

Wisdom of the Institutional Crowd: Implications for Anomaly Returns

AJ Chen

Gerard Hoberg

Miao Ben Zhang*

October, 2023

*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 Aydogan Altı, Frederico Belo, Odilon Camara, Lauren Cohen, Jonathan Cohn, Alan Crane, Zhi Da, Nicholas Guest, David Hirshleifer, Chong Huang, Mete Kilic, Joren Koeter, Kevin Murphy, Chris Parsons, Joel Peress, Jinfei Sheng, Clemens Sialm, Denis Sosyura, Laura Starks, Andrea Tamoni, Paul Tetlock, Sheridan Titman, Selale Tuzel, Mitch Waratchka, Florian Weigert, Feng Zhang, Xiaofei Zhao, and seminar participants at Chapman University, City University of Hong Kong, INSEAD, Georgia State University, UC Irvine, USC Marshall Brownbag, 14th Annual Hedge Fund Research Conference, ASU Sonoran Winter Conference, MFA, UT Austin Alumni Conference, SGF Conference, Connecticut Finance Conference, CETAFE, FutFinInfo, News and Finance Conference, Hong Kong Conference for Fintech, AI and Big Data in Business for valuable comments. We also thank Miyon Sung, Joseph Hedary, Ed Tilihoi, and everyone on the Wall Street Journal team for providing insightful advice, institutional knowledge, and technical 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 ©2023 by AJ Chen, Gerard Hoberg, and Miao Ben Zhang. All rights reserved.

Wisdom of the Institutional Crowd: Implications for Anomaly Returns

ABSTRACT

We study the implications of a novel crowd-sourcing mechanism in which institutional investors communicate with reputable news media to influence the crowd and accelerate return realization. Using over one million Wall Street Journal articles from 1980-2020, we create a new measure of crowd-sourcing based on institutional investor predictions in the news (InstPred). We show that for industries with higher InstPred, (i) value and momentum returns are 34% to 62% larger, and (ii) institutional investors collectively trade anomalies more. Our results are reinforced by quasi-exogenous variation in industries' investor-WSJ connections and cannot be explained by existing measures including sentiment.

This paper proposes a novel measure of institutional investor crowd-sourcing via the Wall Street Journal (WSJ) which strongly amplifies anomaly returns. Industries with higher exposure to institutional crowd-sourcing exhibit higher generalized anomaly returns using a combined portfolio of average anomalies (see Engelberg, McLean, and Pontiff (2018)) and also higher returns for specific anomalies: A one-standard-deviation increase in our crowd-sourcing measure in the cross-section boosts momentum returns by 51%-62% above its benchmark level and value returns by 34%-45% above its benchmark level. The anomaly returns are boosted by an even greater 74%-105% when industries' major institutional investors have strong connections to the WSJ.

The crowd-sourcing mechanism highlights a little-studied investment strategy for boosting information rents among institutions. Institutions possessing trading signals can earn rents both by (i) using the signals to form their own portfolios and (ii) influencing the crowd signal to accelerate their signal's predicted price movements. The former is well-studied. The latter ability to endogenously influence the speed of price realizations can generate large improvements in returns by shortening holding periods. For example, a fund that can speed up price realizations by a factor of two can place twice the number of anomaly bets over a given time span, potentially doubling its annual alpha.

Our crowd-sourcing mechanism and empirical results extend a large body of influential literature on news media and financial market outcomes.¹ These studies illustrate the high impact of news on stock returns for short-term horizons of one or more days. Among those using textual analysis, most document the importance of textual sentiment. We extend this work but take an entirely different approach on both dimensions. We predict *long-term* anomaly returns using theoretically-motivated measures of *interpretable content* instead of textual sentiment. Moreover, our crowd-sourcing mechanism also rationalizes and motivates why institutional investors regu-

¹See Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Fang and Peress (2009), Tetlock (2010), Engelberg and Parsons (2011), Fang, Peress, and Zheng (2014), Peress (2014), Solomon, Soltes, and Sosyura (2014), Engelberg, McLean, and Pontiff (2018), Guest (2021), Jeon, McCurdy, and Zhao (2022), among others.

larly interact with the media and engage in news production. For example, financial news articles frequently report the predictions of investment bankers, fund managers, investment advisors, and their staff.

In our illustrative model, institutional investors possessing signals of a trading strategy share their information via reputable news media to encourage other investors to trade in the same direction. As moving prices for the anomaly portfolios often requires far more capital than any one investor has available, investors are motivated to crowd-source and expedite return realization (Abreu and Brunnermeier (2002, 2003)). Sharing signals via *reputable* media, in particular, can help grow the base of informed investors, avoid cheap talk, and promote more sharing of trading signals. When the informed institutional crowd reaches a critical mass, arbitrage trades accelerate, prices correct, and anomaly returns result. Hence, news articles covering institutional investor predictions foreshadow both anomaly returns and institutional investors' anomaly trading.

The crowd-sourcing mechanism explains three features of our empirical findings that are distinct from the literature. First, it takes time for investors to reach a consensus regarding a profitable anomaly opportunity, which justifies that our empirical measurement based on a 3-month rolling window of news accumulation can predict monthly anomaly returns.² Second, extending the existing literature which focuses on the sentiment of media content, we join a small but growing number of studies focusing on interpretable news content (in our case, content relating to institutional investor predictions).³ Third, the crowd-sourcing mechanism applies better to trading strategies that require large amounts of liquidity (to incentivize crowd-sourcing) and are eminent among institutional investors. This motivates us to examine momentum and value anomalies beyond the baseline combined anomaly, as these two specific anomalies are known to have large systematic components requiring large amounts of liquidity⁴, and they are “everywhere,” as indicated by Asness,

²Our finding of a gradual accumulation of information is also consistent with the concept of gradual price accumulation noted in Da, Guron, and Warachka (2014).

³Other studies in asset pricing focusing on content include Bybee, Kelly, and Su (2023), Hirshleifer, Mai, and Pukthuanthong (2023), and Manela and Moreira (2017).

⁴For example, prior studies show that momentum has a significant industry component

Moskowitz, and Pedersen (2013).

We test our model predictions by constructing a novel measure of institutional investors' crowd-sourcing using over one million articles in the Wall Street Journal from 1979 to 2020. Our measure is constructed in three steps. First, we use Google word2vec embedding technology to measure each article's loadings on institutional investor content and on prediction content. We take the product of these two content loadings to capture the intensity of institutional investors' predictive statements (*InstPred*) in each article. Second, we aggregate the article level measure to the Fama-French 48 industries, where we train a neural network to assign industry tags to WSJ articles based on a subsample of articles with accessible industry labels. Finally, we compute abnormal institutional crowd-sourcing 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 abnormal *InstPred* measure to the standard firm-month return database to test our crowd-sourcing hypothesis.

Our main empirical finding is that abnormal institutional prediction activity in the WSJ crowd-sourced over 3 months strongly amplifies generalized anomaly returns (based on combining 204 anomaly predictors provided by Chen and Zimmermann (2022)) and specific anomaly returns. The amplifications are economically meaningful and particularly large for momentum and value as shown earlier. Results are strong both in cross-sectional Fama-MacBeth regressions and in stringent value-weighted portfolio tests. Additionally, the amplification of *InstPred* on anomaly returns is observed throughout our sample period, even during the financial crisis. This finding is consistent with our crowd-sourcing mechanism not relying on any link to the state of the economy. Finally, we observe results only when *both* institutional investor content and prediction content are highly prevalent (as is the case for *InstPred*), and our results are fully robust to controls for positive tone, negative tone, and uncertain textual tenor that are widely used in existing studies.

We next conduct two tests to examine our model's mechanism in explaining the

(Moskowitz and Grinblatt (1999), Hoberg and Phillips (2018)) and significant factor-based components (Ehsani and Linainmaa (2021)), and that the value anomaly also has a large systematic component (Davis, Fama, and French (2000)).

above results. The first test addresses an endogeneity concern that unobserved industry state variables might be driving our asset pricing results. Our model predicts that the effect of InstPred on anomaly returns should be strong specifically when industries' institutional investors are *connected* to the WSJ. We use institutional investors' *historical* name-mentions in WSJ articles from *unrelated industries* to draw quasi-exogenous variation in institutional investors' WSJ connectedness in a focal industry. We find that the amplification effects of InstPred on the anomaly returns are indeed larger in industries whose current major institutional block-holders have stronger connectedness to the WSJ. Next, we examine journalists leaving WSJ as another source of quasi-exogenous variation that temporarily “shuts down” our model mechanism. We find that the amplification effects are indeed weaker in industries whose major institutional investors are more exposed to the departure of WSJ journalists. These tests using variation that is closely related to our mechanism, but plausibly exogenous to industries' state variables, demonstrate our crowd-sourcing mechanism in driving the asset pricing results.

Our second test of the model mechanism examines institutional investor trading behavior. Our crowd-sourcing mechanism predicts that institutional investors are more likely to trade on anomalies when a consensus is about to be reached (after enough accumulation of institutional predictions in the media). We use the Thomson-Reuters Institutional Holdings (13F) database and show that institutional investors, especially non-passive investors, indeed trade more aggressively on anomalies when InstPred is higher. These results provide unified support for our model mechanism in both price and quantity asset pricing tests (Kojien and Yogo 2019).

We complete our analysis by exploring the topics institutional investors discuss when making predictions in the WSJ, and their link to momentum and value returns. In practice, institutional investors may discuss momentum or value opportunities in various contexts due to the complexity and multi-dimensionality of business information (Goldstein and Yang (2015)). Hence, it is meaningful to shed light on what economic content is specifically relevant to momentum and value. We select 25 “economic themes” from the taxonomy developed by Bybee et al. (2020) and compute

25 theme-specific InstPred measures at the industry-month level. Applying each theme-specific InstPred in Fama-MacBeth regressions, we find that 15 out of 25 economic themes are important in driving InstPred's effects on the value anomaly, and 6 themes are important for the momentum anomaly. These findings suggest that multiple types of information indeed likely explain these anomaly returns, as emphasized by Goldstein and Yang (2015). Themes relevant to momentum and value also appear to have different foci: themes about economic growth and macro conditions are uniquely important for momentum returns, whereas themes about corporate finance such as corporate earnings and managerial changes are uniquely important for the value premium.

Our paper makes three contributions to the literature. First, we show that a novel news-based measure of crowd-sourcing by institutional investors strongly amplifies anomaly returns. Prior studies on media and asset pricing typically focus on the short-term return responses to the news over a horizon of a couple of days, making it difficult to connect past news with future long-term anomaly returns. Guided by a crowd-sourcing mechanism, we show the accumulation of news about institutional investors' predictions over the past 3 months can significantly amplify anomaly returns in the future.⁵

Our second contribution is that our conceptual framework and empirical findings suggest a new view of the relationship between institutional investors and news media. Prior studies on media and asset pricing primarily view investors as responding to the news while treating the news itself as essentially exogenous.⁶ Our framework highlights that institutional investors can also contribute to news to shorten holding

⁵See Da, Engelberg, and Gao (2011), Ben-Rephael, Da, and Israelsen (2017), and Lee, Ma, and Wang (2015) for crowd-sourcing in other contexts using Google Search, Bloomberg, and SEC Edgar, respectively.

⁶For 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 (Peress (2008), Fang and Peress (2009), Tetlock (2010), and Huberman and Regev (2001)). News has also been extensively studied to lead to stock market reactions (Peress (2014), Engelberg, McLean, and Pontiff (2018), Guest (2021), Jeon, McCurdy, and Zhao (2022), among others). A notable exception is Ahern and Sosyura (2014), who illustrate that in the process of mergers, bidding firms endogenously control the amount and timing of their news to influence their stock prices.

periods and expedite profit realization.

Our third contribution is to develop a framework for extracting interpretable content themes embedded in the crowd signal. Examining InstPred jointly with economic themes from Bybee et al. (2020), we analyze which economic themes are most prevalent when institutional investors crowd-source on specific anomalies. We believe that technologies enabling sharp thematic content analysis in this framework can enable future research that further explores the roots of asset pricing anomalies.⁷

We end this section with a summary of limitations. First, we focus on just one news outlet, the Wall Street Journal (see, for example, Dougal et al. (2012) and Guest (2021)). While WSJ is widely regarded as a major reputable media source for financial and business news, we expect institutional investors might also crowd-source trading signals via other reputable media outlets. We are limited by data availability, but to the extent that crowd-sourcing spans multiple outlets, the consequence would be that our results are likely understated. Second, while our empirical findings support our model predictions in both anomaly returns and the quantity of institutional investor trades, we cannot fully rule out endogeneity concerns. Our findings using quasi-exogenous variation in institutional investors' connections to the WSJ, such as the departure of connected journalists, provide compelling evidence for our mechanism and against unobservable industry drivers. Yet, future research further exploring causality in this setting remains fruitful.

