Textual Analysis of News and the Cross-section of Stock Returns

by

Ran Tao

Thesis submitted for the degree of Doctor of Philosophy

ICMA Centre, Henley Business School

Tuesday 4th May, 2021

University of Reading

Declaration of Original Authorship

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Abstract

This thesis contributes to the growing literature on the textual analysis of news and the cross-section of stock returns throughout three main chapters based on the Dow Jones Newswire Archive. First, I consider in an empirical setting how the presence/absence of news coverage affects lottery-type stocks, measured by the maximum daily returns (MAX) in the prior month. I find an augmented negative relationship between MAX stocks without news and expected returns, whereby MAX with news coverage generates return momentum. The differing future return relationships between MAX stocks with and without news appears to be best explained by information uncertainty mitigation upon news arrival. Overall, this finding suggests that news plays a role in resolving information uncertainty in the stock market.

Next, by applying a "co-coverage" concept, I propose to identify each firm's news-based peers (NBPs) and thereby construct a time-varying firm-centric grouping aiming to augment existing industry classification schemes. The advantage of NBP-augmented schemes over traditional industry classifications is to better capture the up-to-date relative importance of the economic links between a base firm and its peers. I show that the base firm's share price responds more favourably to the return shocks of its NBPs than of its traditional industry peers that are not NBPs. The response persists for several months, suggesting that the up-to-date relative importance of the economic links is not immediately clear to investors. Additional

empirical tests show that the persistent response, referred to as NBP momentum, can partially unify several lead-lag return momentums previously documented and is consistent with the investor attention hypothesis. Taken together, these results suggest that monitoring news co-coverage is a key to understanding the lead-lag return momentums documented in the literature.

Finally, I distinguish news coverage as being either slowly or quickly incorporated into contemporaneous stock prices. The return spread between stocks classified according to these two types of news yields a statistically significant profit of 139 basis points per month. This abnormal return cannot be explained by other well-known risk factors and is robust to allow for trading costs. Overall, this research refines the role of news regarding information dissemination in the financial markets.

Acknowledgements

I would like to dedicate this thesis to the following people for their support and accompanying me at every step of my PhD journey.

First, I am indebted to my PhD advisors Prof. Adrian R. Bell and Prof. Chris Brooks for all encouragement and guidance throughout my PhD journey. They are always willing and open to discuss and give high-level suggestions on my investigative pieces and career development. Taking up directly the role of an independent researcher is the best way for one to feel the responsibility and become capable of and skilful in serving this role. This process, despite being painful, lets me grow and eventually be able to stand on my own feet. I also thank the ICMA Centre's financial support for the purchase of the Dow Jones Newswire Archive on which this thesis is based. The entire empirical analysis of the thesis cannot be vigorously elaborated and thoroughly conducted without this dataset.

I would like to express my gratitude to my dear family, especially to my father who always trusts and supports my career decisions. Without their continuous support, I cannot fully realize my potential as a doctoral researcher.

I thank Prof. James Steeley and Dr. Andrew Urquhart for agreeing to be my PhD examiners. I also especially thank Prof. Chris Adcock, Prof. C.S Agnes Cheng, Prof. Steven Young, Dr. Simone Varotto, Dr. Tony Moore, Prof. Marcel Prokopczuk, Dr. Xin Liu, and Dr. Andrew Yim for the time they generously spent on reviewing earlier drafts of the chapters in this thesis. I would also like to thank the participants at the seminars and conferences where I presented my work for the various feedback that substantially improved the quality of this work, including the 2019 BAFA Doctoral Session, 2019 FMA Europe Doctoral Session, 2019 BAFA Asset Pricing and Corporate Finance Conference, and 2020 AFA PhD Poster Session. My thanks also go to the anonymous associate editors and referees for their insightful comments on my papers “When is a MAX not the MAX? How News Resolves Information Uncertainty”, which is included in this thesis as Chapter 2 and is published in *Journal of Empirical Finance* and “Tomorrow’s Fish and Chip Paper? Slowly-incorporated News and the Cross-section of Stock Returns”, which is the Chapter 4 and is now published in *European Journal of Finance*. The proofreading by Barbara Docherty is gratefully acknowledged.

My sincere thanks to the fellow PhD colleagues including those who have completed the degree and those who are still pursuing it. My PhD journey would be much boring without your company. The list of thanks includes Zhenghong Ding, Erick Fung, Dina Ghanma, Tian Han, Cai Liu, Yongda Liu, Shuyuan Qi, Robert Wichman, Xiaohan Xue, and Melanie Zhang.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General Background

Predicting stock returns is always an important topic that motivates most asset pricing studies. Starting with the pioneering work of Fama and French (1992), Grinblatt and Titman (1992) and beyond, a rich body of literature has documented perhaps hundreds of stock return predictors. Over the past decades of research, we have learned and come to understand that cross-sectional variation in average stock returns is somehow predictable. By sorting stocks on observable firm characteristics into different portfolios, researchers calculate the average returns of an investment strategy that goes long stocks and short stocks at the end of each period. These long-short strategies can be used to identify abnormal returns that are not explained by existing theories.

The dispute, however, lies in the explanation of these predictors. Academics sharply disagree about the causes of cross-sectional predictability. The traditional risk-bearing theory believes that a predictable stock return is a result of exposure to certain observable or even latent risk factors (see e.g., Hou, Xue, and Zhang (2015) and Fama and French (2015)). The emerge of anomalies is only due to market friction. Behavioural finance, on the other hand, predicts that investors' biased belief fosters the mispricing of stock returns. For instance, one of the prominent behavioural theories is the cumulative prospect theory by Kahneman and Tversky (1979). The theory predicts that, when investors evaluate risk, they often deviate from the expected utility and overweight (underweight) the small (large) chance of large pay-offs. As a result, it offers a unifying theory to reconcile a number of seemingly unrelated anomalies, such as lower return of distress stocks, conglomerate firms traded at a discount, and higher return of low idiosyncratic volatility stocks. These biases are correlated with anomaly variables which, when new information arrives in the market, will be corrected and lead to subsequent predictable returns.

While two competing theories continue to motivate not only hundreds of new anomalies but also thousands of new risk factor candidates, the majority of these candidate variables are correlated (Kelly, Pruitt, and Su, 2019; Freyberger, Neuhierl, and Weber, 2020), which deters further understanding of the anomaly evolution. News, on the other hand, is less likely to face endogeneity issues and could be a key driver of other characteristics. Thus, investigating the role of news in the anomaly literature is likely to offer additional insights. For example, Engelberg, McLean, and Pontiff (2018) investigate 96 anomalies' predictability on news days, measured by corporate public news and earnings announcements, and on no-news days, respectively. They find that the anomaly expected returns are much higher on these news days and the effect appears to be in line with the investor mispricing pattern, rather than exposure to risk factors. Lochstoer and Tetlock (2020) decompose five well-known anomalies into discount rate news and cash flow news. While discount rate news predominantly predicts the market-level expected returns, cash flow news is the main determinant of cross-sectional stock expected returns.

How to identify firm-specific information is an important empirical question. The literature has long been accepted that quarterly earnings announcements are imperfect approximations. However, Tetlock, Saar-Tsechansky, and Macskassy (2008) document that firm-specific news articles, extracted from the Dow Jones Newswire Archive (DJNS hereafter), predict firms' future earnings, analysts' forecast revisions, and stock expected returns. This highly influential paper suggests that incremental information is embedded among these news articles arriving every day, rather than merely restricted to earnings announcements that are released quarterly.

In fact, the very first empirical study of DJNS news articles is Chan (2003). Using DJNS news headlines to match individual stocks, he finds that these stocks with news-driven price shocks tend to exhibit return momentums, whereas those with non-news-driven price shocks are followed by return reversals. The profound impact of this paper not only lies in generalising the post-earnings announcements drifts (known as PEADs) to post-news drifts, but also highlights the promising implications of DJNS news articles in the cross-sectional stock return studies. In that case, the news seemingly explains the momentum phenomenon, one of the most distinctive and well-known anomalies in the literature.

It should be clear that newswires are very different from newspapers. In particular, the DJNS news used in this thesis comprises *DJNS intraday news* articles and *the Wall Street Journals*, which are mainly subscribed to by professional investors. In contrast, a newspaper such as *The New York Times* and *The Washington Post* aim at the general public, who are most likely less sophisticated investors. Prior studies also confirm this difference. For example, Fang and Peress (2009) note in the literature review section, "*'news' and 'coverage' are indeed different: Many stocks with news (headlines in the Dow Jones Newswire) remain neglected by mass media; in addition, while newswires are released in real-time and contain genuine information, this is unlikely to be the case for mass print media due to publication lags.*" Consequently, it is important to note that some findings from this thesis might not be cross-validated using general newspapers.

1.2 Intended Contribution and Outline of the Thesis

This thesis attempts to advance the anomaly literature by taking a closer look at how DJNS news plays a role in different anomaly settings. The scope of my empirical work rather comprehensively examines a large set of anomaly variables, but aims to refine and explain one category of anomalies in each chapter. To do so, I first construct a firm-level news database based on the Dow Jones Newswire Archive between 1979 and 2017. The dataset includes news sentiment, the number of news items per month, article topics, etc.[1]

The first set of empirical analysis investigates lottery-type stocks. A well-known asset pricing anomaly "MAX", measured by the maximum daily returns in the past month, can depict the stocks' lottery features and investors' gambling behaviours. Prior studies document that MAX stocks tend to have a negative cross-sectional return pattern due to investor lottery preference (Bali, Cakici, and Whitelaw, 2011). According to the cumulative prospect theory of Kahneman and Tversky (1979), investors tend to overweight a small probability of extremely large pay-offs. Hence, stocks with such features are likely to be overvalued with a subsequent return reversal. Motivated by MAX stocks exhibiting high information uncertainty and equity's option-likeness (Barinov, 2018), I consider an empirical setting on how the presence of DJNS news affects these lottery-type stocks. I extend prior findings by showing an augmented negative relationship between MAX stocks without news and future expected returns. More importantly, MAX stocks with news generate return momentum. My empirical results suggest that news does have pricing effects on lottery-type stocks, which is best explained by information

[1] "News sentiment" refers to the sentiment score of news content assessed using computational tools. A news article with high news sentiment scores tends to contain good news, the details of which are elaborated in chapter4.

uncertainty mitigation when news arrives in the market.

The second set of empirical analysis looks into the lead-lag return momentum. A large body of literature shows that the peer firms' past returns can predict a base firm's future expected returns. For instance, there are lead-lag return effects among the same industry (Moskowitz and Grinblatt, 1999); supplier-customer chain (see Cohen and Frazzini (2008) and Menzly and Ozbas (2010)); firm within same product market (Hoberg and Phillips, 2018); or single- and multi-segment firms operating in the same sector (Cohen and Lou, 2012). The empirical phenomenon is consistent with investor inattention patterns where they tend to underreact to a public news announcement (Hong and Stein, 1999; Barberis, Shleifer, and Vishny, 1998). While a large body of literature continues to document new industry schemes with associated asset pricing implications to compete with existing schemes, relatively little work has focused on "sticky peers" in these industry schemes. Taking the Standard Industry Classification (SIC) code as an example, it is unlikely to timely reflect the updates and changes among those emerging industries, such as HighTech sector. By applying a "co-coverage" concept to DJNS news, I construct a time-varying firm-centric grouping scheme (news-based peers, NBPs) aiming to augment existing industry schemes. I show that the base firm's share prices respond more favourably to the stock returns of its NBPs than of its traditional industry peers that are not NBPs. The additional results show that NBPs can partially digest several lead-lag return momentums previously documented in the literature.

Chapter 4 investigates the role of news sentiment in between past stock returns and subsequent investor reactions. Specifically, I consider an empirical setting where short-term return reversals are combined with the aggregated monthly news sentiment scores. I show that predictable future stock returns after the news are more likely to emerge when *ex ante* monthly aggregated news sentiment scores and contemporaneous returns are mismatched – referred to as slowly incorporated news stocks. In contrast, those stocks that quickly incorporate news into contemporaneous stock prices do not generate predictable future returns. A trading strategy of investing those slowly incorporated news stocks generates 139 basis points per year and is robust after considering transaction costs. Overall, my research refines the role of news regarding information dissemination in the financial markets.

Chapter 5 provides some suggestions for future research. While this thesis only looks into the relation between firm-level news and different anomalies, other types of news are worth exploring, such as macroeconomic news announcements, industry news events, and overnight news. For instance, Hirshleifer, Lim, and Teoh (2009) document that a firm's post-earnings announcements drift tends to be stronger when same-day other industry news arrives beside the base firm. This phenomenon is termed as the attention distraction effect: investors reallocate attention to other attention-grabbing events. This might be worth applying to an empirical setting where lottery-type stocks are combined with the same-day industry news. Similarly, it might also be interesting to consider in an empirical setting where overnight news is linked with those anomalies, as little prior studies have examined this type of news.

Other potential extensions can be attempted as well. For example, the empirical setting of this thesis is only based on the US equity market, which has limited implications for international investors. Perhaps, replicating anomalies in different markets is the best way to examine whether these findings are robust and not subject to data snooping. In future research, it may be worth constructing an alternative news data set using local newspapers from non-US countries.

