADVS" with "ADVISORS", and "TR" with "TRUST". We also replace instances of a trailing "L" with space. We then delete any extraneous space. We count how many times each individual cleaned fund name appears in each individual WSJ article.[31] We apply the same procedures to obtain word counts of mutual fund names (from CRSP mutual fund database), hedge fund names (from Thomson Lipper Hedge Fund database), and investment bank names (from Corporate Finance Institute) in WSJ articles.

E Additional Tables

[31] In order to check the quality of our name purge, we also go over the top hits with highest word count in the WSJ corpus and manually remove outliers that can generate false positives.

Electronic copy available at: https://ssrn.com/abstract=4207132

Table IA.1: **Robustness: Baseline Fama-MacBeth Regressions Excluding Industries of Institutional Investors**

This table reports the robustness checks for Table 6 by excluding SIC 4-digit industries that include institutional investors. We obtain institutional investors' CIK identifiers during 1999-2018 SEC filings from Kim, Wang, and Wang (2022), and then we link CIK to GVKEY to obtain institutional investors' SIC 4-digit industry codes. We exclude SIC 4-digit codes that starts with 6 and have been the industry code for an institutional investor during 1999 and 2018. These industry codes include 6020, 6035, 6141, 6172, 6199, 6200, 6211, 6282, 6311, 6321, 6324, 6331, 6361, 6411, 6552, 6722, 6726, 6797, 6798, and 6799. We run Fama-MacBeth regression of next-period monthly stock returns on the interaction between anomaly predictors (*PastRet* and *BM*) and WSJ institutional investor prediction measure (*InstPred*). Monthly stock returns are annualized by multiplying 1,200 for ease of interpretation. See Table 5 for variable definitions. All non-interactive independent variables are standardized to have mean 0 and standard deviation of 1. t-statistics are adjusted using Newey-West with two lags and reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from January 1981 to December 2020.

	(1)	(2)	(3)	(4)	(5)
PastRet×InstPred	1.60***	1.29**			
	(2.70)	(2.34)			
BM×InstPred			1.28***	1.28***	
			(3.48)	(3.68)	
InstPred	-0.42	-0.25	-0.55	-0.53	
	(-0.81)	(-0.56)	(-0.99)	(-1.11)	
PastRet	3.87**	2.54*		2.44*	2.57*
	(2.56)	(1.86)		(1.82)	(1.90)
BM		3.07***	4.39***	3.33***	3.05***
		(4.58)	(5.55)	(4.76)	(4.49)
Size		-0.84		-0.84	-0.91
		(-0.92)		(-0.91)	(-0.98)
Investment		-2.72***		-2.73***	-2.76***
		(-7.75)		(-7.75)	(-7.77)
Profitability		3.74***		3.79***	3.76***
		(5.23)		(5.29)	(5.23)
SUE		3.21***		3.21***	3.23***
		(12.86)		(12.82)	(12.81)
Articles		-0.40		-0.40	-0.31
		(-0.74)		(-0.75)	(-0.57)
Observations	1,640,735	1,640,735	1,640,735	1,640,735	1,640,735

Electronic copy available at: https://ssrn.com/abstract=4207132

Table IA.2: **Portfolio Sorts on Anomalies Only**

This table reports our replication of momentum and value anomalies without conditional on InstPred. Each month, we sort stocks into two size groups based on NYSE median market capitalization. Independently, we sort stocks into three groups by NYSE anomaly predictors (past returns from $t-12$ to $t-1$ in Panel A and book-to-market in Panel B). We next compute value-weighted excess returns within each of the 6 portfolios and then take simple averages of the returns between large- and small-cap portfolios within each of the 3 anomaly-InstPred portfolios. Monthly excess returns are annualized by multiplying 1,200. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from January 1981 to December 2020. Our long-short value anomaly is 98% correlated with the HML factor from Kenneth French's website, which has an average annualized return of 2.47% (t-statistics = 1.51) in our sample period.