This paper is organized as follows. Section 1 presents a simple conceptual framework. Section 2 describes our data and measure. Sections 3 and 4 show our empirical findings and tests of the mechanism. Section 5 explores the thematic content related to momentum and value anomalies, and Section 6 concludes.

⁷A 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, which finds evidence of slow information diffusion using news headlines regarding momentum. Hillert et al. (2014) find that firms with high media coverage exhibit stronger momentum, suggesting that media attention can impact investor behavior, thus supporting overreaction-based theories of momentum. Using a high-frequency decomposition of daily stock returns, Jiang et al. (2021) find evidence of pervasive underreaction to firm news. Our study complements this literature on many dimensions.

1 Conceptual Framework

In this section, we describe a simple framework to characterize how institutional investors use news media to disseminate tradeable information and crowd-source price correction.⁸ Our framework adopts the basic setup of Abreu and Brunnermeier (2002), *AB model* hereafter, but adds news media into their setting.

Consider a market in which the prices of certain stocks deviate from their fundamental value. When mispricing is corrected, an anomaly return results.⁹ 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 liquidity providers who absorb trading orders and stabilize stock prices to a certain extent. Each arbitrageur is assumed to be infinitesimal, and the total mass of arbitrageurs is assumed to be 1.

At time 0, $\delta < 1$ fraction of arbitrageurs are informed that certain stocks are mispriced, where δ is common knowledge among informed arbitrageurs.¹⁰ For instance, they learned that there is an underreaction to past stock returns, leading to a profitable momentum strategy, or an overvaluation of growth stocks, leading to a profitable value strategy.

There are two key assumptions about arbitrageurs in the AB model. First, all ar-

⁸Our use of an implicit coordination-based framework is well-motivated by Kovbasyuk and Pagano (2022), who model informed investors' decision to advertise positions to accelerate price correction. The authors find that advertising becomes ineffective and coordination becomes more relevant when investors have larger numbers of assets to advertise. This is important in our setting of understanding anomaly portfolios, as they typically require taking positions in many assets.

⁹It is important to note that our mechanism applies irrespective of whether an anomaly is rooted in the correction of mispricing or exuberance that increases mispricing (van Binsbergen et al. (2023) and Lou and Polk (2022)). As long as arbitrageurs perceive predictable returns as a profitable trading opportunity, they have incentives to crowd-source and trade on that signal. Indeed, in our empirical tests, we show that our model predictions were supported in both the momentum anomaly, which is viewed as the exuberance of mispricing by van Binsbergen et al. (2023), and the value anomaly, which is viewed as the correction of mispricing. To keep our model simple for illustration, we follow Abreu and Brunnermeier (2002) and Engelberg, McLean, and Pontiff (2018) and assume the hypothetical anomaly in our model is rooted in the correction of mispricing.

¹⁰Hirshleifer, Subrahmanyam, and Titman (1994) seminal study trading behavior and equilibrium information acquisition when some investors receive common information before others. To focus on arbitrageurs' interaction in the media after time 0, we assume no exogenous arrival of information to arbitrageurs after time 0. See Abreu and Brunnermeier (2002) for a dynamic setting with the sequential arrival of private information.

bitrageurs are risk-neutral but face capacity constraints. Hence, trading orders from one or a few arbitrageurs cannot move the price. Instead, price correction occurs only when $\kappa \leq 1$ fraction of arbitrageurs trade in the same direction.¹¹ When it happens, the aggregate order imbalance of arbitrageurs exceeds the absorption capacity of behavioral traders, resulting in price correction. The price of a mispriced stock thus appears constant during the buildup period and corrects at the moment when κ fraction of arbitrageurs have placed their orders. This critical mass requirement introduces an implicit coordination element among arbitrageurs.

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 the inability to deploy capital to other trading strategies once the arbitrageur places the buy orders, or implicit costs such as the relative performance evaluation of fund managers that incentivizes managers to realize profits before an evaluation (see more examples in Abreu and Brunnermeier (2002)). The holding costs along with the critical mass requirement provide an incentive for informed arbitrageurs to crowd-source and 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 mispricing information to uninformed arbitrageurs.¹² Without loss of generality, we assume that one

¹¹For 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 orders to the maximum capacity.

¹²Following 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. This logic is theoretically supported by Van Bommel (2003) (see Section 4 of the paper), who document an informative equilibrium in the presence of reputation incentives in a related setting that also features wealth-constrained informed investors.

news article comes out each period with the information shared by an informed arbitrageur. Importantly, we assume that one piece of news does not perfectly transfer the information to all uninformed arbitrageurs at once. Otherwise, price correction occurs immediately after the first arbitrageur shares her 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 contains 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 trading signal comes out over time, uninformed arbitrageurs progressively become informed and trade on the information accordingly. As a result, 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 of 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*

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

Proposition 1 makes an important empirical prediction that anomaly returns are realized only when enough 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 behavior 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.*

Discussion of the simple model: 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 occur precisely at t^* . Instead, as each observes an uncorrelated noise signal, their trades will become more intense as time approaches t^* .

We note that our framework assumes that informed institutions truthfully reveal their signals to the WSJ. Van Bommel (2003) discusses the effect of investors' reputational incentives on mitigating price manipulation and cheap talk. We emphasize that reputational incentives for well-regarded media such as WSJ provide additional discipline for verifying the quality of news and avoiding fake news.

Finally, our model assumes that informed arbitrageurs have equal access to the news media for sharing information. In practice, building connections with the news media may take time, and arbitrageurs may have heterogeneous access to the media. We expect that our predicted mechanism is likely to be more prominent when arbi-

trageurs are more connected with the news media. This motivates our tests based on plausibly exogenous variation in the connectedness between arbitrageurs and the news media. Such tests can help distinguish our model from other mechanisms for anomaly returns. We explore this later.

2 Data and Measures

2.1 Data

Our news data set consists of the 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.¹³ Second, WSJ is widely regarded as authoritative and independent, making the quoted information providers accountable for any spreading of fake news.¹⁴ Hence, informed institutional investors can find it worthwhile to share their information with WSJ without worrying excessively that their signals might be discredited as cheap talk.¹⁵ Third, the WSJ article full-text 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.

We start by transforming raw article text using standard procedures (e.g., see Bybee et al., 2020). We set all characters to lowercase, remove common stop words, and words with fewer than 4 letters, and we separate text 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., light lemmatization). We then obtain bi-grams

¹³According 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>

¹⁴For instance, WSJ is one of four news medias and the only business-focused news media in the U.S. that reached the prestigious “newspapers of record by reputation” status. See https://en.wikipedia.org/wiki/Newspaper_of_record.

¹⁵Abreu and Brunnermeier (2002) argue that without an institution monitoring the credibility of news, there may exist an equilibrium in which no informed trader publicizes the private information.

of all pairs of adjacent uni-grams, and our final processed vocabulary includes uni-grams and bi-grams over our 40-year sample. As we are interested in industry-level economic news, we exclude WSJ articles with subject tags corresponding to non-economic content such as books, sports, entertainment, lifestyles, arts, and reviews. We also use the journal section tags to further exclude sections pertaining to Books, Bookshelf, Off Duty, Life & Arts and Golf Journal.

We next classify articles into industries. For articles that are about specific publicly traded firms, our data provides linked tickers. We use CRSP SIC codes and map these firms to Fama-French 48 (FF48) industries. 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 and their topical attributes. Internet Appendix A summarizes this procedure. We exclude articles that are not assigned to a dominant FF48 industry during the prediction process. Our final sample includes 1,018,718 industry-tagged WSJ articles.

2.2 Measuring Institutional Investors' Information Sharing

We use the text of WSJ articles to quantify institutional investors' predictive statements. This reflects the sharing of tradeable information at the article level. We then aggregate to FF48 industries and merge with our monthly stock return database.

2.2.1 "Institutional Investor" and "Prediction" Content 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.¹⁶ The use of Google News as input to Google's word2vec model ensures that our analysis is consistent with the contextual style of our newspaper corpus. The word2vec

¹⁶The 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). This embedding method has been applied in recent financial studies of risk exposure (Hanley and Hoberg, 2019) and corporate culture (Li et al., 2020).

procedure generates a list of words that are most likely to co-appear in news articles relevant to our seed word “institutional investor.”

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 include many investment banks, hedge funds, mutual funds, and words that are likely to appear in articles relevant to institutional investors. The full list of the 250 words is in Internet Appendix B.¹⁷

Our second step quantifies each WSJ article’s relatedness to institutional investors using the 250 keywords from above. We compute a cosine similarity score between each WSJ article and these 250 words. Cosine similarities have been widely used in academic studies (e.g., Bhattacharya (1946), Salton and McGill (1983), and Hoberg and Phillips (2016)). The result is a score bounded in $[0,1]$ for each WSJ article.

We also construct an analogous score for the unigram “prediction” using the word “prediction” as the seed word for the Google News word2vec embedding model. Table 1 lists the top 50 related terms for “prediction.” These include words that frequently appear in predictive statements such as “forecast,” “projection,” “estimate,” and “assertion” (the full list of 250 words is in Internet Appendix B).

Our main measure for capturing a WSJ article’s relatedness to institutional investor predictions 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 2 provides examples of WSJ articles that score high on the resulting “institutional investor & prediction” (InstPred) measure. In these examples, analysts or managers of institutions share their views on an industry’s

¹⁷An alternative approach is to measure the occurrence of institutional investor company names in each news article. Yet, this name-based approach is much noisier because (i) this approach cannot capture the many cases where the articles indicate institutional investor predictions without directly mentioning company names, and (ii) we do not have the names of all institutional investors nor the full array of denotations reporters use to reference the investors cited in each article. Nevertheless, we construct a names-based measure and display results in the Internet Appendix Table IA.1. We find that the names-based measure generates similar baseline results as our word2vec-based measure. However, the higher level of noise in this alternative measure leaves it with significantly less power for our later tests. We thus use the word2vec measure as our baseline measure throughout the paper as our objective is to test for mechanisms using the maximum amount of power available.

trajectory, illustrating our intuition for our InstPred measure.

Panel A of Table 3 provides the summary statistics for our three article-level measures. On average, 1% of the words in WSJ articles load on the institutional investor vocabulary, and 0.8% on predictive statements. Both variables have medians greater than zero suggesting that most articles mention at least one word from each list, and WSJ articles are thus informative on both. Regarding standard textual themes 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 InstPred is only mildly (20%-34%) correlated with these measures.

2.2.2 Standardized Institutional Investor Prediction for Industries

We next aggregate the article-level InstPred score to each FF48 industry in each month to facilitate our analyses of monthly stock returns. To mitigate the concern that some industries persistently have higher InstPred scores than others, we standardize industry-month InstPred scores so that a high score indicates abnormally high media coverage of the given theme relative to the industry's long-term average. By doing so, we identify specific industries and periods when each theme is particularly salient to WSJ readers relative to what has occurred in the past.

We obtain standardized InstPred using a 2-step procedure. First, for each industry i in each month t , we compute the average InstPred score (Q_{it}) over all articles mapped to the industry in the month. Second, we standardize Q_{it} by computing its mean and standard deviation over the thirteen observations $Q_{i,t-24}, \dots, Q_{i,t-13}$, and Q_{it} itself. The standardized measure is then $Z_{i,t} = \left[\frac{12}{13} Q_{i,t} - \frac{1}{13} (\sum_{k=13, \dots, 24} Q_{i,t-k}) \right] / \sigma_{i,t}$.¹⁸ Our use of the ex-ante window spanning months $(-24, -13)$ ensures that information is standardized relative to the level of media coverage from a “clean period” that was over one year in the past. A high value of $Z_{i,t}$ indicates that the InstPred theme is highly present in WSJ articles that cover industry i in month t .

¹⁸Our standardization includes Q_{it} in the calculation of the standard deviation to ensure that $Z_{i,t}$ is bounded (it 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 large outliers would be present.

Finally, our hypothesis and model suggest that price correction and institutional trading occur after media content accumulates over time (after t^* periods in the model). Hence, we construct (*InstPred*) as the 3-month rolling average of $Z_{i,t-2}$, $Z_{i,t-1}$ and $Z_{i,t}$. Our choice of a 3-month window is arbitrary. Yet fund managers have to file quarterly reports for performance evaluation, and a quarter can be a natural window for transmitting anomaly signals. We also experiment with various window lengths from 1 month to 36 months, and we find that WSJ InstPred significantly boosts momentum and value returns when the rolling window size is between 2 months and 12 months (t -statistics > 3). See details in Figure 1.