This thesis also includes two sets of appendices covering the additional analysis from Chapter 3 and 4, respectively. In Appendix A.1, I outline the detailed steps of NBP construction. The example of newsworthy links is included in Appendix A.2. A comparison of article-level versus sentence-level news co-coverage is outlined in

Appendix A.3. The other set of appendices include the additional data collection process of the Google Search Volume Index and Bloomberg Attention Index.

All in all, throughout the three main chapters, a variety of empirical analysis shows that DJNS news indeed drives different short-term stock return patterns in these research settings. Taken together, the thesis is expected to offer a further understanding of DJNS news pricing effects and associated underlying explanations in different anomaly settings.

Chapter 2

When is a MAX not the MAX? How News Resolves Information Uncertainty[1]

2.1 Introduction

Recent literature has depicted investor gambling behaviour and lottery preferences in various empirical settings (e.g., Kumar (2009b),Bali et al. (2011),and Barberis, Mukherjee, and Wang (2016)). The MAX effect, documented by Bali et al. (2011), is shown to be an effective proxy. Intuitively, the effect is a cross-sectional stock return pattern: stocks with high maximum daily returns (denoted as MAX) in the prior month tend to experience negative expected returns, whereas low MAX stocks typically demonstrate high returns in the following month. The monotonically decreasing relationship between the MAX and expected returns is consistent with Cumulative Prospect Theory (Tversky and Kahneman, 1992) where investors tend to overweight a small probability of an extremely large pay-off and

[1]An edited version of this chapter appears in the Journal of Empirical Finance (Tao, Brooks, and Bell, 2020b).

push up the contemporaneous prices of these stocks with lottery-like features due to their preferences. As a result, overvalued high MAX stocks tend to exhibit low expected returns in the next month and vice versa.

While the empirical MAX finding is well-acknowledged, the literature has not yet distinguished these extreme daily return events that occur in the absence (or presence) of any public news stories, namely in the business press. As the pricing effect of professional newswires has received incremental attention recently in the financial markets (e.g., Chan (2003),Tetlock et al. (2008),Fang and Peress (2009),Tetlock (2010),Shi, Liu, and Ho (2016),Jiang, Li, and Wang (2017),Wang, Zhang, and Zhu (2018) and Bali, Bodnaruk, Scherbina, and Tang (2017a)), it is important to explore and discuss what the overall effect of news coverage will be on the future performance of MAX stocks. Therefore, in this chapter, I take a unique perspective to examine MAX effects conditioning upon the arrival of news reports in a 3-day window on each MAX day and I ask the question, when the business press is associated with a MAX, do the previously documented cross-sectional return patterns (i.e. MAX effects) still exist?

There are two important building blocks of this chapter: First, investors' gambling behaviour is highly correlated with information uncertainty and equity's option-likeness (e.g., Barinov (2018)). Second, the business press can resolve information uncertainty (e.g., Tetlock (2010),Shi et al. (2016),and Bushee, Core, Guay, and Hamm (2010)). It can then be argued that negative future returns are more likely to occur for non-information-related MAX stocks because they appear to embody more uncertainty. In contrast, it might be difficult to observe such patterns for stocks with information-related maximum daily returns. More intuitively, suppose that there are two stocks, A and B, which both achieve 20% returns on a day of the prior month, but the difference between them is that the extreme daily return event for A is covered by news articles, while for stock B it is not. As a result, when investors are informed by these news articles about stock A, the future impact of the MAX event becomes easy-to-value/less uncertain and it would not make sense for them to speculate on this lottery-like stock as they had planned. In other words, investor lottery demand is likely to be attenuated as the information uncertainty is at a low level. As for stock B, investors can only infer the cause and continue to act as short-term speculators. Therefore, stocks like B exhibit strong lottery-like features.

To examine these conjectures, I utilize a unique U.S. stock-level news dataset between 1979 and 2016. This dataset particularly fits my research purposes for several reasons: First, since it contains both *Dow Jones Intraday News* and news from *The Wall Street Journal* produced by the Dow Jones Newswire Archive,[2] the data can provide broad coverage of various topics for different corporate actions including quarterly earnings, takeovers, analyst recommendations, insider buying and selling, dividend news, bond and stock registrations, labour, unions and strikes, etc. Second, news reports in the dataset are timely and newsworthy so that material news events can be promptly read by investors. Thus, the economic linkages between these abnormally high returns and news reports are more likely to be confirmed. Finally, the large circulation of news collections ensures a high volume of readership and attention by investors. In other words, because of the news coverage, a wide range of investors are aware of these unexpected firm events.

I design a research setting in which the presence of stock-level news articles is combined with lottery-like stock investing. The main conjecture to examine is whether investors prefer gambling on MAX stocks with news (denoted as MAX_{news}) or without news (denoted as MAX_{nonews}). I hypothesize that the MAX effect is more profound among MAX_{nonews} stocks, whereas MAX_{news} stocks might see very

[2]Dow Jones Newswire is a global leading real-time news product and is commonly used in many research papers – (e.g., Tetlock (2010),Tetlock (2011),Engelberg, Reed, and Ringgenberg (2012),Engelberg et al. (2018), and Boudoukh, Feldman, Kogan, and Richardson (2018)).

little evidence of this activity. If the intuition is correct, I should observe that only the expected returns of MAX_{nonews} stocks decrease with their growing prior-month maximum daily returns in an empirical setting. To test this conjecture, in each month, I partition stocks into either the MAX_{news} portfolio, where the MAX can be linked to at least one news article in the prior month, or the MAX_{nonews} portfolio, where the MAX is not associated with any news coverage during the past month. Next I construct deciles within each portfolio: I sort ten different groups based on the MAX value in each month. The return performance of each decile is then traced in the following month and I report the time-series average value of expected returns, respectively.

Three main sets of results are provided in this chapter. First, the empirical evidence confirms my intuition. I do observe that the MAX_{nonews} and MAX_{news} classifications predict that future returns will be in opposite directions. The empirical findings are striking: the MAX effect, calculated by the return spread between the highest and the lowest decile, is -1.74% per month from the MAX_{nonews} portfolio, whereby the return spread of the MAX_{news} stocks is 1.34%, indicating a positive return momentum pattern. Furthermore, Fama-MacBeth regressions suggest that the magnitude of augmented MAX effects by the MAX_{nonews} portfolio is much larger than the original MAX one. Also, I note when comparing the MAX and MAX_{news} portfolios that the signs of the coefficients have different directions.

Next, I conclude that the opposite direction of MAX_{news}'s expected returns is likely to be explained by information uncertainty mitigation instead of investor attention or divergence of opinions. As for investor attention theory, news-induced attraction is unlikely to explain the future return patterns in MAX_{news} stocks. In fact, Bali, Hirshleifer, Peng, and Tang (2019) document that retail investor attention is an important driver of lottery stocks' over-valuation. My finding is a positive future return relationship for MAX stocks with news coverage. To rule out the investor attention channel, I further control for three investor attention proxies in the regression including the volume of DJNS news reports, the number of analysts covering the stock, and the absolute value of standardized unexpected earnings. The results are not influenced by these controls. To investigate the alternative explanation that a divergence of opinions drives MAX_{news}'s return momentum, I examine the other side of the stock return distribution – minimum daily returns (denoted MIN). If news articles induce a divergence of opinions regarding a stock's fundamentals, and the stock should therefore offer high future returns to compensate for higher risks, I should observe that both MAX and MIN stocks' returns behave in the same manner. However, my results clearly reject this premise. Finally, motivated by the role of news stories regarding resolving information asymmetry (Tetlock, 2010), I conjecture that MAX_{news} stocks are associated with less information uncertainty upon the arrival of stock-level news. My empirical results are supportive: time-varying tests clearly show that the 5-day return volatility declines by 1.95%(0.91%) when the MAX day is (isnot) covered by DJNS news reports. Bivariate portfolio tests report that all MAX_{news} future stock returns are insignificantly different from zero under high information uncertainty, indicating that news erases the lottery-like features of these stocks.

Obviously, to what extent investors prefer engaging in lottery-type behaviour depends on their risk attitudes. When investors are overoptimistic, they tend to overestimate the probability of gains and underestimate the probability of losses. Thus, they are more likely to bet on a high MAX stock. In contrast, investors who are overpessimistic are unlikely to gamble in the market. Instead, they are probably looking for safe, high-quality assets. I explore this issue by interacting my baseline results with six different market-aggregated investor sentiment indexes. The empirical results show that the MAX_{nonews} portfolio performs better when investor sentiment is high, whereby the return momentum exhibited by the MAX_{news} portfolio is largely driven by low sentiment periods. This evidence generally supports the findings of previous studies: MAX effects are stronger during high investor sentiment periods (Fong and Toh, 2014) and investors tend to exhibit a strong "flight-to-quality" when sentiment is weaker (Bethke, Gehde-Trapp, and Kempf, 2017).

In addition, a number of robustness checks are performed. First, there is a concern that MAX_{news} stocks might be associated with noisy news. For example, some news articles contain less information content. Therefore, it is hard to conclude that all MAX_{news} stocks are information-related. To address this issue, I narrow down the topic of the news articles to only major corporate events and I reduce the matching window to only trading daytime (i.e. 9:30 am to 16:00 pm) on day t. The results remain quantitatively similar. Furthermore, I am concerned that the expected return of my portfolios may have seasonal effects. For instance, a large number of firms might choose to disclose important information such as 10-Ks in the same month. Seasonal anomalies such as the "January Effect" might partially drive the return momentum of the MAX_{news}. However, the seasonality tests that I perform eliminate these possibilities. Lastly, I also attempt to address the concern that the MAX_{news} results are merely a replication of size effect, rather than a new phenomenon: MAX_{news} stocks on average have bigger market capitalization than MAX_{nonews} stocks, similar to the conclusion by Bali et al. (2011) that the lottery demand effect is more pronounced among small stocks. I find strong evidence against the suggestion that the MAX effect is predominantly driven by size.

My study contributes to the literature in several aspects. First, to the best of my knowledge, this chapter is the first which comprehensively examines the role of news in the study of MAX stocks. The nearly 40 years' of news data at my disposal allows me to extend the sample period where both pre- and post-internet investor gambling behaviour can be investigated. Second, my work can be considered a significant extension and generalisation of Nguyen and Truong (2018). While they observe no evidence of the MAX effect when MAXs are driven by earnings announcements (denoted EA), I find that the important distinction is not between MAXs associated with earnings (or not), but rather between MAXs associated with news more widely (or not). Third, the results also confirm the news momentum pattern documented in the previous literature (e.g., Jiang et al. (2017),Wang et al. (2018)). I note that MAX_{news} stocks yield a significant return momentum of 1.34% per month (equivalent to 16.08% per year). Collectively, the results in this chapter highlight that news plays an important role in explaining investor lottery preferences.

The remainder of the chapter is organized as follows: Section 2.2 discusses the related literature and develops the main hypotheses. Section 2.3 introduces the DJNS news data, lottery-like proxies, and portfolio construction; Section 2.4 presents three main sets of empirical analyses plus additional robustness checks. Finally, Section 2.5 provides a summary and concluding discussion.

2.2 Literature Review and Hypotheses Development

This chapter touches on several strands of the existing literature. It relates to a growing field regarding the role of news articles and media coverage. The business press (e.g., the DJNS Newswire) can impact capital markets in three non-mutually exclusive ways. First, professional news coverage plays a dissemination role by broadcasting events to investors in general. Second, it also has an information creation role by producing new information for the market. Third, it can serve as a proxy for investor attention due to investor behavioural bias. The prior literature has documented the different roles of news coverage regarding its pricing (mispricing) effect on financial markets in various empirical settings. For example, using news articles from the Dow Jones Newswires Archive, Chan (2003) finds that no-news stocks tend to exhibit short-lived return reversals and news stocks tend to exhibit post-news drift. Tetlock et al. (2008) show that firm-level news stories can predict future SUE (Standardized Unexpected Earnings) and stock returns. Subsequently, Tetlock (2010) documents that news can resolve information asymmetry. Bushee et al. (2010) explore the role of business news coverage in terms of information dissemination. Fang and Peress (2009) find that stocks without news coverage exhibit higher returns in the future. They attribute this anomaly to an "investor recognition" hypothesis, where stocks with low investor recognition need to provide a higher return to compensate investors. Moreover, Hillert, Jacobs, and Müller (2014) find that a high volume of media coverage leads to strong momentum profitability in a longer formation period based on an unique newspaper dataset and they further attribute this to media-induced investor overconfidence. Collectively, these studies suggest that news coverage has pricing (mispricing) effects in different empirical settings.

The literature also examines the link between the arrival of public news and the idiosyncratic volatility puzzle. For example, Shi et al. (2016) argue that the negative relationship between idiosyncratic volatility and expected returns could be a result of public news arrival, where bad (good) news can simultaneously increase (decrease) contemporaneous idiosyncratic volatility and decrease (increase) expected returns. DeLisle, Mauck, and Smedema (2016) find that the high volatility of no-news stocks leads to lower expected returns and that news stocks' volatility is positively priced. Moreover, Bali et al. (2017a) document that unusual firm-level news flow contributes to volatility shocks and temporarily increases the level of investor disagreement about a firm's fundamentals. While high idiosyncratic volatility poses a barrier to short selling by pessimistic investors in the short-run, stock prices tend to initially rise but then trend downward towards consensus opinions, which results in stock return underperformance. Since maximum daily returns and idiosyncratic volatility are nearly 90% correlated (e.g., Bali et al. (2011) and Egginton and Hur (2018)), it seems interesting to extend this strand of research by studying how news plays a role in the negative relationship between MAXs and expected returns.