	Panel A: Momentum Anomaly			
	Low PastRet	Med PastRet	High PastRet	H−L
Annualized Returns	5.58	8.76***	11.17***	5.59**
	(1.62)	(3.59)	(3.98)	(2.39)
	Panel B: Value Anomaly			
	Low BM	Med BM	High BM	H−L
Annualized Returns	7.59**	9.29***	9.74***	2.15
	(2.55)	(3.64)	(3.67)	(1.29)

Electronic copy available at: https://ssrn.com/abstract=4207132

Table IA.3: **Instit. Investor WSJ Connectedness (Triple Interactions)**

This table reports the following triple interaction regression between Anomaly, InstPred, and HighConnect, which is a dummy variable indicating if the FF48 industry has above median plausibly exogenous variation in investor-WSJ connectedness as defined in Table 10. See Section 4 and Table 10 for details. t-statistics are adjusted using Newey-West with two lags and are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Sample period is from January 1981 to December 2020.

$$
\begin{aligned}
ret_{i,t+1} = {} & \beta_1 Anomaly_{i,t} \times InstPred_{i,t} \times HighConnect_{i,t-3} \\
& + \beta_2 Anomaly_{i,t} \times InstPred_{i,t} + \beta_3 Anomaly_{i,t} \times HighConnect_{i,t-3} + \beta_4 InstPred_{i,t} \times HighConnect_{i,t-3} \\
& + \beta_5 Anomaly_{i,t} + \beta_6 InstPred_{i,t} + \beta_7 HighConnect_{i,t-3} \\
& + \beta_8 X_{i,t} \times HighConnect_{i,t-3} + \beta_9 X_{i,t} + \epsilon_{i,t+1},
\end{aligned}
$$

	(1)	(2)	(3)	(4)
PastRet×InstPred×HighConnect	2.74**	2.23**		
	(2.33)	(2.06)		
BM×InstPred×HighConnect			1.73**	1.65***
			(2.55)	(2.60)
PastRet×InstPred	0.03	-0.13		
	(0.04)	(-0.20)		
BM×InstPred			0.66	0.72*
			(1.53)	(1.80)
InstPred×HighConnect	0.47	0.74	-1.03	-0.58
	(0.41)	(0.75)	(-0.95)	(-0.62)
PastRet×HighConnect	-0.17	-0.67		0.26
	(-0.19)	(-0.78)		(0.36)
BM×HighConnect		0.36	-0.21	0.09
		(0.79)	(-0.35)	(0.17)
HighConnect	0.26	-0.83	-0.55	-0.87
	(0.29)	(-0.78)	(-0.60)	(-0.84)
InstPred	0.01	0.06	0.51	0.36
	(0.02)	(0.12)	(0.87)	(0.71)
PastRet	3.97***	2.77**		2.39*
	(2.64)	(2.00)		(1.74)
Size×HighConnect		0.70		0.71
		(1.36)		(1.39)
Investment×HighConnect		0.14		0.14
		(0.33)		(0.34)
Profitability×HighConnect		-0.99*		-1.03*
		(-1.67)		(-1.75)
SUE×HighConnect		0.08		0.06
		(0.27)		(0.18)
Articles×HighConnect		-0.45		-0.42
		(-0.45)		(-0.42)
BM		2.75***	4.24***	3.30***
		(4.35)	(5.73)	(4.95)
Size		-0.67		-0.66
		(-0.73)		(-0.72)
Investment		-2.77***		-2.78***
		(-7.37)		(-7.41)
Profitability		4.26***		4.33***
		(5.60)		(5.69)
SUE		3.25***		3.28***
		(11.81)		(11.88)
Articles		-0.61		-0.61
		(-1.03)		(-1.04)
Observations	1,857,988	1,857,988	1,857,988	1,857,988

Electronic copy available at: https://ssrn.com/abstract=4207132