Using the above standardization and rolling-window procedures, we construct several additional textual measures from WSJ articles. These include the standardized “institutional investor” theme, standardized “prediction” theme, standardized number of WSJ articles, and standardized themes regarding tone and uncertainty. We then merge all industry-month measures to firm -months for analysis.

Table 4 reports summary statistics for our firm-month sample. The InstPred measure has a mean of 0.161 and is indeed bounded between $-\frac{12}{\sqrt{13}}$ and $\frac{12}{\sqrt{13}}$ indicating no outliers. InstPred has low correlations with existing cross-sectional return predictors including book-to-market, past returns, size, investment, profitability and standardized unexpected earnings. InstPred is mildly (13%) correlation with the standardized number of WSJ articles. We include this as a control in our regressions.

2.3 Anomalies

We rely on a comprehensive set of asset pricing anomalies that Chen and Zimmermann (2022) publicized at www.openassetpricing.com.¹⁹ The dataset includes 204 anomaly predictors at the stock-month level. From the website, we also obtain each anomaly’s sample selection criteria and portfolio sorting frequency. Chen and Zimmermann (2022) conduct extensive replication of these anomalies in-sample and confirm that almost all significantly predict returns in their original sample periods.

We use all 204 anomaly predictors and apply sample selection criteria from the

¹⁹We downloaded the March 2022 release version on August 4, 2023.

original studies. If the original study used only NYSE stocks, for example, we use this same sample for that anomaly. We follow Engelberg, McLean, and Pontiff (2018) and create a combined anomaly measure. Specifically, we sort stocks into quintile portfolios based on original sorting frequencies, and define the extreme quintiles as the long and short legs of each anomaly. For each stock-month, we calculate the probability it appears in the long portfolios and the probability it appears in the short portfolios. Taking the net of the long and short appearance probabilities gives our combined anomaly measure which indicates whether the stock in the month is more likely to be in the long or the short leg of the 204 anomalies.

In addition to the combined anomaly measure, we also individually consider two popular anomalies—momentum and the value premium. These two anomalies are especially appealing for contextualizing our model’s mechanism for two reasons. First, the crowd-sourcing mechanism applies more to trading strategies that require large amounts of liquidity to move prices (to incentivize crowd-sourcing) and that are eminent among institutional investors. Both of these anomalies are known to have large systematic components requiring large amounts of liquidity²⁰, and they are “everywhere,” as indicated by Asness, Moskowitz, and Pedersen (2013). Second, momentum and value returns are known to be highly negatively correlated in time series. Hence, one would expect the boosting of one of these anomalies to preclude manifestation of the other. As a result, testing whether InstPred can boost the *level* of both anomalies sets a high bar on testing our hypothesis and its degree of generality. We measure the momentum variable as each stock’s past return from $t - 12$ to $t - 2$ and value as the log book-to-market ratio. Overall, we report results using the combined anomaly variable, momentum, and value.

Finally, we use Compustat data to obtain additional firm financials, CRSP for monthly stock returns, and the Thomson-Reuters Institutional Holdings (13F) database for each stock’s institutional ownership. We restrict our final sample to common

²⁰For example, prior studies show that momentum has a significant industry component (Moskowitz and Grinblatt (1999), Hoberg and Phillips (2018)) and significant factor-based components (Ehsani and Linainmaa (2021)), and that the value anomaly also has a large systematic component (Davis, Fama, and French (2000)).

shares (CRSP shrcd 10 or 11) that are traded on NYSE, Amex, or Nasdaq. We also require stocks to have a positive book value of equity, and we exclude penny stocks with a price of less than one dollar. Internet Appendix C provides the definition of these variables and our firm control variables including size, investment, profitability, and standardized unexpected earnings (SUE).

3 Evidence on Anomaly Returns

This section presents our main stock return results. We first present monthly cross-sectional Fama-MacBeth (1973) regressions and 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:

$$ret_{i,t+1} = \beta_1 Anomaly_{i,t} \times InstPred_{i,t} + \beta_2 Anomaly_{i,t} + \beta_3 InstPred_{i,t} + \beta_4 X_{i,t} + \epsilon_{i,t+1},$$

where *Anomaly*_{*i,t*} is the stock's combined anomaly predictor (described in Section 2.3), or past cumulative return from $t - 12$ to $t - 2$ 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 prediction 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.²¹

²¹In the Internet Appendix Table IA.2, we also control for past changes in the stocks' institutional ownership and find nearly identical results.

To ease interpretation, we standardize all non-interactive independent variables to have mean 0 and standard deviation of 1. The term $Anomaly_{i,t} \times InstPred_{i,t}$ 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 5 presents the baseline Fama-MacBeth regression results. Column (1) shows that the combined anomaly is significantly stronger when WSJ InstPred is higher. A one-standard-deviation increase in InstPred increases the yearly anomaly return by 1.04% with a t -statistic of 3.43.²² Compared to the benchmark anomaly of 7.24% when InstPred is at its mean, a one-standard-deviation increase in InstPred boosts anomaly by 14% ($= 1.04\%/7.24\%$) of the benchmark level. Column (2) shows similar results after further controlling for other stock characteristics known to predict returns: A one-standard-deviation increase in InstPred corresponds to additional anomaly returns of 1.01% per year (t -statistics 3.69) or 17% of the benchmark level.

We next examine the two most well-known anomalies traded by institutional investors: momentum and value. Consistent with our prediction that the crowd-sourcing mechanism applies more to these systematic and eminent anomalies, Columns (3)-(6) show that InstPred more sharply boosts these two anomalies. A one-standard-deviation increase in InstPred boosts the momentum anomaly by 51% and 62% of the benchmark momentum anomaly without and with controls, respectively. A one-standard-deviation increase in InstPred boosts the value anomaly by 34% and 45% of the benchmark value anomaly without and with controls, respectively.

These baseline results show that InstPred has an economically large impact on anomaly returns. They support our model prediction that institutional investors' predictive statements via news media are an important synchronization device that predicts price corrections and anomaly returns.

²²Because our independent variables are standardized, an anomaly coefficient indicates the annual return difference between two stocks with a one-standard-deviation difference in past returns.

3.1.2 Results using Permutations of WSJ InstPred

We next sharpen our understanding of the baseline return results by examining the importance of having both the “institutional investor” and “prediction” themes mentioned in WSJ news for boosting anomalies. We construct four permutations in Table 6. 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. To test this conjecture, we construct two additional WSJ thematic variables, Inst&HiPred and Inst&LoPred, by multiplying each article’s “institutional investor” intensity with an above or below median “prediction” intensity, respectively. We observe that although Inst&HiPred significantly boosts anomalies in Columns (1)-(3) of Panel A, Inst&LoPred does not boost any anomaly, as shown in Columns (4)-(6) of Panel A. This result indicates that institutional investor content without the presence of prediction content is inadequate to boost anomaly returns. In Panel B, we report similar findings for Pred&HiInst and Pred&LoInst. Hence, prediction content without the presence of institutional investor content is also inadequate, indicating that both types of content are necessary to boost anomaly returns.

These results support our conceptual framework, which predicts price corrections when institutional investors communicate tradeable information through the news media, indicating the presence of actionable “wisdom from the institutional crowd”.²³

3.1.3 InstPred vs. Number of Articles

We next examine whether InstPred is simply an artifact of the quantity of news. Prior studies have examined stock returns and anomalies on days when a high volume of news about the stocks arrives, i.e., news days (e.g., Tetlock (2010) and Engelberg, McLean, and Pontiff (2018)). We first note that, unlike these existing studies that

²³In the Internet Appendix Table IA.3, we exclude the financial industries that contain institutional investors, and we find very similar results to our baseline findings in Table 5. Hence, our baseline findings are not driven by InstPred capturing news *about* institutional investors themselves.

focus on daily returns, our anomaly-thesis pertains to longer-term monthly returns, as does the broader anomaly literature.

To examine this issue, we add a key control for an interaction between the anomaly variables and the abnormal number of WSJ articles about the industry (as we do for InstPred). Columns (1) and (3) of Table 7 show that while InstPred reliably boosts anomaly returns, the number of articles does not. This is consistent with Engelberg, McLean, and Pontiff (2018), who find that having news on a particular day only increases anomalies on the same day the news is measured. As we impose a one month lag and average news content over a long 3 month ex-ante period, and additionally focus on industry-level signals, it is intuitive that our results are distinct. Moreover, our tests are geared toward a unique hypothesis rooted in crowd sourcing and long-term anomalies. Overall, our finding that InstPred is robust to accounting for the number of articles also suggests that standard attention effects likely do not explain our results.

3.1.4 InstPred vs. Sentiment

Although the use of media article tone (or sentiment) is pervasive in the literature, our institutional crowd-sourcing framework does not make any direct predictions regarding sentiment. First, our framework is based on the ability of institutional investors to speed information production, which we measure as the intensity of institutional investors making predictions in the media (InstPred). In this setting, unsigned informative content, positive content, and negative content all provide context and help to accelerate information production. Indeed, consistent with this expectation, we show that our measure of InstPred is not materially correlated with measures of sentiment (see Table 3).

Second, our framework predicts that InstPred can *boost* anomalies. Hence, our empirical tests compare long-short anomaly returns when InstPred is high to long-short returns when InstPred is low. Because anomalies have long and short legs, it is unclear whether sentiment/tone, without being attached to a specific leg of the anomaly, can boost anomaly returns. Indeed, Columns (4)-(9) of Table 7 run a

horse race between InstPred and tone variables in boosting anomalies. We find that neither positive nor negative tone boosts anomalies, while InstPred robustly boosts anomaly returns even after controlling for the interaction between tone and anomaly predictors.²⁴ We conclude that effects of InstPred are not subsumed by sentiment effects.

Third, our framework also does not make any predictions regarding interactions between our InstPred measure and sentiment measures, again, because an anomaly strategy has both long and short legs. It is thus unclear why a more positive-toned InstPred, without being attached to a specific leg of an anomaly, should boost the anomaly returns. We nevertheless create positive and negative tone versions of InstPred by interacting our baseline InstPred measure with positive and negative tone measures. Consistent with our prediction that InstPred is unique and distinct from sentiment, we find that these sentiment-interacted measures of InstPred do not boost returns beyond InstPred. One way tone might matter in our framework would be to sign the tone of WSJ articles while specifically tagging article content as being related to the long or short leg of specific anomalies. Such a pursuit likely stretches our existing data too thin, and is outside the scope of our study, as our focus is on measuring information acceleration in the general context of all anomalies.

In Columns (10)-(12), we further examine uncertainty news content and find that InstPred's effects also are not subsumed by uncertainty. Overall, the robustness of our findings to these controls illustrates that our results are novel.

3.2 Portfolio Sorts

We construct portfolios following Fama and French (1993, 2015). Specifically, we construct breakpoints for portfolios using only NYSE stocks. In each month, we sort stocks into 2 groups based on the NYSE median market capitalization. Independen-

²⁴This finding does not contradict prior literature on tone and stock returns. First, existing results examining tone focus on unconditionally predicting stock returns, rather than boosting long-short anomaly returns (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). Second, prior studies on tone focus on short-term return predictions, e.g., within days after the news, and focus on firm-specific rather than industry-wide news.

dently, we sort stocks into 3 groups based on the 30% and 70% percentiles of the anomaly predictors among NYSE stocks. 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 $2 \times 3 \times 3$ size-anomaly-InstPred portfolios. We then construct value-weighted excess returns for each portfolio. Finally, we compute the returns for the 3×3 anomaly-InstPred portfolios by averaging the returns of the large-cap and small-cap portfolios within each anomaly-InstPred category.

Table 8 shows the results. Panel A shows the combined anomaly conditional on InstPred. When InstPred is low, we observe a long-short anomaly return of 7.68% per year (t -statistics = 5.19). As we move from low-InstPred to high-InstPred, the long-short portfolio returns increase monotonically to a remarkable 12.08% per year (t -statistic = 6.77) for the high-InstPred group.

Panel B shows the momentum anomaly conditional on InstPred. When 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 significant 6.84% per year (t -statistic = 2.70) for the high-InstPred group.

Panel C shows 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 -statistic = 0.08) when 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.²⁵

We next explore the impact of InstPred on the long-short anomaly returns over time. Our thesis based on institutional investor signal crowd-sourcing does not re-

²⁵It 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 overall significance in each long-short portfolio return in Panel B of Table 8. In the Internet Appendix Table IA.4, 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 the HML factor with a correlation of 98%.

quire our return results to be state-dependent. Figure 2 plots ten-year smoothed portfolio returns (see Linnainmaa and Roberts (2018)) for the combined anomaly (Panel A), the momentum anomaly (Panel B), and the value anomaly (Panel C) in High-InstPred and Low-InstPred portfolios. As the red dashed line (High-InstPred portfolio) is notably above the blue solid line (Low-InstPred portfolio) throughout our entire sample period with few exceptions for all anomaly measures, we conclude that InstPred amplifications are not materially 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). Overall, our calendar time value-weighted portfolio results reinforce our baseline findings established using Fama-MacBeth regressions.