One necessary condition to develop my hypothesis is that high MAX stocks also have high information uncertainty and act as option-like equity. Using several firm-specific volatility/uncertainty measures including idiosyncratic volatility, analyst disagreement, analyst forecast error, and volatility of earnings and cash flows, Barinov (2018) shows that MAX stocks are positively correlated with these volatility/uncertainty measures. Furthermore, Barinov (2018) also documents a positive relationship between high MAX stocks and the option-likeness of equity, proxied by credit ratings and O-scores (a measure of the expected probability of bankruptcy): high MAX stocks tend to have worse credit ratings and a higher probability of bankruptcy. They act like a call option on the asset with strike prices equal to the debt. If firm-specific volatility/uncertainty increases, the value of the option (i.e. the MAX stock) is likely to increase, all else equal. Therefore, Barinov (2018) observes that the MAX effect is stronger for stocks with high information uncertainty.

Although the three different roles that news coverage plays has been thoroughly discussed by the prior literature, it is unclear whether and how the press influences the market pricing of the maximum extreme daily returns. With respect to the information dissemination role, broader dissemination under each MAX day should increase the visibility of the underlying corporate news events if they exist. As the cost of information acquisition reduces, investors are more likely to be aware of these events and the information uncertainty issue should be attenuated. As a result, one can expected much weaker lottery demand effects of these MAX stocks due to low information uncertainty. However, it is also likely that the newswire contains a few “news flashes”. For example, a news item could just provide a summary of the top-ten hot stocks in a day, one of which has a 20% extreme daily return.

Without editorial content, this news story only disseminates bottom-line information regarding the extreme return of a stock and information uncertainty issues will remain. In this case, it is hard to conclude that lottery demand for MAX stocks will be eliminated.

Second, the news coverage might impact the future performance of MAX stocks through its information creation role. On the one hand, the newswire contains a large number of full press articles with credible and timely detailed content that helps investors better understand these extreme daily return events. With all information collected from different sources and put together, journalists may effectively digest a complex MAX event and package a news story in a way that is more easily understood by investors. As a result, this news coverage helps investors understand the implication of the MAX and potentially reduce the information uncertainty. On the other hand, different business reporters may have different opinions and comments regarding the same events. While some might publish positive views and opinions about a firm, others could hold negative expectations regarding the future performance of MAX events. This news-induced divergence of opinions will lead to high idiosyncratic volatility, temporarily preventing pessimistic investors from short selling before stock prices revert to the consensus (Bali et al., 2017a). Therefore, one could expect that stock prices will experience a rise and then a fall.

Lastly, there are reasons to believe that business news coverage is subject to behavioural forces, which will cause the market's overreaction or underreaction. For example, journalists can produce a sensational news report in order to catch investors' attention or cater to investors' demand for each MAX day. Such news articles can distort the truth and mislead investors to take inappropriate actions – overvalue a firm. As a result, it can be observed that stock prices continue to rise in the following period (e.g., Hillert et al. (2014)). Meanwhile, the newswire could also foster investor underreaction (e.g., Chan (2003), Jiang et al. (2017), and Wang et al. (2018)). Unlike sophisticated institutional investors who closely monitor press distribution channels, retail investors only view newswire services periodically and adjust their portfolio accordingly, causing a delayed reaction. Furthermore, limited attention theory predicts that firms with certain characteristics such as small size, lower analyst coverage and with less institutional ownership tend to have an inferior information environment (Da, Gurun, and Warachka, 2014b). It takes a longer time for these firms to fully incorporate the information embedded in the news articles.

Two broad conclusions can be drawn from the prior literature. First, investors' gambling behaviour is highly correlated with information uncertainty and equity's option-likeness. Second, the different roles that the business press plays has received incremental attention in the literature, but evidence regarding whether it increases or decreases information uncertainty is mixed. Thus, this study aims to provide direct evidence on how news coverage influences stocks' information uncertainty and further deters or exacerbates investors' lottery demand.

Ex ante, it is unclear what the overall effect of news coverage will be on the future performance of MAX stocks. The general tendency of the research discussed above is that stocks with news coverage tend to produce higher future expected returns, whereas stocks without news coverage tend to have lower future expected returns (e.g. Chan (2003), Tetlock (2010), Frank and Sanati (2018), and Jiang et al. (2017)). Two recent works focusing on news coverage and lottery-like stocks are noteworthy: Nguyen and Truong (2018) document that lottery-type investment will be attenuated when the MAX is triggered by earnings announcements. Presumably, if the business press covers these earnings announcements events and disseminates this information to the masses on each MAX day, it is likely that a similar or perhaps enhanced positive return pattern will be observed. The other work is Shi et al. (2016), who argue that the negative relationship between idiosyncratic volatility and expected returns could be a result of public news arrival. In other words, good news can decrease idiosyncratic volatility in the current month and increase stock returns in the following months. When turning to MAX stocks, it is very likely that news coverage on each MAX day constitutes good news. Thus, it seems reasonable to infer that MAX_{news} stocks will have lower firm-specific volatility and higher future returns.

Collectively, this discussion motivates my first hypothesis. The overall effect of news coverage on MAX stocks is as follows:

Hypothesis 1 MAX_{news} stocks tend to produce higher future expected returns, whereas MAX_{nonews} stocks tend to have lower future expected returns.

My second hypothesis is relevant to the mechanisms through which the effect of news coverage is responsible for the opposite performance of MAX_{news} and MAX_{nonews} stocks. News helps reduce uncertainty and incorporate information into stock prices, there should be direct implications for stock return volatility and stock price efficiency. As a prior study has documented that MAX stocks also have higher information uncertainty and act as an option-likeness of equity (Barinov, 2018), one can expect that lottery-type demand will be attenuated among stocks with lower information uncertainty/firm-specific volatility. Meanwhile, the previous discussion has summarized three different roles of news coverage through which it can mitigate the stock's information uncertainty. The intuition behind this is that the news coverage's dissemination role could increase the visibility of extreme daily return events and underlying corporate events if they exist. This reduces information asymmetry for market participants, especially uninformed investors, who will revise their beliefs upon news arrival. Consistent with this theory, Tetlock (2010) finds that a ten-day volume-induced momentum exists only on news days. Moreover, information uncertainty can also be mitigated through the press' information creation role. As various information is aggregated by journalists from multiple sources and articles offer much detailed content and deep insights, investors can better understand those complex extreme daily return events and rely on this credible content to make investment decisions rather than simply speculating. As a result, it seems sensible to hypothesize that MAX_{news} stocks have lower information uncertainty and deter investor lottery-type demand.

On the other hand, complex news events might lead to investor disagreement regarding the true value of a stock. For example, Bali et al. (2017a) document that stocks with unusual news flows tend to increase firm-specific volatility on which pessimistic investors find it difficult to short sell. So one can observe that the stock initially has a large rise and then underperforms its peers. Furthermore, the newswire's information creation role will lead to an issue in that different journalists might have different views and opinions towards extreme daily return events. The existence of both positive and negative news stories around the future implications of MAX stocks will induce a large idiosyncratic volatility and greater information uncertainty. Therefore, one can predict that MAX_{news} stocks have higher information uncertainty and will exacerbate investor lottery-type demand.

Given the discussion above, my second hypothesis is as follows:

Hypothesis 2 The arrival of news articles mitigates a stock's information uncertainty rather than inducing a divergence of opinions, therefore deterring investor lottery-type demand.

Lastly, the extent to which investors prefer to engage in lottery-type investing is highly associated with their risk attitude. For example, an overoptimistic investor tends to overestimate the probability of gains and underestimate the probability of losses. Thus, he or she is more likely to bet on a high MAX stock. On the other hand, overpessimistic investors tend to pursue safer assets and thus they are unlikely to gamble in the market. To explore the interaction between MAX stocks and associated investor risk preferences, I attempt to utilize market-level investor sentiment indexes as proxies. Other research has identified that the MAX effect is stronger when investor sentiment is more positive (Fong and Toh, 2014), which is explained by the theory that investor optimism induces a high level of lottery-type demand. Such a phenomenon is unlikely to be observed during low sentiment periods when investors tend to be pessimistic and more risk-averse. Therefore, I hypothesize that the augmented MAX effect should also align with this argument. Moreover, investors are more likely to follow a "flight-to-quality" strategy when sentiment is bad. It is then expected that the stock return predictability of MAX_{news} should be more persistent. So my third hypothesis is as follows:

Hypothesis 3 Short-term speculation is more persistent during high sentiment periods whereas long-term investment should be observed when sentiment is weaker.

Testing Hypothesis 3 requires market-aggregated investor sentiment data. As my study focuses on the U.S. market, I am able to obtain a number of available sentiment indices. Previous studies have developed six investor sentiment indices: the Baker and Wurgler (BW) sentiment index from Baker and Wurgler (2006); the FEARS (Financial and Economic Attitudes Revealed by Search) index constructed by Da, Engelberg, and Gao (2014a), who collect a set of high Google search volume financial and economic keywords by households as a measure of retail investor interest; the MS (Manager Sentiment) index from Jiang, Lee, Martin, and Zhou (2019) aggregating firm-level manager sentiment proxied by 10-Ks and 10-Qs; economic indicators, the MCSI (the Michigan Consumer Sentiment Index) and the CBC (the

Conference Board Consumer Confidence Index); and an optimal BW Sentiment Index named HJTZ from Huang, Jiang, Tu, and Zhou (2015).

2.3 Data Description

2.3.1 Dow Jones News and Stock Data

I use a self-constructed U.S. stock-level news data set between July 1979 and December 2016.[3] The underlying news items in this data set are produced by the Dow Jones Newswire and published either on the *Dow Jones Intra News* newswire or in the *The Wall Street Journal* newspaper. Stock-level news is collected when the ticker tagged in the article matches the stock's within the valid period. By doing so, I obtain a group of stocks with at least one news release during the lifespan. This provides data for 14,079 stocks.

The dataset I utilize to calculate lottery-like proxies includes all common stocks listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX) and Nasdaq Stock Exchange (NASDAQ), which is downloaded from the Center for Research in Security Prices (CRSP). The common stock universe excludes all stocks such as preferred stocks, warrants, unit or investment trusts, funds, REITs, ADRs and duplicates. The daily and monthly return data covers the period from July 1979 to December 2016 in order to match the sample range of the news data set. The Fama-French risk factors are obtained from Kenneth French's data library including the excess market return (MKT), small-minus-big (SMB), high-minus-low (HML), up-minus-down (UMD); the liquidity factor (PS) is downloaded from Lubos Pástor's data library.

[3] I acknowledge that a number of studies utilise RavenPack, which is a pre-constructed firm-level news database, but I decided to hand-construct the data for two reasons. First, the sample period provided by RavenPack starts from January 2000. The limited data period poses constraints on lottery-type stock studies. Second, RavenPack does not provide the original text, which prevents further studies on textual analysis.

2.3.2 Lottery-like Proxies

Several different lottery-like proxies are employed among the literature. The one utilised in this chapter is the MAX according to Bali et al. (2011), which takes the maximum daily return in the prior month. I also require that there must be at least 15 trading days for each stock in every month. When forming the portfolio, I exclude all below $5 per share stocks at the end of formation period to avoid bid-ask bounce and illiquidity. The calculated MAX effect is therefore stronger in an equally-weighted portfolio than a value-weighted one.[4] In addition, I also compute the average of N maximum daily returns in the prior month, with N ranging from 2 to 5. These alternative MAX proxies do not alter my main finding.

2.3.3 Construction of MAX_{news}

To link the MAX with Dow Jones news reports, I start by setting up a [t-1,t+1] interval for each MAX to capture any news coverage. The three-day window allows me to address the possibility of any potential information linkage and subsequent press coverage.[5] MAX_{news} stocks then refer to those which have at least one news article located within this window in a given month. For example, a stock's maximum daily return in the month of May is May 23rd, precisely. Then the three-day window will be May 22nd to 24th. Any news reports within this time window are considered to be relevant and the stock is then identified as a MAX_{news}. Next, I construct a set decile portfolios, ranking on average returns within the set of MAX_{news} stocks. I apply a calendar-time portfolio approach, and specifically, in each month, all MAX_{news} stocks are sorted into ten different subsets based on their MAX value. As a result, the highest group contains all stocks with the highest maximum daily

[4]This finding is consistent with Bali et al. (2011) when they exclude all below $5 per share stocks.

[5]In robustness checks, I employ different time intervals including a [t-2,t+2] and [t-3,t+3] window. My findings remain qualitatively similar.

stock returns, which are likely to be treated as lottery-type stocks. On the other hand, the lowest group accommodates all the lowest maximum daily return stocks.

——————————— Insert Table 2.1 here ———————————

2.4 Empirical Analysis

2.4.1 Univariate Portfolio Analysis - MAX_{news} and MAX_{nonews}

I am primarily interested in stock characteristics across each MAX_{news} and MAX_{nonews} group. To do so, I perform an univariate portfolio analysis. Specifically, I sort stocks into deciles based on the MAX in the prior month for each month, and then link a variety of stock characteristics for each decile. The average value of each characteristic is calculated across all stocks in each group. The computation is then repeated for every month between August 1979 and December 2016. Finally, I report the time-series averages of the decile-level characteristics during the sample period.