4 Tests of Mechanism

In this section, we conduct two tests of our theoretical mechanism. The first considers plausibly exogenous variations 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 media drive our asset pricing results. The second test examines our model prediction regarding 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

The prior section presented robust evidence that anomaly returns are significantly higher when InstPred is higher, supporting our model predictions. Yet, one concern is that an unobserved industry-level state variable might be driving both InstPred and future anomaly returns. For example, our industry-specific InstPred measure might correlate with a hidden signal regarding industry performance that drives anomaly returns. In this case, InstPred may boost anomaly returns regardless of institutional investor connections to WSJ journalists. To address this concern, we now consider

two sources of plausibly exogenous variation in investors' connectedness to the WSJ.

Variation in industries' investor-WSJ connectedness Our theory's most direct prediction is that the interactions between informed institutions and the WSJ should drive our predictable anomaly returns. A condition for our mechanism to work is that the informed institutions have direct connections to the media. In practice, any given institutional investor can have strong or weak connections to the media. This variation motivates a rather direct test of our proposed mechanism. When an industry's main institutions have strong connections to the WSJ in a given month, the crowd-sourcing mechanism we propose is likely to be stronger in the industry relative to when the main institutions are not connected. InstPred should thus amplify anomaly returns more in these industry-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 measure each Fama-French 48 industry's average *investor-WSJ connectedness* in each month through 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 industry i ranks among the top 10 of all institutional investors in the previous quarter.²⁶

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.²⁷ In order to ensure the variation we use is plausibly exogenous to the economic 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

²⁶Using our approach, each industry has on average 26 major institutional investors in a given quarter.

²⁷Lagging 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.

connectedness measure excludes all content from industry i itself, any state variable relevant to industry i is unlikely to confound the connectedness measure.

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. 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.²⁸

Table 9 displays the results of 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. In particular, among industries within high investor-WSJ connectedness, a one-standard-deviation increase in InstPred boosts the momentum anomaly by 105% ($= 2.17/2.07$) of the benchmark level and boosts the value anomaly by 74% ($= 2.44/3.30$) of the benchmark level. Internet Appendix Table IA.5 shows that the cross-term coefficients in the two subsamples differ significantly at either the 5% or the 1% level.

Variation in industries' exposure to WSJ journalists' turnover We next explore another source of plausibly exogenous variation relating to WSJ journalist turnover. When a journalist leaves the WSJ, the communication channel for institutions that were previously connected to that journalist might shut down for a period

²⁸Note that high or low investor-WSJ connectedness is not a permanent feature of an industry as it varies over time. For example, the high investor-WSJ connectedness dummy has a quarterly autocorrelation of 0.70.

of time until connections are rebuilt.²⁹ Hence, we expect InstPred to be less effective in amplifying anomaly returns when major institutional investors are highly exposed to WSJ journalist turnover.

To compute an institutional investor's exposure to WSJ journalist turnover, we need two components: the journalists who exit WSJ, and each institution's historical connectedness to the exiting journalists. We measure the time a journalist leaves WSJ as the quarter the journalist exits our database of WSJ articles from January 1984 to December 2020.³⁰ Similar to our work above, we measure an institutional investor's historical connectedness to a journalist based on the occurrences of the institution's name appearing in the articles authored by the journalist during the three years prior to the previous year. An institutional investor's *exposure to journalist turnover* in a quarter is thus the sum of the institution's connectedness to exiting journalists normalized by its connectedness to all journalists.

We compute industry i 's weighted average exposure to journalist turnover in a month by averaging its major institutional investors' exposure to journalist turnover, where we use each major institutional investor's dollar holdings of the industry's stocks in the previous quarter as weights. We then split each of our two subsamples from Table 9 (high versus low investor-WSJ connectedness) into two additional groups based on journalist turnover. Specifically, we group industries into above versus below top tercile exposure to journalist turnover within each of the balanced investor-WSJ connectedness subsamples.³¹ We thus test InstPred's effects separately among *Industries with Low Exposure to Turnover of Connected Journalists* and *Industries with High Exposure to Turnover of Connected Journalists*.

Table 10 shows the results. Columns (1)-(3) show that InstPred significantly amplifies anomaly returns among industries whose major institutional investors are less affected by WSJ journalist turnover. In contrast, Columns (4)-(6) show that the

²⁹We thank Joel Peress for suggesting we consider disruptions to the WSJ to test our mechanism.

³⁰Our database does not provide journalist names for WSJ articles before 1984. We only use journalists who exit before June 2019 to ensure the quality of our measure.

³¹We use the top tercile instead of the median to split the sample because the average journalist turnover rate is only about 3%, and a disproportionately large number of industries are not exposed to any connected journalist turnover in a given period.

amplification effects of InstPred on anomaly returns are statistically and economically insignificant among industries whose major investors are more heavily affected by WSJ journalist turnover. These insignificant results are consistent with our proposed mechanism that the turnover of connected WSJ journalists shuts down the communication channel between institutional investors and journalists and hinders the effectiveness of information dissemination by informed institutions.

Overall, the two quasi-natural experiments in this section illustrate that WSJ articles containing institutional predictions significantly amplify anomalies when an industry's major institutional investors are ex-ante connected to WSJ journalists. These findings support our model's mechanism that institutional investors use reputable media such as the WSJ to crowd-source their trades when they have access to the media, accelerating price adjustments and anomaly returns.

4.2 Evidence on Institutional Investor Trading

After showing the price effects (i.e., returns) above, we now explore InstPred's quantity effects by examining changes in institutional investors' holdings. Our model's second empirical prediction is that institutional investors will trade more aggressively on anomalies when InstPred is high, speeding up price correction. We thus test whether institutional investors indeed change their holdings 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 is compiled from the quarterly filings of 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 remaining 32 percent held by households and non-13F institutions (Kojien and Yogo (2019)). Form 13F reports long positions but 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.³² Bushee (2001) and Bushee and Noe (2000) categorize institutional investors with high portfolio turnover as “transient.” We regard these transient investors as being the actively-trading institutional investors seeking to crowd-source trades as modeled by our theory.³³ Bushee (2001) next 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 actively seek crowd-sourcing regarding anomaly profits. Finally, as there are no predictions regarding funds that simply track indices, following the convention in the literature, we remove the quasi-indexers from our sample.³⁴ 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.

For each stock, we compute quarterly changes in its institutional ownership based on all institutions, high-active institutions, and low-active institutions. We expect the results to be stronger for the most active institutional investors.

We merge monthly anomaly predictors and *InstPred* with the above database of quarterly changes in institutional ownership for stocks and 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 combined anomaly predictor, or past returns, or the natural logarithm of book-to-market ratio, $InstPred_{i,t}$ is the WSJ institutional-

³²<https://accounting-faculty.wharton.upenn.edu/bushee/>

³³Bushee’s website provides time-varying labels and permanent labels of transient for each institutional investor. We choose the permanent labels to mitigate the concern that institutional investors who traded on anomalies in the month are mechanically labeled as active.

³⁴It 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).

investor prediction measure for the stock's FF48 industry, $X_{i,t}$ is our array of control variables, and FE_t is the quarter fixed effects. We standardize all independent variables to have a mean of zero and a standard deviation of one for ease of interpretation where the unit observation is at the firm-quarter level.

Table 11 presents the results. Column (1) shows the overall non-index institutional investor trades for the combined anomaly. We find that when InstPred is high, institutional investors buy stocks that are more likely in the long legs of the 204 anomalies (Chen and Zimmermann (2022)) and sell stocks that are more likely in the short legs. Moreover, confirming the intuition that active institutional investors trade with the institutional crowd signal, we find that results are strongest for high-activity institutional investors in Column (2), but not for low-activity institutional investors in Column (3). The economic magnitude is moderate in nominal terms but larger in relative terms. Column (2), for example, shows that a one-standard-deviation increase in InstPred boosts high-activity institutional investors' anomaly trading by 1.49 basis points. When compared to these investors' unconditional anomaly trading benchmark of 7.99 basis points, a one-standard-deviation increase in InstPred boosts anomaly trading by a rather substantial 19% ($=1.49/7.99$) of the benchmark level.

Similarly, in Columns (4) and (7), we observe that institutional investors buy more past winner and value stocks and sell more past loser and growth stocks, respectively, when WSJ InstPred is high. In Columns (5)-(6) and (8)-(9), we observe a similar pattern that the effects on institutional investors' momentum and value trading are entirely driven by high-activity institutional investors rather than low-activity institutional investors. Compared to benchmark trading by high-activity institutional investors, a one-standard-deviation increase in InstPred boosts momentum and value trading by 8% ($=2.93/36.97$) and 24% ($=1.66/6.85$) of the benchmark levels, respectively.

These unified findings on both quantity and price changes support our proposed mechanism. They are consistent with active institutional investors crowd-sourcing their trades by sharing signals via the news media to accelerate anomaly returns.

5 What Content Indicates Momentum and Value?

In this section, we shed light on two additional questions about InstPred by further exploring the content in our 1 million WSJ articles. 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.

Second, which content themes relating to InstPred specifically boost anomaly returns? In our conceptual framework, when the signals related to a particular anomaly are communicated by arbitrageurs via the media, anomaly returns are predicted to be higher. We can thus use this technique to assess the extent to which specific economic content drives InstPred's ability to boost an anomaly and further infer what explains the anomaly. While the combined anomaly predictor can be driven by various economic forces governing the 204 underlying anomalies and is less meaningful to explore, momentum and value anomalies, on the other hand, have each received widespread research interest in understanding their drivers. We thus apply our technique to specifically assess the extent to which specific economic content drives InstPred's ability to boost momentum and value anomalies.

We select candidate content themes from those provided by Bybee et al. (2020),³⁵ and prune the 180 themes to a set of more plausibly relevant themes for our application. We 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 these two categories, 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 themes provides a relevant and detailed set of economic issues. At the same time, the list is not too large to preclude us from adding all into one

³⁵The topics and the keywords for each topic of Bybee et al. (2020) can be downloaded at <http://structureofnews.com/#>. We thank the authors for making these available.

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:³⁶

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

where $InstPred_{j,t}$ is the InstPred intensity of article j at month t , $Theme_{j,k,t}$ is the 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 suggests that the roots of anomaly mispricing are likely not uni-dimensional.

To explore the second question regarding which content themes specifically boost momentum and value anomaly returns, we construct 25 standardized measures corresponding to when high InstPred articles are focused on each of the 25 topic themes for each of the FF48 industries in each month. These thematic-focused versions of InstPred are first computed at the article level using our standard measure of institutional investor prediction content that is then interacted with one of our 25 interpretable economic content themes (these triple-content metrics are henceforth referred to as InstPredTheme). We standardize and map InstPredTheme to the industry-month level using the same 2-step aggregation and standardization methodology summarized in Section 2.2.2 and used for InstPred itself. Finally, we map the resulting industry-month values to our firm-month return database to facilitate standard monthly Fama-MacBeth return regressions. For each topic theme k , we then run the following Fama-MacBeth regression that interacts the topic theme with key

³⁶Following 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 et al. (2020).

variables in our baseline regressions:

$$\begin{aligned} ret_{i,t+1} = & \beta_1 Anomaly_{i,t} \times InstPredTheme_{i,t} \\ & + \beta_2 Anomaly_{i,t} + \beta_3 InstPredTheme_{i,k,t} + X_{i,t} + \epsilon_{i,t+1}. \end{aligned} \quad (2)$$

We run the above regression for the momentum anomaly and also for the value premium anomaly, using each topic theme one at a time, and for each regression, we record the t -statistic of the key interaction term for the anomaly variable and *InstPredTheme* for each given economic theme. Figure 3 plots the ordered-pairs of two t -statistics for each of the 25 topic themes, with x -axis representing the t -statistics from the value regression and the y -axis representing the t -statistics from the momentum regression. This figure provides an intuitive visualization regarding which economic themes are most important for each anomaly, which are important for both, 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 *InstPredTheme* is high, as 15 of the 25 themes interacted with *InstPredTheme* and book-to-market have t -statistics for β_1 above 3 (Harvey, Liu, and Zhu (2016)). In contrast, the topics that boost momentum are fewer as only 6 of the 25 themes have an analogous t -statistic above 3. Second, the topics for which a higher *InstPredTheme* boosts value and momentum returns are visibly different. Those that facilitate the effects of *InstPredTheme* on the value anomaly but not the momentum anomaly are 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 *InstPredTheme* effects on momentum are more related to issues of macro growth including the economic growth theme, macro data, optimism, and European sovereign debt.