As can be seen in Panel A of Table 2.2, media-relevant statistics are reported including the total volume of news coverage (denoted as MEDIA), sentiment score of news content computed by the Loughran and McDonald (2011) dictionary method (denoted OPT)[6] Noticeably, MEDIA monotonically rises along with the portfolio's index, suggesting that a high return is associated with a high volume of press releases. OPT witnesses an upward trend, albeit with non-linearity, which reflects a more positive word percentage in the highest portfolio. Moreover, the lottery-feature statistics are presented. Consistent with Kumar (2009b), high MAX stocks tend to have lower prices, higher idiosyncratic risk and higher idiosyncratic skewness

[6] The Loughran and McDonald (2011) dictionary method provides word lists in which all words are categorized as having positive or negative sentiment. A sentiment score is then calculated by taking the positive percentage of words minus the negative percentage of words and dividing by the total number of words in a given article.

(denoted PRC, IVOL, and ISKEW, respectively). Furthermore, I report other MAX decile characteristics such as SIZE, BTM, TURN, REV, MOM, BETA, and ILLIQ. These statistics suggest that MAX_{news} stocks entering the highest portfolio tend to have small market capitalization, lower book-to-market ratio, higher turnover ratio, display positive short-term reversal and momentum patterns, have higher positive exposure to market risk, and are more likely to be illiquid, which is consistent with Bali et al. (2011).

In Panel B, the decile-level characteristics of MAX_{nonews} stocks are reported. Interestingly, I find that the MAX value of MAX_{nonews} is consistently lower than that of MAX_{news} across each decile. This suggests that a media-associated MAX is generally stronger than a non-media-associated MAX. Moreover, compared to MAX_{news} stocks, those in the MAX_{nonews} portfolio tend to have smaller SIZE; higher BTM and TURN; lower REV, ISKEW, and BETA; and higher ILLIQ. Overall, the characteristics of MAX stocks associated with news coverage do not exhibit any significant differences compared to the original MAX stocks, although there are a few differences between MAX_{news} and MAX_{nonews} stocks regarding some firm characteristics.

——— Insert Table 2.2 here ———

2.4.2 Baseline Results

To evaluate post-formation return performances, I perform a calendar-time portfolio approach with a one-month window to trace holding period returns across different portfolios. Specifically, in each month, all stocks are sorted into deciles based on the MAX calculated in the prior month. I then trace the post-formation period return for each decile. The final tabulated results are reported by taking the average of time-series returns for the entire sample period across deciles.

I start by replicating the MAX effect during the sample period (i.e. August 1979 to December 2016) on both equally-weighted and value-weighted schemes. Following Bali et al. (2011), the MAX represents a stock-level lottery-feature. The high value of the MAX suggests that investors have high preferences for investing in such stocks. Demand pressure then pushes up the current price of high MAX stocks, which leads to lower future returns in the next month. Consistent with this theory, it can be observed that the stock return monotonically decreases from low MAX stocks to those with high MAX. Looking closely at the hedged return difference between the lowest and highest MAX deciles, I report significant lottery-type demand premiums on both an equally-weighted and a value-weighted basis, namely -0.97% (-0.78%) per month with a t-statistic of -3.48 (-2.38) for the equally-weighted (respectively, value-weighted) scheme in Table 2.3.

Next, I move to examine the performance of MAX_{nonews} and MAX_{news}. The tabulated results are computed by taking the average of time-series raw returns across each portfolio in Table 2.3. As can be seen clearly, the next-month return of the MAX_{nonews} portfolios monotonically decrease, from 0.93% per month for the low decile portfolio to -0.81% for the high decile portfolio. The hedged portfolio return difference is highly significant (t-statistic: -7.13). After controlling for Fama, French and Carhart four factors and liquidity-augmented five factors, it remains statistically significant. Looking at the coefficients, the magnitude of MAX_{nonews} almost doubles when compared to the MAX, suggesting that stocks without any news coverage are more likely to be of the lottery-type. Moreover, the future return of MAX_{news}, conversely, rises from decile 1 to decile 10; the difference is 1.34% per month (t-statistic: 3.08). This indicates that MAX stocks, if covered by news reports, lose their lottery-type features and earn positive returns driven by the information embedded in the news stories.

In addition, I examine the statistical difference between MAX_{nonews} and the

MAX_{news} returns. The hedged portfolio return is significant at the 1% level for both EW and VW weighting schemes, suggesting that the MAX stock with and without news coverage is indeed different. Overall, the baseline results clearly show that the MAX effect will be enhanced if no news coverage is associated with the high maximum daily return, whereby the effect will disappear upon news arrival.

——————————— Insert Table 2.3 here ———————————

2.4.3 Bivariate Portfolio Analysis

I next examine the two types of stocks, MAX_{nonews} and MAX_{news}, after controlling for major stock characteristics, namely: $Size$, BTM, $BETA$, MOM, REV, and $ILLIQ$. The aim of this exercise is to examine whether the cross-sectional return pattern I documented is actually driven by other stock characteristics. I design the empirical exercise as follows: In each month, I sort stocks into deciles based on one of the control variables. Within each decile, I then sort stocks into another set of deciles based on MAX. The future returns of the ten MAX-decile portfolios are then averaged across different deciles of the control variable. For example, the final tabulated future return of the MAX-decile 1 is the average future returns of MAX-decile 1 within each size-decile group when I control for firm size.

I report the results of this test. Due to space constraints, I mainly focus on the difference between MAX_{nonews} and MAX_{news} for both equally- and value-weighted portfolios. Each column represents a control variable. Within each column, the return spread between decile 10 and decile 1, the Fama-French-Carhart four-factor alphas, and the liquidity-augmented Fama-French-Carhart five-factor alphas are presented. As can be seen in Table 2.4, the difference between MAX_{nonews} and MAX_{news} is consistently significant across all specifications.

Collectively, the results show that DJNS news articles are an effective sorting variable between MAX_{nonews} and MAX_{news}, even after controlling for stock characteristics.

——————————— Insert Table 2.4 here ———————————

2.4.4 Fama-MacBeth Regressions

In this section, I continue to study the cross-sectional stock return predictability of the MAX with or without news coverage in the Fama and MacBeth (1973) regression framework. To be more specific, I first examine the interaction term, MAX_{news}, defined as the portfolio of MAX stocks with at least one news coverage item in the [t-1,t+1] window in the prior month, and MAX_{nonews} denotes those MAX stocks without any news coverage in the three-day interval. The regression is constructed in the following way: for each time point, I perform a cross-sectional regression for all observations at that time. The time-series averages of the coefficients are then calculated and reported in Table 2.5.

The regression equations are constructed for each setting as follows:

$$Ret_{i,t+1} = \alpha + \beta_1 * MAX_{i,t} + \beta_2 * X_{i,t} + \epsilon_{i,t} \tag{2.1}$$

$$Ret_{i,t+1} = \alpha + \beta_1 * MAX_{nonews,i,t} + \beta_3 * X_{i,t} + \epsilon_{i,t} \tag{2.2}$$

$$Ret_{i,t+1} = \alpha + \beta_1 * MAX_{news,i,t} + \beta_3 * X_{i,t} + \epsilon_{i,t} \tag{2.3}$$

$MAX_{i,t}$ is computed by taking stock i's highest daily return within each month t, following Bali et al. (2011). $MAX_{nonews,i,t}$ is the MAX stock i which is not associated with any news coverage during the past month t. $MAX_{news,i,t}$ is the month t return on MAX stock i that can be linked to at least one Dow Jones news article in the prior month. In the X vector, a battery of control variables are added to capture various stock characteristics including $SIZE$, BTM, $BETA$, MOM, REV, and $ILLIQ$. It is then expected that the sign of β_1 would be negative in equation (2.2) whereas it would be positive in equation (2.3), as suggested by my initial hypothesis.

In Table 2.5, findings from estimating the three equations are presented. First, I cross validate the MAX effect in equation 1. The coefficient on MAX is negative as expected, and is significant at the 1% level (t-statistic: -5.51). This again confirms the finding of Bali et al. (2011) where stocks with high maximum daily returns in the prior month are more likely to experience negative returns in the next month. Next, I test the MAX_{nonews} variable, which comprises those MAX stocks not associated with news coverage, using equation (2.2). It is evident that the magnitude of the MAX_{nonews} coefficient is much larger than the corresponding MAX coefficient in equation (2.1), which is consistent with the results from previous sections. Astute readers may have noticed that the beta effect is insignificant from zero in the regression analysis. Given the finding of Bali, Brown, Murray, and Tang (2017b) that investors' demand for lottery-like stocks is an important driver of the beta anomaly, it is reasonable to believe that the MAX effect may be replacing the beta effect in this case. Finally, I include the MAX_{news} term in equation (2.3). As a result, the positive coefficient of MAX_{news} (0.0155 with a t-statistic of 1.99) confirms the return momentum pattern as hypothesized.

Overall, the results from the Fama-MacBeth regressions confirm the different directions of stock return predictability for MAX_{nonews} and MAX_{news} stocks.

Insert Table 2.5 here

2.4.5 Robustness Checks

I examine the robustness of the decile-sort portfolio results in Table 2.6. The four panels refer to four alternative measures. There are six columns corresponding to MAX_{nonews}, MAX_{news} and MAX_{nonews}-MAX_{news} on both equally- and value-weighted schemes, respectively.

I attempt to address to what extent MAX_{news} is associated with less informative DJNS news articles. The potential concern is that some MAX_{news} might be constructed from a problematic linkage between MAX and the co-occurrence of less informative or even irrelevant news articles. Without a clear economic path from news to MAX, it is hard to attribute MAX_{news} to information-related MAX stocks. But arguably, I can reject this notion for three reasons. First, if this is a major issue within the news database, I would not observe a significant difference between MAX_{nonews} and MAX_{news} from the empirical results. Second, news articles associated with MAX days are indeed firm-specific. In unreported tests, I find similar results if I employ an alternative news database where stock-level news is collected only upon a firm's name appearing in the first 25 words including headlines and the leading paragraph. Third, I continue to narrow down the target news topics to only six major corporate events including earnings reports, M&A, insider buying and selling, analyst and credit rating regrading, dividend news, and CEO announcements. The linking window from MAX to news employs data from only 9:30 am to 16:00 pm on day t in order to drop overnight news. It is then believed that these MAX_{news} returns are purely informative. In the first panel, it can be noticed that results remain similar.

It is natural to question whether the MAX_{news} effect I document is merely replicating the size effect. Indeed, stocks with news coverage are much bigger (in the range of \$198 million (log market cap 5.29) to \$2186 million (log market cap 7.69)) than those without news coverage in the range of \$44 million (log market cap

3.79) to \$226 million (log market cap 5.00)). This evidence is consistent with the previous literature that news coverage tends to focus on stocks with larger market capitalization (see e.g., Tetlock et al. (2008)). On the other hand, the evidence from the chapter clearly shows that the lottery demand effect is significantly stronger for stocks without news coverage, similar to the argument raised by Bali et al. (2011) that the lottery demand effect is more pronounced among small stocks. But arguably, it is unlikely that the MAX_{news} effect is purely reflecting the size effect for three reasons. First, I show that the t-statistic of the MAX_{news} portfolio remains statistically significant after controlling for firm size in the Fama-Macbeth regression. Second, I perform a 10-by-10 bivariate portfolio analysis between MAX_{nonews}, MAX_{news}, and size. The aim of this exercise is to examine whether the cross-sectional return pattern I documented is actually driven by firm size. By so doing, one can expect that each MAX-decile portfolio will contain stocks with evenly distributed firm size. It can be seen that the return spread between MAX_{nonews} and MAX_{news} remains statistically significant. Finally, I attempt to construct a measure of abnormal news arrival and link it to the MAX day. The rationale behind this exercise is the following: large firms tend to receive more news coverage every day whereas smaller firms have sparse news observations. By benchmarking against a firm's historical news arrival rate (30-day rolling average and 90-day rolling average), I reduce the proportion of firms with a larger market capitalization and a higher level of news coverage in the sample. Eventually, I find the results are quantitatively similar. Overall, it seems that the MAX_{nonews} (MAX_{news}) effect is not just a replication of the size effect.

I now move to discuss several other robustness checks. When linking MAX stocks to news coverage, I have thus far used a [t-1,t+1] window to allow for potential information leakage or journalist delayed reactions. The second panel of Table 2.6 shows that, if I adapt the time interval to [t-2,t+2] or [t-3,t+3], the results are similar, albeit being less significant. This finding is sensible as a larger window will capture a large volume of noisy news, making the result relatively weaker.

Empirical asset pricing often skips a month between the formation and estimation periods to allow investors to re-balance their portfolios. I therefore add a one-month gap in between to make the portfolio construction more practical. The third panel of Table 2.6 reports similar results to the previous one.

In addition, I employ alternative MAX measures following Bali et al. (2011), namely MAX2, MAX3, MAX4 and MAX5. These lottery-like proxies are calculated by taking the average of 2, 3, 4 or 5 maximum daily returns in the prior month, respectively. As for the MAX_{news} effect presented above, I re-identify it as long as at least one news item can be linked to the maximum daily return in the specified window. In the fourth panel of Table 2.6, I confirm that the MAX_{nonews}-MAX_{news} difference portfolio return consistently exists across each MAX measure.

Next, I cross-validate the findings by applying an alternative lottery-like proxy. The essential question here is whether MAX_{nonews} stocks still have stronger lottery demands conditional on the other type of lottery characteristics. If news coverage indeed erases the lottery-feature of stocks on each MAX day, the MAX_{nonews} portfolio should be immune to alternative lottery characteristics settings. To examine this, I introduce two lottery-like proxies. One follows the work by Kumar (2009b), where lottery-type stocks are defined as those with a lower stock price (PRC), higher idiosyncratic volatility (IVOL), and higher idiosyncratic skewness (ISKEW). To construct this proxy, I independently rank all stocks into 50 bins based on IVOL and ISKEW in an ascending order, and sort all stocks based on PRC in a descending order, respectively. The total ranking score (donated as LTRY) is then calculated by summing the ranking indices of the IVOL, ISKEW and PRC portfolios. By doing so, the highest ranking bin contains lower PRC, higher IVOL and ISKEW stocks and they are therefore likely to be treated as lottery-type stocks. In the final panel of Table 2.6, it can be noticed that the spread between MAX_{nonews}-MAX_{news} is still statistically significant. My second lottery-type proxy is the O-score which measures the expected probability of bankruptcy from Ohlson (1980).[7] A higher O-score indicates that a firm is expected to have a higher probability of bankruptcy. Recent work by Barinov (2018) suggests that the MAX effect will be stronger for firms with a high O-score. Inspired by this finding, I re-examine whether the lottery demand of MAX_{nonews} is also stronger conditional upon high O-score stocks. Specifically, in each month, all stocks are sorted into terciles based on the most recent quarterly O-score. The top tercile therefore contains high O-score stocks which are likely to have stronger lottery features. I then construct MAX_{nonews} and MAX_{news} within the tercile and report return performance, respectively. In a line with the baseline result, the MAX_{nonews}-MAX_{news} difference portfolio return still exists in this lottery-like characteristic.

It is also worth considering alternative models to evaluate portfolios' abnormal returns. To do so, I utilise the Fama-French five risk factors from Kenneth French's Data library.[8] and retrieve the Hou-Xue-Zhange four risk factors from the Q-factors Data library.[9] I obtain both sets of alphas by regressing the excess returns of MAX_{news} and MAX_{nonews} on the Fama-French five-factor model (FF5F) and Hou-Xue-Zhang four-factor model (HXZ4F), respectively. The results are included in Table 2.6 of the robustness checks subsection. Overall, the abnormal return spread between MAX_{nonews} and MAX_{news} from FF5F and HXZ4F is quantitatively similar to those from FFC4F and FFC4F plus liquidity factor.[10]

Collectively, all robustness checks support the notion that the MAX_{news} port-

[7] I thank the referee for suggesting this lottery-like proxy for the test.

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

[9] `http://theinvestmentcapm.com/research.html.`

[10] The FF5F model includes market excess return, firm size, value premium, operating profitability, and investment. The HXZ4F model includes market excess return, firm size, investment, and ROE. They are different from FFC4 and FFC4+PS models where momentum and liquidity factors are included.

folio has fewer lottery-like features compared to the MAX_{nonews} portfolio.

——————————— Insert Table 2.6 here ———————————

2.4.6 Is the MAX Effect Caused by Seasonality?

A potential concern is that MAX_{news}'s return momentum is due to seasonality. This may be the case for several reasons. First, most firms tend to release corporate filings such as 10-Ks and 10-Qs in certain months, and the MAX_{news} effect might be driven by the month where a large number of companies update their firm-specific information. Second, seasonal anomalies such as the "January effect" may be attributed to return momentum since previous studies have observed large abnormal stock returns in January.

To address this issue, I replicate my baseline model adjusting for seasonality. Specifically, in each month, all stocks are sorted into deciles based on the MAX calculated in the prior month. The post-formation period return for each decile is traced. I drop one month of observations from January to December, respectively, from the sample before the formation of the portfolio. For example, when the month of January is dropped, I report the time-series average of the post-formation period return for MAX_{nonews} and MAX_{news} for the remaining months. Due to space constraints, only the hedged portfolio return differences between the highest and lowest deciles are tabulated in Table 2.7 for each case.

In Panel A of Table 2.7, it is evident that the augmented MAX effect, MAX_{nonews}, is consistently present which ever month is dropped from the sample. Looking at Panel B, the hedge return spread of MAX_{news} is also statistically significant across each case on an equally-weighted basis. These results suggest that the effect I find is unlikely to be driven by a single month. Moreover, I do observe that both MAX_{nonews} and MAX_{news} earn the lowest next-month return when excluding January. For example, the equally-weighted return of MAX_{news} is only 1.07% per month, compared to the performance of other portfolios ranging from 1.22% to 1.53%, indicating that the "January effect" partially inflates MAX_{news} stock return predictability. However, it appears to be implausible that the entire return momentum is attributable to this seasonal anomaly.

——————————— Insert Table 2.7 here ———————————

2.4.7 Do Only Earnings Announcements Matter?

Prior literature has documented that the MAX effect disappears if the maximum daily return is driven by earnings announcement events – what I term $MAX_{earnings}$ – (Nguyen and Truong, 2018). Thus, it is worth studying to what extent my new measure, MAX_{news}, overlaps with $MAX_{earnings}$. In this subsection, I try to deal with this concern: is the MAX_{news} finding largely driven by earnings announcement events? There are several reasons why I expect this not to be the case.

First, the business press utilised in this chapter is substantially different from earnings announcements events that are commonly applied in the literature and in Nguyen and Truong (2018) in particular. As discussed in the literature review section, the business press (e.g., the Dow Jones Newswire Archive that I use in this study) has several different roles, including information dissemination and information creation, which can impact the capital market through different mechanisms. With respect to its information dissemination role, the business press can cover scheduled corporate events such as earnings announcements and circulate them into the Newswire. In this case, it is fair to say that the news coverage is a proxy for different corporate news events and the firm's quarterly earnings conference accounts for a small subset. Moreover, the news coverage has an information creation role through which journalists can produce comprehensive content and deep insights.

These full articles containing information from multiple sources such as analysts, managers, etc., may offer additional content beyond the corporate filings released by firms. From this perspective, it is fair to conclude that the DJNS news coverage provides investors with a high-dimensional information flow compared with earnings announcement events alone. As the statistics show, 81% of MAX_{news} is not associated with the firm's quarterly earnings events, indicating that EA events do not capture all the effects identified by the DJNS news reports.[11]

Next, I look into each MAX decile and examine the proportion of MAX_{news} and $MAX_{earnings}$ observations. For the highest MAX group (decile 10 specifically), 14.5% of maximum daily returns are linked to EA events whereas 33.5% of MAXs are associated with any type of news coverage. The gap between the two types not only exists within the highest decile but can also be consistently observed among others. Collectively, the percentage of MAX returns linked with MAX_{news} is overwhelming when compared to the $MAX_{earnings}$. Splitting the sample range into pre- (before 1995) and post-internet (1995 onwards) periods, I perform a subsample analysis. During the pre-internet period, it is interesting to see that the percentage of MAX_{news} events is smaller than those of $MAX_{earnings}$ for the top four deciles, while the remaining groups report similar statistics. For example, 6.2% of MAXs can be linked to DJNS news articles, whereas EA events associated MAXs account for 7.1% in the decile 10 group. Given that the Dow Jones Newswire Archive is incomplete and that the internet was underdeveloped before 1995, the pre-internet results are not surprising. In contrast, the number of MAX_{news} events is significantly larger than those of $MAX_{earnings}$, as the former accounts for more than 50% compared to only 20% for the latter in the highest MAX decile. This again confirms the huge difference between the MAX_{news} and the $MAX_{earnings}$.

[11] I link each MAX with earnings announcement events in the [t-1,t+1] windows in the same fashion as the exercise designed for constructing the MAX_{news}.

——————————— Insert Table 2.8 here ———————————

In addition, I re-run the baseline test but drop all MAXs associated with EA events from the MAX_{news} to the MAX_{nonews}. If EA events drive the entire return momentum of the MAX_{news}, it is then expected that the difference in holding period returns between the two types will become statistically insignificant. By contrast, the difference will remain statistically significant if other (non-earnings) types of news also contribute to this phenomenon. In the results from Table 2.9, Panel A presents time-series average future returns across each decile for the whole sample period. Moreover, to address the concern that subsequent earnings-related news coverage could also provoke the MAXs as EA events, I exclude both EA events and DJNS tagged EA news from MAX_{news}.[12] As a result, these two tests report similar outputs as the statistical difference remains significant at 1% for both equally- and value-weighting schemes in the MAX_{nonews} – MAX_{news} column in Panel C. This indicates that the anomaly documented by Nguyen and Truong (2018) not only exists for earnings announcements events but can also be broadly observed when the MAX is associated with the business press. Although I do not have direct evidence on the extent to which the business press influences the pricing of MAX_{news} stocks, their performance could be a result of those articles that reprint corporate filings released by the firm, or a result of full articles that contain additional content and deep insights, or both. As such, it provides a promising avenue for future study such as performing textual analysis to disaggregate reprinted corporate filings and in-depth coverage.

Lastly, in an unreported test, I also perform a horse-race test between MAX_{news} and $MAX_{earnings}$; and between MAX_{nonews} and $MAX_{noearnings}$, respectively. As my approach better disentangles the two anomalies from the MAX, one can expect that

[12]The Dow Jones Newswire Archive provides subject codes for each news item and therefore the topic can be easily identified.

MAX_{news} and MAX_{nonews} will contain less noise and should predict more distinct and sharp future return performance than those of $MAX_{earnings}$ and $MAX_{noearnings}$. When aligning the two samples on the same time period for direct comparison with the results of Nguyen and Truong (2018) – i.e., August 1979 to December 2016, MAX_{nonews} reports an average of –1.74% per month compared with –1.30% for $MAX_{noearnings}$; MAX_{news} yields 1.34%, which is significant at the 1% level, whereas $MAX_{earnings}$ only gives -0.36 %, which is insignificant. I therefore conclude that the important distinction is not between MAXs associated with earnings (or not), but rather between MAXs associated with news more widely (or not).

Overall, it can be concluded that the MAX_{news} effect that I find in this chapter is not purely driven by the $MAX_{earnings}$ documented by the previous literature.

Insert Table 2.9 here

2.4.8 Investor Attention?

Next, I explore the possibility that investor attention drives MAX_{news}. The previous literature documents that investor attention is a meaningful cross-sectional stock return predictor (e.g., Hillert et al. (2014), Kaniel and Parham (2017)). The economic channel between investor attention and higher future returns is that a high volume of media coverage tends to generate high visibility. Once attention has increased, investors, especially retail investors, tend to buy such attention-grabbing stocks.

Arguably, these findings do not fit to the lottery-type stock setting. Bali et al. (2019) document that lottery stock overvaluation is heavily driven by retail investor attention. In other words, a high level of investor attention leads to lower future returns for lottery-type stocks. Given the positive future returns of the MAX_{news} portfolios in the empirical setting, it is fair to conclude that investor attention cannot explain the result.

To further rule out investor attention theory, I construct three investor attention proxies. The first variable is the natural logarithm of the total volume of DJNS news reports in the formation month, donated as $NEWSCVG$. Second, I apply analyst forecasts for one-year-ahead earnings for each firm, $ASTCVG$, which is computed by taking the natural logarithm of the total volume of analyst coverage. Lastly, I utilise the absolute value of the most recent standardized unexpected earnings (SUE) as a proxy for attention, denoted as $ABSSUE$, where the SUE is calculated following Bernard and Thomas (1989). In the Fama-MacBeth regression, I first include these three controls individually and then incorporate them together; the details of this are reported in Table 2.10. As can be seen, MAX_{news} remains consistently statistically significant under each specification, albeit exhibiting relatively lower t-statistics.

Overall, it can be concluded that investor attention theory does not explain differing future returns of MAX_{news}.

——————————— Insert Table 2.10 here ———————————

2.4.9 Information Uncertainty Mitigation?

Motivated by the role of news in resolving information uncertainty (Tetlock, 2010), I attempt to explain the return momentum of MAX_{news} in this subsection. I hypothesize that MAX_{news} stocks eliminate their information uncertainty since associated news coverage discloses the underlying events and and allows public scrutiny. When investors read a news article, they immediately update their estimates of the actual fundamentals of the stock and therefore change their trading behaviour accordingly. Alternatively, if there is no news report accompanying a MAX, investors can only infer the cause and speculate, categorizing the stock as being of the lottery-type.