Interpreting these results through the lens of our model, 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 shocks (e.g., Hong and Stein (1999), Daniel, Hirshleifer, and Subrahmanyam (1998), Jegadeesh and Titman (2011), and Hoberg and Phillips (2018)). Private signals that indicate major

revisions to fundamentals are thus likely valuable for identifying the underreaction in prices and when to trade, especially when large amounts of liquidity are required to move stock prices. Our results also suggest that signals that do not facilitate InstPredTheme to boost momentum are typically related to corporate variables and issues that are more passive in nature.

Value-anomaly themes are more numerous. The economic themes that facilitate the value anomaly but not momentum are more related to longer-term issues such as managerial incentives and effort, value creation through innovation, and the value each manager brings to the firm. Intuitively, this accords with the conventional wisdom that the value anomaly is slower moving than momentum. While these results are supportive of the conclusion that the value anomaly likely has multiple economic roots, they also are consistent with the value premium having a strong link between corporate finance and asset pricing.

Overall, this section provides suggestive evidence regarding thematic content that facilitates how InstPred boosts anomaly returns. Our results suggest that institutional investors have broad information sets, and they can share many different types of signals to induce crowd-sourcing of investment timing. While a full assessment of various asset pricing theories is beyond the scope of our study (which focuses on crowd-sourcing), these results can motivate future research to use our framework to assess the specific predictions of various anomaly theories.

6 Conclusion

This study explores crowd-sourced signals from institutional investors in news and examines how the crowd's signal relates to future anomaly returns. We construct a novel measure that captures institutional investors making predictions (InstPred) based on the full text of over one million WSJ articles from 1979 to 2020. Our main finding is that for industries with higher abnormal InstPred, (i) anomaly returns are significantly larger (e.g., 34% to 62% larger for value and momentum), and (ii) institutional investors collectively trade anomalies more aggressively. Amplified

anomaly returns are obtained when news covering institutional investors' predictions has accumulated over roughly 2 to 12 months, where we choose 3 months for our main analyses. Our focus on long-term anomaly returns distinguishes our work from most prior studies on media and asset pricing, which typically focus on short-term (e.g., daily) returns. Our use of textual analysis to assess interpretable content that is not related to sentiment also differentiates our study, as most existing studies on news and asset pricing focus on the tone of media articles.

Our findings supporting the crowd-sourcing mechanism are reinforced by two customized experiments. (i) Using quasi-exogenous variation in institutional investor connections to the WSJ, we show that our asset pricing results are strongest when major institutional investors are most connected to the WSJ (value and momentum are boosted by an even greater 74%-105%). (ii) We confirm that active institutional investors indeed trade anomalies more aggressively when WSJ articles about the industry exhibit more institutional investors' predictions.

Our broad results are most consistent with a crowd-sourcing mechanism among capacity-constrained institutions, as moving prices of major anomaly portfolios typically requires far more liquidity than any individual institution has available. Our central thesis is that these institutions crowd-source, at least in part, through highly reputable and visible news media such as the Wall Street Journal. Our theoretical framework that extends Abreu and Brunnermeier's (2002) setting predicts that arbitrage returns are realized only after enough news articles with institutional investors' predictions have accumulated, consistent with our empirical findings.

Further analysis of content in WSJ articles indicates that institutions build signals using information that spans a wide array of economic topics. 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 can be invaluable to future research that explores the roots of asset pricing anomalies.

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.
- Ahern, K., and Sosyura, D. "Who writes the news? Corporate press releases during merger negotiations." *Journal of Finance*, 69.1 (2014), 241-291.
- Asness, C.S., Moskowitz, T.J. and Pedersen, L.H. "Value and momentum everywhere." *Journal of Finance*, 68.3 (2013), 929-985.
- Ben-Rephael, A., Da, Z. and Israelsen, R.D. "It depends on where you search: Institutional investor attention and underreaction to news." *Review of Financial Studies*, 30.9 (2017), pp.3009-3047.
- Bhattacharya, A. "On a measure of divergence between two multinomial populations." (1946).
- van Binsbergen, J., Booms, M., Opp, C., and Tamoni, A. "Dynamic asset (mis)pricing: Build-up versus resolution anomalies." *Journal of Financial Economics* 147 (2023), 406–431.
- 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., Su, Y. "Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text." (2023) Working Paper.
- 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.
- Da, Z., Engelberg, J. and Gao, P. "In search of attention." *Journal of Finance*, 66.5 (2011), pp.1461-1499.
- Da, Z., Engelberg, J. and Gao, P. "The sum of all FEARS investor sentiment and asset prices." *Review of Financial Studies*, 28.1 (2015), 1-32.
- Da, Z., Gurun, U. and Warachka, M. "Frog in the pan: Continuous information and momentum." *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.
- Daniel, K. and Moskowitz, T.J. "Momentum crashes." *Journal of Financial Economics*, 122.2 (2016), 221-247.
- Davis, J., Fama, E. and French, K. "Characteristics, covariances, and average returns: 1929-1997." *Journal of Finance* 55 (2000), 389–406.
- Dougal, C., Engelberg, J., Garcia, D. and Parsons, C.A. "Journalists and the stock market." *Review of Financial Studies*, 25.3 (2012), 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., McLean, R.D. and Pontiff, J. “Anomalies and news.” *Journal of Finance*, 73.5 (2018), pp.1971-2001.
- Engelberg, J.E. and Parsons, C.A. “The causal impact of media in financial markets.” *Journal of Finance*, 66.1 (2011), 67-97.
- 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.” *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?.” *Review of Financial Studies*, 27.12 (2014), 3441-3466.
- Garcia, D. “Sentiment during recessions.” *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.
- Goldstein, I. and Yang, L. “Information diversity and complementarities in trading and information acquisition.” *Journal of Finance* 70 (2015), 1723-1765.
- 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”. *Review of Financial Studies*, 32.12 (2019), 4543-4603.
- Harvey, C. R., Liu, Y., and Zhu, H. “... and the cross-section of expected returns.” *Review of Financial Studies*, 29.1 (2016), 5-68.
- Hillert, A., Jacobs, H. and Müller, S. “Media makes momentum.” *Review of Financial Studies*, 27.12 (2014), 3467-3501.
- Hirshleifer, D., Mai, D., and Pukthuanthong, K. “War Discourse and Disaster Premia: 160 Years of Evidence from Stock and Bond Markets.” (2023) Working Paper.
- Hirshleifer, D., Subrahmanyam, A., and Titman, S. “Security analysis and trading patterns when some investors receive information before others.” *Journal of Finance*, 49.5 (1994), 1665-1698.

- 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. and Phillips, G.M. “Text-based industry momentum.” *Journal of Financial and Quantitative Analysis*, 53.6 (2018), 2355-2388.
- 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., Xue, C. and Zhang, L. “Digesting anomalies: An investment approach.” *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.” *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*, 145.2 (2022), 1-17.
- Jiang, H., Li, S.Z. and Wang, H. “Pervasive underreaction: Evidence from high-frequency data.” *Journal of Financial Economics*, 141.2 (2021), 573-599.
- 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.
- Kovbasyuk, Sergei, and Pagano, Marco “Advertising Arbitrage.” *Review of Finance*, 26.4 (2022), pp.799-827.
- Latane, H.A. and Jones, C.P. “Standardized unexpected earnings–1971-77.” *Journal of Finance*, 34.3 (1979), 717-724.
- Lee, C.M., Ma, P. and Wang, C.C. “Search-based peer firms: Aggregating investor perceptions through internet co-searches.” *Journal of Financial Economics*, 116.2 (2015), 410-431.
- Li, K., Mai, F., Shen, R. and Yan, X. “Measuring corporate culture using machine learning.” *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.
- Ljungqvist, A. and Qian, W., 2016. “How constraining are limits to arbitrage?.” *Review of Financial Studies*, 29(8), pp.1975-2028.

- Lou, D., and Polk, C. “Comomentum: Inferring arbitrage activity from return correlations.” *Review of Financial Studies*, 35.7 (2021), 3272–3302.
- Loughran, T. and McDonald, B. “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks.” *Journal of Finance*, 66.1 (2011), 35-65.
- Manela, Asaf, and Alan Moreira. ”News implied volatility and disaster concerns.” *Journal of Financial Economics* 123.1 (2017): 137-162.
- 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: <https://ssrn.com/abstract=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.
- Salton, G. and McGill, M.J. “Introduction to modern information retrieval.” mcgraw-hill (1983).
- 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.” *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?.” *Review of Financial Studies*, 23.9 (2010), 3520-3557.
- Van Bommel, Jos “Rumors.” *Journal of finance*, 58.4 (2003), 1499-1519.

Figure 1: **Effect of WSJ InstPred on Anomalies by Measurement Window.** The figure displays the strength of the signal from the WSJ “institutional investor prediction” (InstPred) in boosting anomaly returns as we change the window for constructing the InstPred measure. See Section 2.2.2 for more details on constructing InstPred and the measurement window. The solid navy line below reports the coefficient of the t -statistic of the interaction term $Anomaly \times InstPred$ in Fama-MacBeth regressions in Section 3.1.1 as we increase its measurement window for $InstPred$ from 1 month to 36 months. $Anomaly$ is a stock-month level index as the probability the stock appears on the long side of 204 anomaly strategies (from Chen and Zimmermann (2022)) minus the probability the stock appears on the short side in the month following Engelberg, McLean, and Pontiff (2018). The dotted line and dashed line report analogous t -statistics for the interaction terms $BM \times InstPred$ and $PastRet \times InstPred$, respectively. BM is the natural logarithm of book-to-market ratio. $PastRet$ is the return from $t - 12$ to $t - 1$.

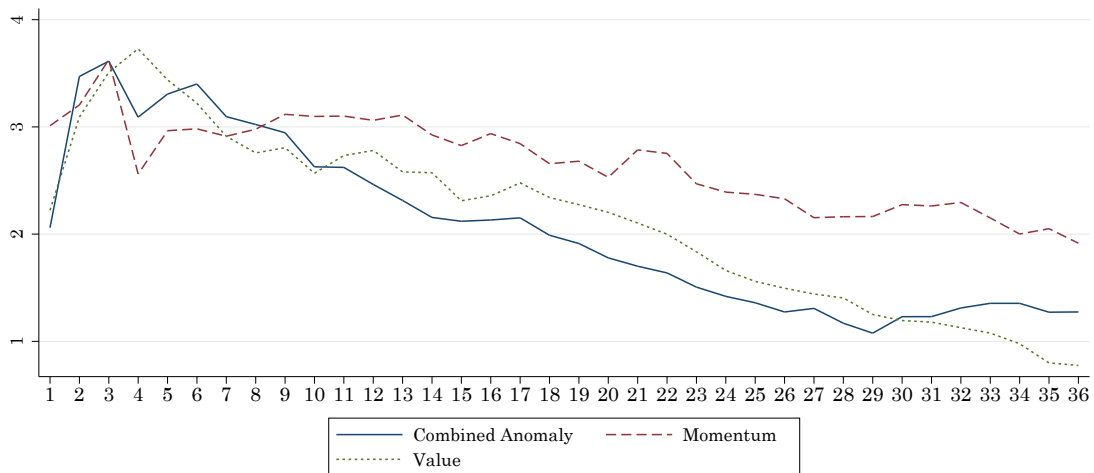
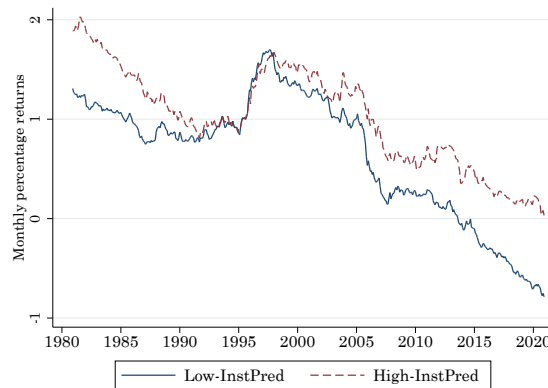


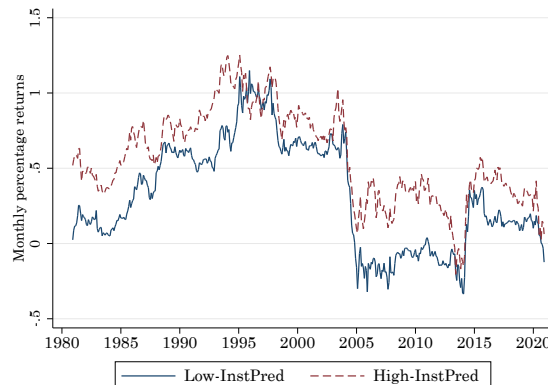
Figure 2: Anomaly Returns in Low-InstPred and High-InstPred Portfolios.

The figure plots the smoothed monthly percentage returns of the combined anomaly (in Panel A), momentum anomaly (in Panel B), and value anomaly (in Panel C) in the Low-InstPred and High-InstPred portfolios. The stock-month level predictor of *combined anomaly* is an index as the probability the stock appears on the long side of 204 anomaly strategies (from Chen and Zimmermann (2022)) minus the probability the stock appears on the short side in the month following Engelberg, McLean, and Pontiff (2018). See Table 8 for details of the portfolio formation. Each point in the line represents the average quantity for a ten-year window centered around the date indicated by the x -axis (Linnainmaa and Roberts (2018)). For example, the point in June 1990 represents the average return from July 1985 through June 1995.

Panel A: Combined Anomaly



Panel B: Momentum



Panel C: Value



Figure 3: WSJ Content Themes and the InstPred Effect on Anomalies. The figure displays pairs of Fama-MacBeth t -statistics of the interaction between an anomaly and our measures of institutional investor prediction content that is focused on an interpretable economic content theme (*InstPredTheme*). We consider 25 WSJ economic content themes for each FF-48 industry based on word lists from Bybee et al. (2020). The regressions are implemented analogous to our baseline results in the Table 5 except that we include *InstPredTheme* instead of *InstPred*. We run the regression specification in equation (2) using one *InstPredTheme* at a time. The x -axis displays the t -statistics of the interaction term between book-to-market ratio and *InstPredTheme*, while y -axis represents the t -statistics of the interaction term between past returns and *InstPredTheme*. See details in Section 5.

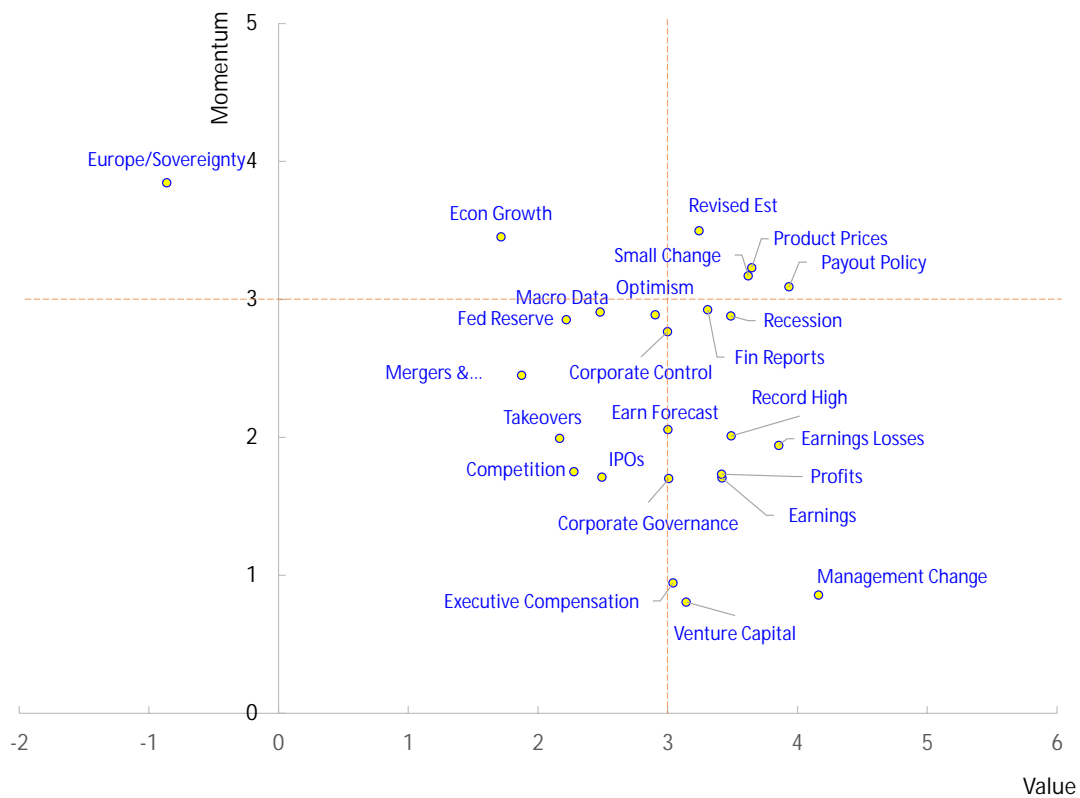


Table 1: Top 50 Keywords for “Institutional Investor” and “Prediction”

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 B 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

Table 2: **Sample WSJ Articles for Institutional Investor Prediction**

This table shows a sample of WSJ articles that have high “Institutional Investor & Prediction” (InstPred) scores. 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: [Bearishness Paid Off for Mr. Odey. Now He’s Bullish](#), 2009-04-16

“Crispin Odey, the London **hedge-fund manager** who gained fame and large returns last year by shorting U.K. banks, says the recent market rally could be the first signs of a new bull market. ... His bullish **assessment** makes him the latest in a string of high-profile investors to suggest the markets are on the way up... ‘Stock markets have shot up, led by the **financials** and the **base material sectors**,’ he said. ... Anthony Bolton, **Fidelity International’s** legendary stock investor, said six months ago that he was starting to buy, reiterating his stance last month. Sandy Nairn, a respected stock investor at **Edinburgh Partners**, also said last month that investors should begin reinvesting in shares. ...”

Example 4: [3-D Printing Expands to Metals, Showing Industrial Promise](#), 2016-11-11

“Those who didn’t **anticipate** the rise of 3-D printing of metals may have missed it because of disappointment in the polymer-based products made by consumer-oriented companies like MakerBot, said Cathie Wood, CEO of **investment manager ARK Invest**. ‘The **industrial part** of this is really heating up,’ said Ms. Wood, who earlier this year started a fund that only invests in 3-D printing companies. ...”

Table 3: **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.

<i>Panel A: Summary Statistics</i>						
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

<i>Panel B: Pearson Correlation Coefficients</i>					
	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

Table 4: **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 $t-13$. *Anomaly* is the probability the stock appears on the long side of 204 anomaly strategies (from Chen and Zimmermann (2022)) minus the probability the stock appears on the short side in the month following Engelberg, McLean, and Pontiff (2018). *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.

<i>Panel A: Summary Statistics</i>						
Variable	Mean	Std.Dev.	Minimum	Median	Maximum	Obs.
Monthly return	0.011	0.162	-0.981	0.001	19.884	1,936,301
InstPred	0.161	0.811	-3.092	0.132	3.276	1,936,301
Articles	0.073	1.124	-2.874	-0.045	3.284	1,936,301
Anomaly	-0.060	0.069	-0.456	-0.054	0.230	1,936,301
BM	-0.598	0.936	-11.308	-0.505	5.685	1,936,301
PastRet	0.161	0.729	-0.996	0.064	98.571	1,936,301
Size	12.236	2.151	4.676	12.093	21.170	1,936,301
Investment	0.148	0.383	-0.587	0.059	5.307	1,936,301
Profitability	0.143	0.428	-5.370	0.199	2.705	1,936,301
SUE	-0.123	1.634	-48.858	0.000	20.568	1,936,301

<i>Panel B: Pearson Correlation Coefficients</i>									
	Monthly return	InstPred	Articles	Anomaly	BM	PastRet	Size	Investment	Profitability
InstPred	-0.007								
Articles	-0.007	0.126							
Anomaly	0.041	0.000	-0.039						
BM	0.027	-0.025	-0.047	0.437					
PastRet	0.004	-0.060	0.002	0.142	0.024				
Size	-0.006	-0.038	-0.088	-0.188	-0.271	-0.002			
Investment	-0.021	0.046	0.029	-0.463	-0.158	-0.051	0.066		
Profitability	0.015	0.024	-0.002	0.079	0.092	-0.021	0.236	0.065	
SUE	0.010	-0.028	-0.001	0.111	0.026	0.113	-0.019	-0.053	-0.060

Table 5: **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 (*Anomaly*, *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 4 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. The sample period is from January 1981 to December 2020.

	(1)	(2)	(3)	(4)	(5)	(6)
Anomaly×InstPred	1.04*** (3.43)	1.01*** (3.69)				
PastRet×InstPred			1.93*** (3.40)	1.56*** (3.03)		
BM×InstPred					1.38*** (3.61)	1.41*** (4.02)
InstPred	-0.00 (-0.01)	0.20 (0.45)	0.15 (0.29)	0.40 (0.92)	-0.18 (-0.33)	0.03 (0.06)
Anomaly	7.24*** (9.73)	5.87*** (9.90)				
PastRet		1.23 (0.97)	3.80*** (2.63)	2.50* (1.87)		2.38* (1.81)
BM		0.81 (1.36)		2.79*** (4.17)	4.08*** (5.23)	3.10*** (4.46)
Size		0.21 (0.23)		-0.64 (-0.75)		-0.63 (-0.73)
Investment		-0.10 (-0.32)		-2.56*** (-7.79)		-2.57*** (-7.78)
Profitability		2.93*** (4.38)		3.56*** (5.07)		3.60*** (5.11)
SUE		2.73*** (12.53)		3.27*** (14.18)		3.27*** (14.12)
Articles		-0.70 (-1.34)		-0.72 (-1.35)		-0.72 (-1.35)
Observations	1,936,301	1,936,301	1,936,301	1,936,301	1,936,301	1,936,301

Table 6: **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 (*Anomaly*, *PastRet* and *BM*) and permutations of the WSJ institutional investor prediction measures. Monthly stock returns are annualized by multiplying 1200 for ease of interpretation. The table reports results when we develop WSJ themes aimed at isolating separate effects from the institutional investor theme and the prediction theme. Panel A reports results based on permutations of the prediction theme in InstPred. Inst&HiPred (Inst&LoPred) is the institutional investor theme loading for the given article multiplied by a dummy regarding if the article has an above (below) median value for the prediction theme relative to other articles from the same month. Panel B reports results based on permutations of the institutional investor theme in InstPred. HiInst&Pred (LoInst&Pred) is analogously defined as the prediction theme multiplied by the above (below) median institutional investor theme dummy. All regressions control for stock characteristics including size, investment, profitability, SUE, and articles. See Table 4 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. The sample period is from January 1981 to December 2020. There are 1,936,301 observations in each column.

<i>Panel A: Permutations of "Prediction" in InstPred</i>						
<i>WSJ:</i>	Inst&HiPred			Inst&LoPred		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Anomaly</i> × <i>WSJ</i>	1.15*** (4.37)			-0.40 (-1.61)		
<i>PastRet</i> × <i>WSJ</i>		1.48*** (2.87)			0.08 (0.16)	
<i>BM</i> × <i>WSJ</i>			1.47*** (4.59)			-0.36 (-1.21)
<i>WSJ</i>	0.34 (0.78)	0.59 (1.32)	0.17 (0.36)	0.16 (0.38)	0.28 (0.68)	0.17 (0.39)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
<i>Panel B: Permutations of "Institutional Investor" in InstPred</i>						
<i>WSJ:</i>	HiInst&Pred			LoInst&Pred		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Anomaly</i> × <i>WSJ</i>	0.81*** (3.40)			-0.33 (-1.34)		
<i>PastRet</i> × <i>WSJ</i>		1.42*** (3.16)			-0.29 (-0.65)	
<i>BM</i> × <i>WSJ</i>			1.23*** (3.83)			-0.57** (-2.01)
<i>WSJ</i>	0.47 (1.07)	0.71 (1.63)	0.32 (0.69)	0.13 (0.22)	0.11 (0.18)	0.27 (0.43)
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Table 7: 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 (*Anomaly*, *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 4 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. The sample period is from January 1981 to December 2020. There are 1,936,301 observations in each column.