To test my hypothesis, I first examine the time-varying information uncertainty around MAX days. Stock return volatility (denoted as RETVOL) is utilised as a proxy. If the volatility substantially decreases from the pre-MAX_{news} period to the post-MAX_{news} time as compared to those of MAX_{nonews} days, the evidence may be supportive to the notion that news resolves uncertainty. Specifically, I calculate the RETVOL using daily return data before and after the MAX day for both MAX_{nonews} and MAX_{news} stocks. For example, when the window is considering 5 days before and 5 days following, $RETVOL_{t-5,t-1}$ and $RETVOL_{t+1,t+5}$ are computed. I then take the natural logarithm of $RETVOL_{t-5,t-1}$ against $RETVOL_{t+1,t+5}$. A negative value implies that information uncertainty is reduced. From Panel A of Table 2.11, it can be observed that the ratios of MAX_{news} are consistently larger than those from the MAX_{nonews} in three different examining windows (i.e. 5-day, 10-day, and 15-day).

Next, I manage to provide further evidence by performing a bivariate portfolio study. If news indeed resolves uncertainty after the MAX day, I should expect that the return difference between decile 10 and 1 of MAX_{news} stocks to be at least non-negative for a high information uncertainty group. Moreover, the future return of MAX_{nonews}-MAX_{news} is also expected to be larger under high information uncertainty. Following Zhang (2006) and Kumar (2009a), I employ four different information uncertainty proxies including idiosyncratic volatility (IVOL), earnings volatility (EVOL), cash flow volatility (CFVOL), and analyst forecast dispersion (DISP). [13] IVOL is obtained by calculating the standard deviation of residuals from a rolling monthly OLS regression between daily stock returns and the market excess return (MKT), small-minus-big (SMB), high-minus-low (HML), and up-minus-down (UMD) risk factors. EVOL is measured based on the standard deviation of earnings

[13]These proxies have been employed in the study of uncertainty resolution and shown to be effective (e.g. Patton and Verardo (2012)).

from the past 20 quarters.[14] CFVOL is defined as the standard deviation of cash flows from operations in the past 20 quarters.[15] DISP is the logarithm of the standard deviation of analyst forecasts obtained from Thomson Reuters Datastream for one-year-ahead earnings per share scaled by the prior year-end price.

The test proceeds as follows: in each month, all stocks are sorted into low, middle and high information uncertainty groups based on the uncertainty proxies. Within each group, I further construct MAX_{nonews} and MAX_{news} stocks using the same methods. The holding period returns are collected and tabulated by calculating time-series averages within the whole sample. For brevity, I only report the top and bottom information uncertainty groups; columns for each of MAX_{nonews}, MAX_{news} and MAX_{nonews}-MAX_{news} are presented in Table 2.11 for the top and bottom terciles, respectively.

In the Panel B of Table 2.11, I report the time-series averages of the hedged return differences for MAX_{nonews}, MAX_{news} and MAX_{nonews}-MAX_{news} within the lowest and highest information uncertainty groups. First, it is evident that all MAX_{news} stocks report insignificant future returns among high information uncertainty groups. This indicates that these stocks do not have lottery-features due to the DJNS news coverage. When no news can be linked to the MAX stocks, they continue to exhibit strong lottery-features (i.e., negative future returns). Moving on to compare the return spread between MAX_{nonews} and MAX_{news}, it is interesting to see that the difference is negative and statistically significant among high IVOL stocks, whereby for the low uncertainty group this is not the case. This result can be interpreted as follows: stocks with high information uncertainty are most likely to be influenced upon the arrival of news. When investors are informed by these stock-level news articles, the fundamentals become easy-to-value. Thus, the differ-

[14]The earnings measure is the diluted earnings per share (Compustat quarterly data item 9).

[15]The cash flow per share is calculated as: (Compustat quarterly data item 108)/(Compustat quarterly data item 61 * Compustat quarterly data item 17).

ence between the two types becomes substantially large. Similarly, the EVOL also confirms my hypothesis. For example, the MAX_{nonews}-MAX_{news} within high EVOL earns -0.78% in next month, while this figure for the low EVOL grouping is only -0.25% and insignificant from zero. As for the cash flow volatility, the MAX_{nonews}-MAX_{news} is -0.68% for low CFVOL but earns -1.02% when the CFVOL is high on an equally-weighted basis. The empirical evidence, however, provides mixed results where DISP is not in line with the initial hypothesis. Presumably, a possible explanation is that the total number of observations substantially shrank after conditioning on DISP. As Kumar (2009a, p.1376) argues in his paper: "a large number of firms with high uncertainty do not have analyst coverage and must be excluded from the analysis".

——————————— Insert Table 2.11 here ———————————

Arguably, given analyst forecast dispersion, it is fair to say that news coverage could temporarily increase the level of investor disagreement and lead to higher information uncertainty. For example, Bali et al. (2017a) document that unusual firm-level news flow sparks a divergence of opinions. By using an advanced textual analysis to identify relevant public information, Boudoukh et al. (2018) find that a news story is a meaningful component of stock return variance. As a result, the stock return predictability of MAX_{news} could result from mixed empirical evidence. In the next subsection, I continue to study the channel between lottery-like investing and news coverage.

2.4.10 MIN_{news} and VaR_{news}

A natural way to examine whether information uncertainty mitigation theory or divergence of opinions theory explains the MAX_{news}'s stock return predictability is to look at the other side of the stock return distribution – the minimum daily return (denoted as MIN) in the prior month. If the arrival of news flow induces a divergence of opinions, it would further increase stock idiosyncratic volatility. As a result, stocks' future returns should be higher to compensate for this incremental risk in the following month. Similarly, the same pattern should also be observed by the MIN_{news} stocks in the sense that they will compensate investors for higher volatility too. On the other hand, I would expect that MAX_{news} and MIN_{news} predict very different subsequent investor trading behaviours if news coverage mitigates information uncertainty and incorporates its embedded information.

To examine the MIN and its stock return predictability, I adapt my baseline research setting. Specifically, in each month, all stocks are ranked into deciles based on the MIN calculated in the past one-month window.[16] The future return performance of each decile is traced. To construct MIN_{news}, I follow the previous empirical settings in this chapter where the linkage between minimum daily returns and DJNS news reports is built within a [t-1,t+1] time interval. Finally, I calculate the time-series average of MIN_{news} holding period returns in my sample.

As can be seen in Table 2.12, both equally- and value-weighted MIN_{news} predict negative future returns. For instance, the spread of High-Low is -1.78% per month (t-statistic: -3.79) for equally-weighted stocks, and the value-weighted spread is -1.66% (t-statistic: -3.23). When the stock return predictability of MAX_{news} is positive, the opposite direction between the two types suggests that news coverage is more likely to resolve information uncertainty instead of increasing it.

——————————— Insert Table 2.12 here ———————————

In addition, I study the left-tail of the stock return distribution with different measures as a further robustness check. Specifically, I try to replicate the work of Atilgan, Bali, Demirtas, and Gunaydin (2020). They document left-tail momentum

[16]The MIN is in absolute terms. Therefore, a high MIN means extremely low stock returns.

in a cross-sectional setting by utilizing Value-at-Risk (VaR) for each stock. The empirical evidence points out that high (low) VaR stocks predict lower (higher) future returns. I again adapt their measure and construct VaR_{news} portfolios as an additional proxy for lower stock returns. Specifically, the VaR is calculated by rolling the past one-year daily returns (250 trading days) with at least 200 non-missing observations. Due to limited space within this chapter, I only compute the first percentile of the daily returns (denoted as VaR1). Using the method from my previous settings to construct VaR_{news}, I report the time-series average of the return spread between the highest and lowest VaR group and risk-adjusted returns including the Fama-French-Carhart four-factor alpha and liquidity-augmented Fama-French-Carhart five-factor alpha.

The empirical results show that VaR_{news} is still aligned with MIN_{news}, though the magnitude of future returns significantly reduces. This is plausible because return momentum would be expected to decline gradually over a longer time horizon. Overall, the empirical evidence in this subsection supports the theory that news coverage mitigates information uncertainty.

2.4.11 Investor Sentiment

The previous literature has identified that the MAX effect largely concentrates during high investor sentiment periods using the Baker and Wurgler (2006) sentiment index (Fong and Toh, 2014). In this subsection, I explore how investor sentiment plays a role in news and no-news MAX stocks.

I employ six sets of investor sentiment. First, the Baker and Wurgler (BW) sentiment index is included. BW is a market-aggregated sentiment index based on six different sources, namely: the closed-end fund discount, NYSE share turnover, the number and average of first-day returns on IPOs, the equity share in new issues, and the dividend premium. Second is the FEARS (Financial and Economic Attitudes Revealed by Search) index constructed by Da et al. (2014a), who aggregate a set of high Google search volume financial and economic keywords by households as a measure of retail investor interest. Third, I add an MS (Manager Sentiment) index from Jiang et al. (2019), which is built by aggregating firm-level manager sentiment proxied by 10-Ks and 10-Qs. Fourth is MCSI which names the Michigan Consumer Sentiment Index compiled by the University of Michigan Survey Research Center. Next, the CBC (the Conference Board Consumer Confidence Index) is included. Finally, I add HJTZ, an optimal BW sentiment index which has reduced noise compared with the original one by Huang et al. (2015).[17]

To examine the performances of MAX stocks during high/low investor sentiment periods, I first partition each index into high and low periods based on when it is above/below its sample mean. I then trace the holding period return for MAX_{nonews} and MAX_{news} in each month and report the time-series average coefficients. To summarize, the detailed performance of each decile is excluded and I only report the hedged portfolio return spread between the highest and lowest deciles for each of the six investor sentiment indexes on both an equally- and a value-weighted basis.

In Table 2.13, I report all the time-series average return spreads of the MAX_{nonews} and MAX_{news} portfolios conditioning on different investor sentiment. Noticeably, the hedged return spread between the highest and lowest MAX_{nonews} is consistently stronger within high sentiment periods. For example, the BW reports -2.04% per month (t-statistic equals -6.58) under the equally-weighted scheme when sentiment is high, whereby it only achieves -1.08% from a low sentiment period, with a t-statistic of -3.79. I conclude that this finding is robust across different sentiment

[17] The Baker and Wurgler sentiment index can be downloaded from their website, `http://people.stern.nyu.edu/jwurgler/`; the FEARS sentiment index is available from Zhi Da's website, `https://www3.nd.edu/~zda/`; the MS and HJTZ sentiment index can be found from Guofu Zhou's website, `http://apps.olin.wustl.edu/faculty/zhou/`; the MCSI is downloaded from `http://www.sca.isr.umich.edu/`; The CBC is obtained from Bloomberg after adjusting seasonality.

proxies and sample ranges. For instance, due to data constraints, the FEARS index is only available from July 2004 to December 2011. As a result, the empirical finding supports the theory that high sentiment induces stronger MAX effects, which is consistent with previous studies.

In contrast, the predictive power of MAX_{news} is stronger when sentiment is low. Five out of six sentiment measures report statistical significance during the low sentiment periods. Looking at the BW index, the average return in the next month is 1.80% (1.50%) on an equally-weighted (value-weighted) basis within the low periods, while those statistics are only 0.85% (-0.37%) respectively for the high sentiment periods. Presumably, it is likely that investors are more desperate to demand good news when sentiment is at a relative low level, which is consistent with the "flight-to-quality" theory (Bethke et al., 2017). As a result, a persistent MAX_{news} can be observed.

To conclude, I confirm the theory proposed by other studies where the MAX effect is stronger when market-aggregated sentiment is high. Yet MAX_{news} stocks witness the opposite: they mainly lead to return momentum during low sentiment periods.

———————— Insert Table 2.13 here ————————

2.5 Summary

In this chapter, I first consider an empirical setting in which lottery-like stocks, proxied by their maximum daily returns (MAX) over a month, are combined with news reports. Using a comprehensive U.S. stock-level Dow Jones news data set between 1979 and 2016, I distinguish MAX stocks according to whether they can be linked to DJNS' news coverage. MAX stocks associated with news coverage (denoted MAX_{news}) exhibit a monotonically increasing future return relationship from the lowest MAX portfolio to the highest, indicating a return momentum pattern. While the augmented MAX effect, which I term MAX_{nonews} (MAX stocks without news coverage), is more than double that compared to the original noted effect.

I argue that the opposite direction of the MAX_{nonews} and the MAX_{news} stocks' expected returns is likely to be explained by information uncertainty mitigation theory, where the arrival of news items resolves the stock's future uncertainty, and therefore deters investors from subsequent gambling behavior. A time-varying comparison between the pre-MAX and post-MAX period clearly shows that stock return volatility substantially decreases when MAX days are accompanied by news reports. Further empirical evidence also rejects the alternative theory where MAX-associated news articles induce a high level of investor disagreement and divergence of opinions, as MAX_{news} and MIN_{news} exhibit opposite future return directions.

In addition, I also explore how investor sentiment plays a role in lottery-type investments. Consistent with previous studies, the augmented MAX effect contributed by the MAX_{nonews} is persistent when investor sentiment is high. In contrast, the return momentum of the MAX_{news} largely concentrates during low sentiment periods, suggesting that investors are more likely to undertake a "flight-to-quality" when sentiment is bad.

To conclude, this chapter refines the role of news coverage in resolving information uncertainty among lottery-type stocks. When a MAX is covered by news reports under the spotlight, a MAX is no longer the MAX.