<i>OtherTheme</i> :	Articles			Positive Tone			Negative Tone			Uncertainty		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
<i>Anomaly</i> × <i>InstPred</i>	0.95*** (3.38)			0.94*** (3.28)			0.84*** (3.01)			0.88*** (3.07)		
<i>Anomaly</i> × <i>OtherTheme</i>	-0.20 (-0.64)			0.20 (0.72)			0.07 (0.28)			0.41 (1.46)		
<i>PastRet</i> × <i>InstPred</i>		1.39*** (2.70)			1.42*** (2.61)			1.38*** (2.69)			1.46*** (2.65)	
<i>PastRet</i> × <i>OtherTheme</i>		0.33 (0.57)			0.34 (0.58)			0.86 (1.43)			-0.01 (-0.02)	
<i>BM</i> × <i>InstPred</i>			1.49*** (4.17)			1.34*** (3.67)			1.21*** (3.52)			1.37*** (3.69)
<i>BM</i> × <i>OtherTheme</i>			-0.33 (-0.88)			0.08 (0.23)			0.27 (0.71)			0.28 (0.79)
<i>InstPred</i>	0.17 (0.40)	0.44 (1.01)	-0.01 (-0.01)	-0.20 (-0.40)	-0.01 (-0.01)	-0.37 (-0.72)	0.19 (0.42)	0.45 (1.01)	0.06 (0.13)	-0.09 (-0.18)	0.08 (0.16)	-0.23 (-0.45)
<i>OtherTheme</i>	-0.67 (-1.30)	-0.96* (-1.79)	-0.60 (-1.10)	0.72 (1.31)	0.98* (1.85)	0.86 (1.52)	0.15 (0.26)	-0.04 (-0.07)	0.01 (0.02)	0.72 (1.29)	0.77 (1.38)	0.74 (1.31)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 8: **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 a combined anomaly predictor based on the stock's net probability of appearing in the long and short legs of 204 anomaly strategies, Panel B reports the results of using past returns as the predictor of the momentum anomaly, and Panel C 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 described in Table 4 (combined anomaly predictor in Panel A, past returns from $t - 12$ to $t - 1$ in Panel B, and book-to-market in Panel C). 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. The sample period is from January 1981 to December 2020.

<i>Panel A: Combined Anomaly</i>				
	Low Anomaly	Med Anomaly	High Anomaly	H-L
Low InstPred	4.27 (1.37)	9.75*** (3.67)	11.95*** (4.79)	7.68*** (5.19)
Med InstPred	5.94* (1.84)	10.81*** (4.11)	14.22*** (5.66)	8.28*** (5.56)
High InstPred	1.55 (0.47)	9.52*** (3.67)	13.63*** (5.55)	12.08*** (6.77)
<i>Panel B: Momentum</i>				
	Low PastRet	Med PastRet	High PastRet	H-L
Low InstPred	6.59* (1.88)	8.27*** (3.27)	10.29*** (3.57)	3.70 (1.50)
Med InstPred	7.16** (2.02)	9.39*** (3.73)	12.22*** (4.22)	5.06** (2.16)
High InstPred	3.52 (0.97)	8.48*** (3.38)	10.35*** (3.65)	6.84*** (2.70)
<i>Panel C: Value</i>				
	Low BM	Med BM	High BM	H-L
Low InstPred	7.76** (2.58)	9.38*** (3.60)	7.91*** (2.79)	0.16 (0.08)
Med InstPred	8.72*** (2.90)	10.01*** (3.70)	10.71*** (3.67)	1.99 (1.15)
High InstPred	5.58* (1.83)	9.14*** (3.44)	9.89*** (3.48)	4.31** (2.21)

Table 9: **The Role of Institutional Investors' WSJ Connectedness**

This table reports our baseline Fama-MacBeth regression (in Table 5) 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. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from July 1981 to June 2019.

	Industries with Low Investor-WSJ Connectedness			Industries with High Investor-WSJ Connectedness		
	(1)	(2)	(3)	(4)	(5)	(6)
Anomaly×InstPred	0.46 (1.52)			1.87*** (4.67)		
PastRet×InstPred		-0.14 (-0.22)			2.17** (2.34)	
BM×InstPred			0.71* (1.81)			2.44*** (4.72)
InstPred	0.35 (0.73)	0.04 (0.08)	0.32 (0.64)	0.07 (0.08)	0.83 (0.92)	-0.11 (-0.13)
Anomaly	6.44*** (11.37)			5.81*** (9.49)		
PastRet	1.21 (0.89)	2.79** (1.98)	2.41* (1.71)	1.39 (1.00)	2.07 (1.34)	2.61* (1.82)
BM	0.68 (1.16)	2.78*** (4.36)	3.36*** (4.98)	0.98 (1.50)	3.01*** (4.21)	3.30*** (4.27)
Size	0.31 (0.33)	-0.67 (-0.73)	-0.66 (-0.72)	0.82 (0.91)	0.03 (0.03)	0.04 (0.05)
Investment	-0.09 (-0.26)	-2.83*** (-7.41)	-2.84*** (-7.45)	-0.05 (-0.12)	-2.61*** (-6.43)	-2.62*** (-6.39)
Profitability	3.64*** (4.83)	4.32*** (5.59)	4.40*** (5.68)	2.42*** (3.62)	3.01*** (4.31)	3.03*** (4.31)
SUE	2.78*** (10.26)	3.28*** (11.82)	3.31*** (11.91)	2.78*** (11.89)	3.30*** (13.70)	3.30*** (13.71)
Articles	-0.47 (-0.85)	-0.58 (-1.04)	-0.59 (-1.05)	-1.02 (-1.10)	-1.13 (-1.19)	-1.09 (-1.15)
Observations	989,589	989,589	989,589	868,184	868,184	868,184

Table 10: **The Role of Connected-Journalist Turnover**

This table reports our baseline Fama-MacBeth regression (in Table 5) for two subsamples based on industries' exposure to turnovers of connected WSJ journalists. Turnover is defined as a journalist leaving the WSJ in the past year. We measure each institutional investor's exposure to the journalist turnover based on the number of occurrences the institution was reported by the journalist in the prior three years. Finally, for each FF48 industry, we aggregate its major investors' exposure to journalist turnover weighted by the dollar value of their holdings in the industry. In each month and within each high and low investor-WSJ connected industries, we divide the sample into two groups based on the top tercile of the industries' exposure to connected WSJ journalists, resulting in *Industries with Low Exposure to Turnover of Connected Journalists* and *Industries with High Exposure to Turnover of Connected Journalists*. 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.

*, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1987 to June 2019.

	Industries with Low Exposure to Turnover of Connected Journalists			Industries with High Exposure to Turnover of Connected Journalists		
	(1)	(2)	(3)	(4)	(5)	(6)
Anomaly×InstPred	1.02*** (2.98)			0.78 (1.30)		
PastRet×InstPred		1.83*** (2.77)			0.43 (0.39)	
BM×InstPred			1.31*** (3.37)			0.20 (0.28)
InstPred	-0.34 (-0.58)	-0.14 (-0.23)	-0.33 (-0.54)	0.49 (0.52)	0.48 (0.50)	0.51 (0.56)
Anomaly	6.30*** (10.22)			5.11*** (6.49)		
PastRet	0.30 (0.18)	1.57 (0.92)	1.42 (0.85)	2.46 (1.59)	2.93* (1.77)	3.45** (2.13)
BM	0.01 (0.02)	2.07*** (2.70)	2.41*** (3.00)	0.80 (1.19)	2.45*** (3.36)	2.43*** (2.73)
Size	0.51 (0.54)	-0.36 (-0.38)	-0.35 (-0.37)	-0.18 (-0.19)	-0.95 (-0.99)	-0.94 (-0.98)
Investment	-0.07 (-0.18)	-2.69*** (-6.67)	-2.68*** (-6.67)	0.04 (0.10)	-2.25*** (-5.24)	-2.24*** (-5.22)
Profitability	2.13*** (2.61)	2.76*** (3.27)	2.83*** (3.35)	3.35*** (4.14)	3.99*** (4.80)	3.99*** (4.74)
SUE	2.68*** (10.47)	3.23*** (12.33)	3.25*** (12.32)	2.97*** (8.98)	3.50*** (10.16)	3.49*** (9.90)
Articles	-1.28* (-1.79)	-1.32* (-1.83)	-1.38* (-1.90)	0.07 (0.07)	-0.13 (-0.13)	-0.18 (-0.18)
Observations	1,127,038	1,127,038	1,127,038	483,462	483,462	483,462

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 (*Anomaly*, *PastRet* and *BM*) and current WSJ institutional investor prediction 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 and Table 4 for definitions of other variables. 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 the stock level are presented in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from March 1981 to December 2018.

	All (1)	HiActive (2)	LoActive (3)	All (4)	HiActive (5)	LoActive (6)	All (7)	HiActive (8)	LoActive (9)
Anomaly×InstPred	1.49*** (2.59)	1.49*** (2.83)	-0.00 (-0.02)						
PastRet×InstPred				2.67** (2.32)	2.93*** (2.62)	-0.26 (-1.12)			
BM×InstPred							1.18** (2.02)	1.66*** (3.20)	-0.47 (-1.63)
InstPred	-1.41** (-2.41)	-1.54*** (-2.84)	0.13 (0.49)	-1.34** (-2.29)	-1.46*** (-2.68)	0.11 (0.43)	-1.36** (-2.33)	-1.47*** (-2.72)	0.11 (0.44)
Anomaly	6.16*** (8.49)	7.99*** (11.98)	-1.82*** (-5.96)						
PastRet	34.37*** (19.77)	35.44*** (19.43)	-1.07*** (-3.65)	35.61*** (21.56)	36.97*** (21.19)	-1.35*** (-4.51)	35.22*** (20.30)	36.54*** (20.00)	-1.32*** (-4.43)
BM	3.72*** (4.96)	3.95*** (6.47)	-0.23 (-0.65)	5.96*** (9.19)	6.85*** (12.95)	-0.88*** (-2.85)	5.95*** (9.15)	6.85*** (12.92)	-0.89*** (-2.87)
Size	-6.28*** (-15.16)	-5.12*** (-14.13)	-1.16*** (-6.11)	-7.07*** (-17.73)	-6.14*** (-17.81)	-0.93*** (-4.93)	-7.06*** (-17.72)	-6.13*** (-17.79)	-0.93*** (-4.93)
Investment	-3.54*** (-4.64)	-3.45*** (-5.06)	-0.09 (-0.30)	-6.08*** (-9.07)	-6.73*** (-11.25)	0.65** (2.55)	-6.07*** (-9.05)	-6.72*** (-11.22)	0.65** (2.53)
Profitability	-2.35*** (-4.26)	-2.39*** (-5.21)	0.05 (0.15)	-1.69*** (-3.11)	-1.53*** (-3.39)	-0.16 (-0.53)	-1.63*** (-3.00)	-1.47*** (-3.26)	-0.16 (-0.53)
SUE	9.50*** (16.51)	11.06*** (20.42)	-1.56*** (-6.43)	9.79*** (16.89)	11.44*** (20.92)	-1.65*** (-6.80)	9.85*** (16.91)	11.51*** (20.92)	-1.66*** (-6.84)
Articles	-1.41** (-2.29)	-1.30** (-2.40)	-0.10 (-0.34)	-1.31** (-2.13)	-1.20** (-2.21)	-0.10 (-0.35)	-1.41** (-2.30)	-1.33** (-2.45)	-0.08 (-0.27)
Observations	592,072	592,072	592,072	592,072	592,072	592,072	592,072	592,072	592,072

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 et al. (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. Coefficients of Instpred are multiplied by 100 for viewing convenience.

Theme	InstPred	Inst	Pred
Payout policy	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 sovereignty	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

Internet Appendix

Wisdom of the Institutional Crowd: Implications for Anomaly Returns

AJ Chen, Gerard Hoberg, and Miao Ben Zhang

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, the 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 the Google word2vec model to generate a list of vocabularies that are likely to co-appear in news articles relevant to the institutional investors. Using the 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

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 vectors 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 is 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.

We use the Google News word2vec keywords for institutional investors (instead of searching for institutional investor names as above) to identify an article's institutional investor focus for three reasons. First, some institutional investor names are common words that can create widespread measurement errors in article searches.³⁷ 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.*"³⁸ 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.

³⁷Examples include Boston Co Inc, Trust Co, and Society Corp.

³⁸The article can be found at <https://www.wsj.com/articles/SB104948462248570600>.

Nevertheless, we construct a name-based InstPred measure to examine the robustness of our baseline findings. To do so, 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. To account for similar yet not exact denotations of the institutions' names in the WSJ articles, we conduct a Named Entity Recognition (NER) technique to identify all entity names in the WSJ articles and select the names that have high similarity to our list of institutional investor names. Finally, we include a list of names that describe the institutional investors, such as "institutional investor", "hedge fund", "insurance company". We search whether each WSJ article and label the article as mentioning institutional names if it includes any of these keywords. Finally, we use this name-based dummy variable to replace the word2vec-keyword-based cosine similarity, and construct the name-based InstPred. We report the baseline Fama-MacBeth regression results in the Internet Appendix Table IA.1 that the name-based InstPred strongly boosts anomaly returns.