Table 2.1: Summary Statistics

Table 2.1 reports the summary statistics of each variable used in this chapter. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). MEDIA is the total volume of the MAX associated DJNS news reports. OPT is the net value of the positive word percentage minus the negative on a daily basis calculated by the Loughran and McDonald (2011) dictionary method respectively, and is normalized cross-sectionally. SIZE is computed by taking the logarithm of prices multiplied by the number of shares outstanding in a prior month. BTM is calculated by taking logarithm of book value over market capitalization updated at the every end of June. TURN is calculated by taking the logarithm of average trading volume over the past month. REV is the one-month stock return on a prior month. MOM is the 12-month cumulative stock returns after skipping the one-month recent stock return. PRC is the stock price in a prior month. IVOL and ISKEW are the standard deviation and the skewness of the residual by regressing a stock one-month return against corresponding Fama-French three factors. BETA is calculated following Scholes and Williams (1977) and Dimson (1979). ILLIQ is computed following Amihud (2002). The sample runs from August 1979 to December 2016.

	MAX	**SIZE**	**MEDIA**	**OPT**	**REV**	**TURN**	**MOM**	**BTM**	**IVOL**	**ISKEW**	**PRC**	**BETA**	**ILLIQ**	**SUE**
Mean	6.77	6.70	2.71	0.01	3.22	0.17	22.14	-0.80	0.02	0.28	32.37	1.09	0.09	-0.01
Stdev	8.27	2.24	3.57	0.85	17.70	1.09	78.90	0.88	0.02	0.99	42.61	1.87	2.41	0.63
Min	0.00	-3.70	1.00	-11.79	-87.84	-8.52	-99.91	-8.86	0.00	-4.52	5.00	-154.04	0.00	-2.47
0.25	3.02	5.42	1.00	-0.25	-4.49	-0.45	-9.41	-1.28	0.01	-0.27	13.00	0.25	0.00	0.00
0.50	4.80	6.76	2.00	0.00	1.73	0.25	9.95	-0.72	0.02	0.22	23.74	0.98	0.00	0.00
0.75	8.00	8.13	3.00	0.36	8.79	0.89	36.18	-0.23	0.03	0.77	40.30	1.80	0.01	0.00
Max	1277.78	13.57	585.00	13.43	1598.45	7.15	9857.14	4.55	2.80	4.74	3430.00	178.30	568.19	2.47

Table 2.2: Univariate Portfolio Analysis

Table 2.2 reports the univariate portfolio analysis of MAX_{news} and MAX_{nonews} stocks. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). I identify whether the MAX is accompanied by DJNS news reports as long as they lie within a [t-1,t+1] 3-day window. MEDIA is the total volume of the MAX associated DJNS news reports. OPT is the net value of the positive word percentage minus the negative on a daily basis calculated by the Loughran and McDonald (2011) dictionary method respectively, and is normalized cross-sectionally. SIZE is computed by taking the logarithm of prices multiplied by the number of shares outstanding in a prior month. BTM is calculated by taking logarithm of book value over market capitalization updated at the every end of June. TURN is calculated by taking the logarithm of average trading volume over the past month. REV is the one-month stock return on a prior month. MOM is the 12-month cumulative stock returns after skipping the one-month recent stock return. PRC is the stock price in a prior month. IVOL and ISKEW are the standard deviation and the skewness of the residual by regressing a stock one-month return against corresponding Fama-French three factors. BETA is calculated following Scholes and Williams (1977) and Dimson (1979). ILLIQ is computed following Amihud (2002). The sample runs from August 1979 to December 2016.

Panel A: Univariate Portfolio Analysis of MAX_{news}

RANK	Low	port2	port3	port4	port5	port6	port7	port8	port9	High
MAX_{news}	1.78	2.70	3.34	3.98	4.69	5.52	6.56	8.01	10.44	20.70
MEDIA	2.53	2.65	2.64	2.61	2.58	2.59	2.61	2.68	2.80	3.40
OPT	0.70%	0.77%	0.72%	0.32%	0.92%	0.90%	0.53%	0.73%	0.51%	1.17%
SIZE	7.67	7.69	7.43	7.14	6.88	6.63	6.35	6.09	5.77	5.29
BTM	-0.68	-0.75	-0.77	-0.78	-0.80	-0.82	-0.83	-0.86	-0.87	-0.83
TURN	-2.82	-2.68	-2.57	-2.47	-2.35	-2.24	-2.12	-2.01	-1.85	-1.52
REV	-1.95	-1.08	-0.55	0.00	0.61	1.40	2.40	4.14	7.05	20.14
MOM	16.85	15.39	16.73	18.40	20.82	22.49	25.34	27.89	29.17	28.30
PRC	42.97	42.25	39.17	36.19	33.26	30.95	28.21	25.81	23.55	21.30
IVOL	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.05
ISKEW	-0.21	-0.12	-0.05	0.02	0.09	0.17	0.28	0.44	0.70	1.44
BETA	0.49	0.72	0.86	0.96	1.04	1.15	1.25	1.35	1.49	1.59
ILLIQ	0.04	0.04	0.05	0.06	0.06	0.08	0.11	0.12	0.17	0.21

Panel B: Univariate Portfolio Analysis of MAX_{nonews}

RANK	Low	port2	port3	port4	port5	port6	port7	port8	port9	High
MAX_{nonews}	1.34	2.37	3.06	3.72	4.41	5.19	6.13	7.37	9.30	15.66
SIZE	5.00	5.42	5.34	5.19	5.03	4.87	4.67	4.48	4.22	3.79
BTM	-0.31	-0.39	-0.43	-0.46	-0.49	-0.53	-0.56	-0.60	-0.62	-0.63
TURN	-1.36	-1.07	-0.97	-0.87	-0.80	-0.70	-0.62	-0.53	-0.43	-0.28
REV	-2.44	-1.62	-1.03	-0.52	0.04	0.73	1.53	2.74	4.76	12.08
MOM	18.45	17.63	18.05	18.82	19.86	21.25	23.12	24.90	26.53	27.21
PRC	118.00	75.37	63.96	40.93	33.50	28.41	22.20	19.43	19.27	13.67
IVOL	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.05
ISKEW	-0.33	-0.10	-0.02	0.04	0.09	0.16	0.23	0.32	0.46	0.86
BETA	0.26	0.46	0.56	0.65	0.72	0.81	0.89	0.99	1.08	1.22
ILLIQ	0.06	0.08	0.11	0.13	0.15	0.19	0.21	0.25	0.29	0.51

49

Table 2.3: MAX_{nonews} and MAX_{news} Lottery Demand Investment

Table 2.3 Panel A reports average monthly returns for MAX stocks, MAX stocks without DJNS news reports and MAX stocks associated with DJNS news reports on both an equally-weighted and a value-weighted basis. Panel B reports the raw return, Fama-French-Carhart four-factor alpha returns and liquidity-augmented Fama-French-Carhart five-factor alpha returns between the highest and lowest MAX stocks. The liquidity factor is the Pástor and Stambaugh (PS) liquidity factor. MAX_{nonews}-MAX_{news} is the difference between the returns of the MAX_{nonews} and MAX_{news} portfolios. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). I identify the MAX as being accompanied by DJNS news reports as long as the coverage lies within a [t-1,t+1] window. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighting in **bold** indicates a coefficient that is significant at the 10%, 5% or 1% level, respectively.

Panel A		Low	port2	port3	port4	port5	port6	port7	port8	port9	High
MAX	EW	**1.22%**	**1.39%**	**1.45%**	**1.39%**	**1.37%**	**1.29%**	**1.14%**	**1.03%**	**0.76%**	0.25%
	t-stat	(8.15)	(7.20)	(6.92)	(6.17)	(5.69)	(4.98)	(4.11)	(3.43)	(2.26)	(0.70)
	VW	**1.20%**	**1.10%**	**1.20%**	**1.01%**	**1.25%**	**1.11%**	**1.11%**	**0.94%**	**0.86%**	0.42%
	t-stat	(7.26)	(5.97)	(5.84)	(4.56)	(5.17)	(4.28)	(3.83)	(3.01)	(2.43)	(1.10)
MAX_{nonews}	EW	**0.93%**	**1.06%**	**1.03%**	**0.91%**	**0.89%**	**0.73%**	**0.54%**	0.27%	0.03%	**-0.81%**
	t-stat	(6.71)	(5.77)	(5.15)	(4.23)	(3.84)	(3.04)	(2.10)	(0.98)	(0.09)	(-2.53)
	VW	**0.91%**	**0.83%**	**0.81%**	**0.66%**	**0.61%**	**0.58%**	0.39%	0.04%	-0.01%	**-0.66%**
	t-stat	(5.53)	(4.41)	(3.88)	(2.91)	(2.52)	(2.29)	(1.41)	(0.12)	(-0.04)	(-1.78)
MAX_{news}	EW	**1.56%**	**1.75%**	**1.93%**	**2.02%**	**1.91%**	**2.23%**	**2.11%**	**2.64%**	**2.41%**	**2.90%**
	t-stat	(7.74)	(7.34)	(7.82)	(7.29)	(6.65)	(6.87)	(5.79)	(6.78)	(5.36)	(6.05)
	VW	**1.05%**	**1.21%**	**1.34%**	**1.29%**	**1.21%**	**1.72%**	**1.20%**	**1.75%**	**1.91%**	**1.74%**
	t-stat	(5.00)	(5.33)	(5.42)	(4.60)	(4.38)	(5.28)	(3.36)	(4.31)	(4.11)	(3.40)

Panel B	MAX		MAX_{nonews}		MAX_{news}		MAX_{nonews}-MAX_{news}	
	EW	VW	EW	VW	EW	VW	EW	VW
High - Low	**-0.97%**	**-0.78%**	**-1.74%**	**-1.57%**	**1.34%**	0.70%	**-3.08%**	**-2.27%**
	(-3.48)	(-2.38)	(-7.13)	(-4.88)	(3.08)	(1.44)	(-9.19)	(-5.55)
FFC4F	**-1.14%**	**-0.87%**	**-1.99%**	**-1.69%**	**1.13%**	0.35%	**-3.12%**	**-2.05%**
	(-7.28)	(-4.06)	(-13.45)	(-7.96)	(3.03)	(0.80)	(-9.08)	(-4.82)
FFC4F + PS	**-1.17%**	**-0.94%**	**-1.97%**	**-1.68%**	**1.03%**	0.23%	**-3.00%**	**-1.91%**
	(-7.39)	(-4.37)	(-13.24)	(-7.83)	(2.76)	(0.52)	(-8.76)	(-4.51)

Table 2.4: Bivariate Portfolio Analysis

Table 2.4 reports the average monthly return difference between MAX stocks without DJNS news reports and MAX stocks associated with DJNS news reports after controlling for stock characteristics. I report the t-statistics for the hedge return difference, Fama-French-Carhart four-factor alphas and liquidity-augmented Fama-French-Carhart five-factor alpha returns between the highest and lowest MAX stocks. The liquidity factor is the Pastor-Stambaugh (PS) liquidity factor. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). I identify the MAX as being accompanied by DJNS news reports as long as the coverage lies within a [t-1,t+1] window. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighting in bold indicates a coefficient that is significant at the 10%, 5% or 1% level.

	SIZE		BTM		BETA		MOM		REV		ILLIQ	
	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW
Low	-0.13%	0.15%	-0.11%	0.04%	-0.08%	0.25%	-0.17%	0.15%	-0.13%	0.05%	0.01%	0.22%
port2	-0.05%	0.13%	0.05%	0.21%	-0.05%	-0.04%	0.05%	0.05%	-0.02%	-0.05%	0.11%	0.13%
port3	0.03%	-0.11%	-0.01%	0.02%	0.08%	0.03%	0.13%	0.27%	-0.13%	0.04%	0.03%	-0.07%
port4	0.06%	0.23%	-0.11%	-0.04%	-0.07%	0.02%	-0.06%	-0.15%	-0.03%	-0.02%	-0.03%	0.05%
port5	0.01%	-0.18%	-0.13%	-0.10%	-0.05%	0.02%	-0.03%	-0.07%	-0.09%	0.05%	-0.05%	-0.22%
port6	-0.02%	-0.04%	-0.16%	-0.12%	-0.11%	0.02%	0.04%	-0.01%	0.05%	0.26%	-0.36%	-0.03%
port7	-0.15%	-0.02%	-0.07%	0.05%	-0.10%	-0.17%	-0.21%	-0.02%	-0.11%	0.06%	-0.21%	0.00%
port8	-0.31%	-0.16%	-0.30%	-0.15%	-0.24%	-0.47%	-0.14%	-0.07%	-0.14%	-0.35%	-0.36%	-0.14%
port9	-0.48%	-0.17%	-0.26%	-0.38%	-0.29%	-0.40%	-0.28%	-0.06%	-0.47%	-0.21%	-0.35%	-0.16%
High	-0.77%	-0.45%	-0.79%	-0.38%	-0.85%	-0.51%	-0.90%	-0.97%	-0.69%	-0.59%	-0.80%	-0.49%
High - Low	**-0.64%**	**-0.60%**	**-0.68%**	**-0.43%**	**-0.77%**	**-0.76%**	**-0.73%**	**-1.12%**	**-0.56%**	**-0.64%**	**-0.81%**	**-0.71%**
t-stat	(-4.04)	(-2.82)	(-4.42)	(-1.66)	(-5.69)	(-2.97)	(-5.02)	(-4.73)	(-3.85)	(-2.56)	(-4.99)	(-3.05)
FFC4F	**-0.83%**	**-0.60%**	**-0.82%**	**-0.54%**	**-0.86%**	**-0.70%**	**-0.89%**	**-1.18%**	**-0.67%**	**-0.72%**	**-1.00%**	**-0.71%**
t-stat	(-5.09)	(-2.76)	(-5.28)	(-2.05)	(-6.22)	(-2.68)	(-6.01)	(-4.84)	(-4.51)	(-2.82)	(-6.11)	(-3.01)
FFC4F + PS	**-0.76%**	**-0.54%**	**-0.76%**	**-0.47%**	**-0.81%**	**-0.65%**	**-0.84%**	**-1.07%**	**-0.62%**	**-0.63%**	**-0.97%**	**-0.64%**
t-stat	(-4.71)	(-2.47)	(-4.92)	(-1.78)	(-5.91)	(-2.48)	(-5.70)	(-4.44)	(-4.19)	(-2.48)	(-5.85)	(-2.69)

Table 2.5: Fama-MacBeth Regression

Table 2.5 reports the Fama and MacBeth (1973) regression under different specifications. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). MAX_{news} denotes the MAX associated with DJNS news reports whereby MAX_{nonews} is not associated DJNS news reports. I identify the MAX as being accompanied by DJNS news reports as long as they lie within the [t-1,t+1] window. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, ***,**,* indicate a coefficient significant at the 1%, 5% or 10% level, respectively.