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, and 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.³⁹ 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

³⁹https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

in pattern recognition, combined with text-based topical modeling.

Specifically, we use topical modeling 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 principal components analysis for numerical data, LDA identifies verbal themes that best explain the variation in text across our sample. This step allows 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 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, op-

penheimer, 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, instinct, 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, 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, market-

watch}

2. 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, findings, 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, overesti-

mated, 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}

C Definitions of Financial Variables

The variables used in this study are defined as follows:

- **Combined Anomaly** is defined as the probability of a stock in the long quintile portfolio of 204 anomalies on www.openassetpricing.com March 2022 release version (Chen and Zimmermann (2022)) minus the probability of the stock in the short quintile portfolio following Engelberg, McLean, and Pontiff (2018).
- **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).

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 Institutional 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.⁴⁰ 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

⁴⁰In 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.

Table IA.1: **Robustness: Fama-MacBeth Regressions using Name-Based InstPred**

This table reports Fama-MacBeth regression of next-period monthly stock returns on the interaction between anomaly predictors (*Anomaly*, *PastRet* and *BM*) and a name-based version of WSJ institutional investor prediction measure (*InstPred*). See the Internet Appendix A for the construction of name-based InstPred. Monthly stock returns are annualized by multiplying 1,200 for ease of interpretation. See Table 4 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. The sample period is from January 1981 to December 2020.

	(1)	(2)	(3)	(4)	(5)	(6)
Anomaly × InstPred	0.83*** (3.34)	0.89*** (3.84)				
PastRet × InstPred			1.41*** (3.15)	1.15*** (2.77)		
BM × InstPred					1.07*** (3.27)	1.18*** (3.88)
InstPred	0.12 (0.24)	0.15 (0.35)	0.40 (0.79)	0.49 (1.19)	-0.17 (-0.32)	-0.02 (-0.04)
Anomaly	7.02*** (9.44)	5.75*** (9.86)				
PastRet		1.22 (0.96)	3.48** (2.44)	2.20* (1.68)		2.39* (1.82)
BM		0.75 (1.26)		2.75*** (4.09)	3.94*** (5.14)	2.97*** (4.31)
Size		0.22 (0.25)		-0.63 (-0.73)		-0.60 (-0.70)
Investment		-0.06 (-0.19)		-2.54*** (-7.70)		-2.56*** (-7.74)
Profitability		2.88*** (4.31)		3.55*** (5.09)		3.53*** (5.03)
SUE		2.73*** (12.58)		3.26*** (14.19)		3.27*** (14.15)
Articles		-0.49 (-0.95)		-0.50 (-0.95)		-0.51 (-0.95)
Observations	1,936,301	1,936,301	1,936,301	1,936,301	1,936,301	1,936,301

Table IA.2: **Robustness: Fama-MacBeth Regressions Controlling for Past Institutional Trades**

This table reports Fama-MacBeth regression of next-period monthly stock returns on the interaction between anomaly predictors (*Anomaly*, *PastRet* and *BM*) and WSJ institutional investor prediction measure (*InstPred*) controlling for the interaction between anomaly predictors and past changes institutional investors' ownership of the stocks (*InstTrade*). Monthly stock returns are annualized by multiplying 1,200 for ease of interpretation. 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. We compute the *InstTrade* as the most recent quarterly changes in the ownership in the month. Because institutional investors file their 13F form to the SEC in the next 45 days after the quarter ends, we merge stock returns with the 13F information with a five-month lag. That is, we regard the Q1 13F filings as the most recent information to the public in May, June, and July of the year, which correspond to next month's stock returns in June, July, and August. *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. *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. The sample period is from September 1981 to May 2019 (due to the matching to the 13F holdings data).

<i>InstTrade</i> :	All (1)	HiActive (2)	LoActive (3)	All (4)	HiActive (5)	LoActive (6)	All (7)	HiActive (8)	LoActive (9)
Anomaly× <i>InstPred</i>	1.29*** (4.59)	1.29*** (4.58)	1.29*** (4.58)						
Anomaly× <i>InstTrade</i>	-0.55*** (-2.81)	-0.26 (-1.33)	-0.41** (-2.49)						
PastRet× <i>InstPred</i>				1.71*** (3.23)	1.75*** (3.30)	1.72*** (3.23)			
PastRet× <i>InstTrade</i>				0.29 (0.62)	-0.20 (-0.42)	0.44 (1.28)			
BM× <i>InstPred</i>							1.58*** (4.23)	1.60*** (4.25)	1.59*** (4.26)
BM× <i>InstTrade</i>							-0.47** (-2.24)	-0.24 (-1.09)	-0.39** (-2.09)
<i>InstPred</i>	0.05 (0.11)	0.05 (0.10)	0.05 (0.12)	0.39 (0.85)	0.39 (0.85)	0.37 (0.80)	-0.14 (-0.28)	-0.15 (-0.30)	-0.14 (-0.28)
<i>InstTrade</i>	-0.63** (-2.40)	-0.42 (-1.47)	-0.39* (-1.83)	-0.10 (-0.37)	0.16 (0.53)	-0.22 (-1.05)	-0.51* (-1.89)	-0.20 (-0.69)	-0.41* (-1.85)
Anomaly	6.27*** (11.17)	6.38*** (11.42)	6.31*** (11.26)						
BM	0.41 (0.68)	0.42 (0.69)	0.42 (0.69)	2.51*** (3.71)	2.53*** (3.75)	2.51*** (3.71)	2.95*** (4.18)	2.98*** (4.22)	2.99*** (4.25)
PastRet	1.80 (1.38)	1.79 (1.36)	1.72 (1.30)	3.33** (2.41)	3.37** (2.36)	3.21** (2.31)	3.09** (2.27)	3.06** (2.22)	3.02** (2.18)
Size	0.48 (0.56)	0.44 (0.51)	0.47 (0.55)	-0.46 (-0.55)	-0.51 (-0.61)	-0.45 (-0.53)	-0.42 (-0.50)	-0.47 (-0.56)	-0.42 (-0.50)
Investment	0.07 (0.22)	0.08 (0.26)	0.06 (0.18)	-2.63*** (-7.80)	-2.64*** (-7.78)	-2.64*** (-7.76)	-2.65*** (-7.78)	-2.65*** (-7.78)	-2.64*** (-7.76)
Profitability	2.51*** (3.78)	2.51*** (3.79)	2.50*** (3.77)	3.20*** (4.62)	3.21*** (4.64)	3.21*** (4.61)	3.23*** (4.62)	3.24*** (4.64)	3.23*** (4.61)
SUE	2.70*** (13.46)	2.70*** (13.49)	2.70*** (13.46)	3.22*** (15.64)	3.22*** (15.58)	3.23*** (15.64)	3.25*** (15.62)	3.25*** (15.64)	3.24*** (15.52)
Articles	-0.81 (-1.55)	-0.81 (-1.55)	-0.81 (-1.57)	-0.83 (-1.55)	-0.83 (-1.55)	-0.84 (-1.58)	-0.86 (-1.60)	-0.85 (-1.58)	-0.85 (-1.60)
Observations	1,770,287	1,770,287	1,770,287	1,770,287	1,770,287	1,770,287	1,770,287	1,770,287	1,770,287

Table IA.3: Robustness: Fama-MacBeth Regressions Excluding Industries of Institutional Investors

This table reports the robustness checks for Table 5 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 (*Anomaly*, *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 4 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. The sample period is from January 1981 to December 2020.

	(1)	(2)	(3)	(4)	(5)	(6)
Anomaly × InstPred	0.83*** (3.06)	0.77*** (2.98)				
PastRet × InstPred			1.59*** (2.68)	1.28** (2.32)		
BM × InstPred					1.28*** (3.48)	1.28*** (3.69)
InstPred	-0.32 (-0.61)	-0.32 (-0.72)	-0.43 (-0.83)	-0.27 (-0.60)	-0.56 (-1.00)	-0.54 (-1.12)
Anomaly	7.52*** (10.66)	5.99*** (10.63)				
PastRet		1.25 (0.96)	3.86** (2.55)	2.53* (1.84)		2.42* (1.80)
BM		1.12* (1.85)		3.07*** (4.59)	4.39*** (5.56)	3.34*** (4.77)
Size		0.09 (0.10)		-0.84 (-0.92)		-0.84 (-0.92)
Investment		-0.06 (-0.20)		-2.73*** (-7.78)		-2.74*** (-7.78)
Profitability		3.10*** (4.47)		3.74*** (5.22)		3.78*** (5.28)
SUE		2.67*** (11.17)		3.22*** (12.87)		3.22*** (12.84)
Articles		-0.36 (-0.68)		-0.40 (-0.74)		-0.40 (-0.75)
Observations	1,640,567	1,640,567	1,640,567	1,640,567	1,640,567	1,640,567

Table IA.4: **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 (combined anomaly predictor in Panel A, past returns from $t - 12$ to $t - 1$ in Panel B, and book-to-market in Panel C). 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. See Table 8 for more details. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The 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.

<i>Panel A: Combined Anomaly</i>				
	Low PastRet	Med PastRet	High PastRet	H-L
Annualized Returns	3.97 (1.26)	10.21*** (4.02)	13.23*** (5.56)	9.27*** (6.36)
<i>Panel B: Momentum</i>				
	Low PastRet	Med PastRet	High PastRet	H-L
Annualized Returns	5.58 (1.62)	8.76*** (3.59)	11.17*** (3.98)	5.59** (2.39)
<i>Panel C: Value</i>				
	Low BM	Med BM	High BM	H-L
Annualized Returns	7.59** (2.55)	9.29*** (3.64)	9.74*** (3.67)	2.15 (1.29)

Table IA.5: Testing High and Low Investor-WSJ Connectedness

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 9. See Section 4 and Table 9 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. The sample period is from January 1981 to December 2020.

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

Table IA.5: Testing High and Low Investor-WSJ Connectedness—Continued

	(1)	(2)	(3)	(4)	(5)	(6)
Anomaly× InstPred× HighConnect	1.75*** (3.42)	1.35*** (2.89)				
PastRet × InstPred × HighConnect			2.74** (2.33)	2.23** (2.06)		
BM × InstPred × HighConnect					1.74** (2.56)	1.66*** (2.61)
Anomaly × InstPred	0.43 (1.35)	0.46 (1.52)				
PastRet× InstPred			0.02 (0.03)	-0.14 (-0.22)		
BM× InstPred					0.66 (1.53)	0.73* (1.81)
InstPred × HighConnect	-1.07 (-0.99)	-0.41 (-0.45)	0.48 (0.42)	0.75 (0.76)	-1.03 (-0.95)	-0.58 (-0.62)
Anomaly× HighConnect	-0.37 (-0.76)	-0.33 (-0.68)				
PastRet× HighConnect		0.22 (0.29)	-0.16 (-0.18)	-0.65 (-0.77)		0.28 (0.38)
BM× HighConnect		0.34 (0.68)		0.35 (0.78)	-0.22 (-0.36)	0.08 (0.15)
Size× HighConnect		0.51 (0.99)		0.70 (1.35)		0.71 (1.38)
Investment× HighConnect		0.04 (0.07)		0.15 (0.36)		0.16 (0.36)
Profitability× HighConnect		-0.94 (-1.58)		-0.97 (-1.65)		-1.02* (-1.73)
SUE× HighConnect		0.06 (0.20)		0.08 (0.27)		0.06 (0.18)
Articles× HighConnect		-0.46 (-0.47)		-0.45 (-0.45)		-0.41 (-0.41)
InstPred	0.58 (1.00)	0.39 (0.77)	-0.01 (-0.01)	0.04 (0.09)	0.50 (0.86)	0.36 (0.71)
HighConnect	-0.58 (-0.65)	-1.05 (-1.03)	0.27 (0.29)	-0.82 (-0.77)	-0.56 (-0.60)	-0.86 (-0.83)
Anomaly	7.54*** (11.19)	6.25*** (11.38)				
PastRet		1.19 (0.89)	3.95*** (2.62)	2.75** (1.98)		2.37* (1.71)
BM		0.67 (1.16)		2.75*** (4.36)	4.25*** (5.75)	3.31*** (4.96)
Size		0.31 (0.33)		-0.67 (-0.73)		-0.66 (-0.72)
Investment		-0.09 (-0.26)		-2.78*** (-7.41)		-2.79*** (-7.45)
Profitability		3.57*** (4.83)		4.25*** (5.59)		4.32*** (5.68)
SUE		2.75*** (10.26)		3.25*** (11.82)		3.28*** (11.91)
Articles		-0.49 (-0.85)		-0.61 (-1.04)		-0.61 (-1.05)
Observations	1,857,773	1,857,773	1,857,773	1,857,773	1,857,773	1,857,773