	Model 1	Model 2	Model 3
CONST	2.0705***	2.2977***	1.3258**
	(3.87)	(4.17)	(2.17)
MAX	**-0.0588*** **		
	(-5.51)		
MAX_{nonews}		**-0.0712*** **	
		(-9.57)	
MAX_{news}			**0.0155** **
			(1.99)
SIZE	-0.0367	-0.0490**	-0.0110
	(-1.44)	(-1.92)	(-0.39)
BTM	0.1393**	0.1488**	0.1783***
	(2.35)	(2.39)	(2.80)
BETA	0.0005	-0.0013	-0.0374
	(0.01)	(-0.02)	(-0.71)
MOM	0.0058***	0.0057***	0.0057***
	(4.51)	(4.43)	(4.36)
REV	-0.0188***	-0.0211***	-0.0280***
	(-4.42)	(-5.57)	(-7.08)
ILLIQ	-0.0715*	-0.0629*	-0.1092***
	(-1.91)	(-1.74)	(-3.01)

Table 2.6: Robustness Checks

Table 2.6 reports robustness checks by various alternative settings. The sample starts from August 1979 to December 2016. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). MAX_{news} denotes the MAX associated with DJNS news reports whereby MAX_{nonews} is no associated DJNS news stocks. I identify the MAX is accompanied by DJNS news reports as long as they lie between [t-1,t+1] window. MAX_{nonews}-MAX_{news} is the difference between the MAX_{nonews} and MAX_{news} portfolio. t-statistics are in parentheses, and highlighted in bold indicating a coefficient significant at the 10% level or better.

		MAX_{nonews}		MAX_{news}		MAX_{nonews}-MAX_{news}	
		EW	VW	EW	VW	EW	VW
Alternative news		**-1.17%**	**-0.94%**	**2.22%**	1.08%	**-3.45%**	**-1.93%**
		(-4.24)	(-2.77)	(2.60)	(1.23)	(-4.05)	(-2.12)
Window for linking news to MAX	[t-2.t+2]	**-1.77%**	**-1.67%**	**0.65%**	0.52%	**-2.42%**	**-2.19%**
		(-7.44)	(-5.24)	(1.68)	(1.15)	(-8.86)	(-6.14)
	[t-3,t+3]	**-1.79%**	**-1.71%**	0.42%	0.50%	**-2.20%**	**-2.20%**
		(-7.65)	(-5.34)	(1.16)	(1.20)	(-9.02)	(-6.71)
Skip a month		**-1.18%**	**-1.35%**	**0.70%**	0.36%	**-1.88%**	**-1.72%**
		(-4.77)	(-4.18)	(1.65)	(0.86)	(-6.11)	(-4.94)
Alternative MAX	MAX5	**-2.13%**	**-2.67%**	-0.20%	0.00%	**-1.92%**	**-2.67%**
		(-8.06)	(-7.29)	(-0.56)	(0.01)	(-7.28)	(-7.51)
	MAX4	**-2.19%**	**-2.52%**	0.03%	0.16%	**-2.21%**	**-2.68%**
		(-7.74)	(-6.65)	(0.07)	(0.38)	(-7.98)	(-7.28)
	MAX3	**-2.11%**	**-2.15%**	0.27%	0.42%	**-2.38%**	**-2.57%**
		(-8.26)	(-6.27)	(0.70)	(0.96)	(-9.36)	(-7.29)
	MAX2	**-2.01%**	**-2.18%**	**0.76%**	0.56%	**-2.77%**	**-2.49%**
		(-8.04)	(-8.59)	(1.93)	(1.30)	(-10.26)	(-6.98)
Alternative lottery-like proxy	High LTRY	**-2.47%**	**-2.08%**	-0.69%	-0.84%	**-1.78%**	**-1.24%**
		(-9.22)	(-5.46)	(-1.50)	(-1.57)	(-3.82)	(-2.03)
	High O-score	**-1.57%**	**-1.64%**	0.22%	0.04%	**-1.79%**	**-1.68%**
		(-4.00)	(-3.38)	(0.32)	(0.05)	(-2.44)	(-2.00)
Alternative risk-adjusted returns	FF5F	**-1.63%**	**-1.36%**	**1.77%**	**1.01%**	**-3.40%**	**-2.37%**
		(-12.34)	(-6.74)	(4.94)	(2.31)	(-9.80)	(-5.45)
	HXZ q-model	**-1.58%**	**-1.15%**	**1.89%**	**0.99%**	**-3.47%**	**-2.13%**
		(-10.34)	(-5.25)	(4.95)	(2.16)	(-9.76)	(-4.81)
Abnormal news flow	30-rollings	**-1.63%**	**-1.41%**	**1.51%**	0.72%	**-3.14%**	**-2.14%**
		(-4.07)	(-2.59)	(2.33)	(1.05)	(-6.23)	(-3.49)
	90-rollings	**-1.62%**	**-1.41%**	**1.21%**	0.45%	**-2.83%**	**-1.86%**
		(-4.17)	(-2.66)	(1.94)	(0.68)	(-5.82)	(-3.00)

Table 2.7: MAX and Seasonality

Table 2.7 reports average monthly returns for MAX stocks with/without DJNS news reports when the formation period drops certain months. For example, Jan means that all observations on January are excluded during the formation of portfolio; Feb means including all observations except February, etc. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). I identify a MAX is accompanied by DJNS news reports as long as the coverage lies within a [t-1,t+1] window. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighting in bold indicates a coefficient significant at the 10% level or better.

Panel A: MAX_{nonews}

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
EW	**-1.86%**	**-1.78%**	**-1.71%**	**-1.75%**	**-1.73%**	**-1.75%**	**-1.58%**	**-1.71%**	**-1.74%**	**-1.71%**	**-1.79%**	**-1.72%**
t-stat	(-7.28)	(-7.38)	(-6.60)	(-7.03)	(-6.75)	(-6.78)	(-6.16)	(-6.61)	(-6.73)	(-6.69)	(-7.34)	(-6.71)
VW	**-1.70%**	**-1.64%**	**-1.50%**	**-1.66%**	**-1.62%**	**-1.58%**	**-1.36%**	**-1.56%**	**-1.57%**	**-1.47%**	**-1.63%**	**-1.59%**
t-stat	(-4.95)	(-5.15)	(-4.51)	(-4.95)	(-4.80)	(-4.64)	(-4.01)	(-4.57)	(-4.61)	(-4.34)	(-4.92)	(-4.65)

Panel B: MAX_{news}

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
EW	**1.07%**	**1.45%**	**1.45%**	**1.22%**	**1.31%**	**1.47%**	**1.53%**	**1.45%**	**1.33%**	**1.30%**	**1.30%**	**1.24%**
t-stat	(2.37)	(3.26)	(3.20)	(2.80)	(2.83)	(3.22)	(3.32)	(3.13)	(2.89)	(2.83)	(2.86)	(2.73)
VW	0.39%	0.72%	0.68%	0.59%	0.66%	0.79%	**0.88%**	0.69%	0.74%	0.59%	0.80%	0.74%
t-stat	(0.81)	(1.48)	(1.35)	(1.21)	(1.27)	(1.54)	(1.73)	(1.34)	(1.46)	(1.17)	(1.56)	(1.45)

Table 2.8: Percentage of MAX_{news} and $MAX_{earnings}$ across MAX Portfolios

Table 2.8 reports the number and percentage of MAX stock-months associated with DJNS news coverage and Earnings Announcements events across each MAX portfolio. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). DJNS denotes the MAX associated with DJNS news reports. EA is the MAX associated with Earnings Announcements events. I identify a MAX accompanied by DJNS news reports or Earnings Announcements events as long as it lies within a [t-1,t+1] window. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighting in bold indicates a coefficient significant at the 10% level or better.

	Panel A: 1979-08 to 2016-12					Panel B: 1979-08 to 1995-12					Panel C: 1996-01 to 2016-12				
	Nobs	DJNS	Pct	EA	Pct	Nobs	DJNS	Pct	EA	Pct	Nobs	DJNS	Pct	EA	Pct
Low MAX	165,728	36,582	22.1	6,343	3.8	74,454	3,827	5.1	2,276	3.1	91,790	32,834	35.8	4,082	4.4
port 2	169,609	41,440	24.4	7,861	4.6	75,854	5,063	6.7	2,996	3.9	94,250	36,461	38.7	4,878	5.2
port 3	169,458	41,727	24.6	8,929	5.3	75,815	5,128	6.8	3,465	4.6	94,151	36,651	38.9	5,488	5.8
port 4	169,186	41,851	24.7	9,788	5.8	75,528	4,929	6.5	3,642	4.8	94,149	36,980	39.3	6,168	6.6
port 5	168,627	41,318	24.5	10,758	6.4	75,433	4,692	6.2	4,017	5.3	93,706	36,681	39.1	6,769	7.2
port 6	167,739	41,428	24.7	11,529	6.9	74,711	4,322	5.8	4,063	5.4	93,524	37,169	39.7	7,501	8.0
port 7	167,101	41,771	25.0	12,754	7.6	74,726	4,135	5.5	4,321	5.8	92,877	37,719	40.6	8,463	9.1
port 8	165,692	43,160	26.0	14,479	8.7	73,891	3,970	5.4	4,604	6.2	92,278	39,235	42.5	9,907	10.7
port 9	164,142	45,735	27.9	17,352	10.6	73,322	3,863	5.3	4,811	6.6	91,294	41,934	45.9	12,571	13.8
High MAX	161,554	54,168	33.5	23,378	14.5	72,259	4,452	6.2	5,162	7.1	89,772	49,775	55.4	18,256	20.3

Table 2.9: MAX_{news} without Earnings Announcements

Table 2.9 reports the MAX associated with DJNS news coverage after dropping Earnings Announcement events across each MAX portfolio. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). DJNS denotes the MAX associated with DJNS news reports. EA events refers to Earnings Announcements events and EA DJNS News is Dow Jones-tagged earnings-related news reports. I identify the MAX as accompanied by DJNS news reports as long as it lies within a [t-1,t+1] window. MAX_{nonews}-MAX_{news} is the difference between the MAX_{nonews} and MAX_{news} portfolios. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighted in bold indicating a coefficient significant at the 10% level or better.

	Exclude EA events in MAX_{news}						Exclude EA events & EA DJNS News in MAX_{news}					
	MAX_{nonews}		MAX_{news}		MAX_{nonews}-MAX_{news}		MAX_{nonews}		MAX_{news}		MAX_{nonews}-MAX_{news}	
	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW
High - Low	**-1.38%**	**-1.14%**	**0.85%**	0.17%	**-2.24%**	**-1.31%**	**-1.34%**	**-1.07%**	0.75%	0.11%	**-2.09%**	**-1.18%**
	(-5.62)	(-3.57)	(1.79)	(0.32)	(-6.01)	(-2.94)	(-5.44)	(-3.42)	(1.56)	(0.20)	(-5.54)	(-2.60)
FFC4F	**-1.61%**	**-1.25%**	0.68%	-0.03%	**-2.29%**	**-1.22%**	**-1.57%**	**-1.17%**	0.55%	-0.10%	**-2.12%**	**-1.07%**
	(-11.24)	(-6.25)	(1.60)	(-0.07)	(-5.97)	(-2.62)	(-10.95)	(-5.93)	(1.30)	(-0.19)	(-5.45)	(-2.26)
FFC4F + PS	**-1.61%**	**-1.26%**	0.59%	-0.18%	**-2.20%**	**-1.08%**	**-1.57%**	**-1.18%**	0.46%	-0.25%	**-2.03%**	**-0.93%**
	(-11.13)	(-6.26)	(1.38)	(-0.37)	(-5.72)	(-2.32)	(-10.86)	(-5.93)	(1.07)	(-0.50)	(-5.19)	(-1.96)

Table 2.10: MAX_{news} and Investor Attention

Table 2.10 reports the Fama and MacBeth (1973) regression for MAX_{news} under different investor attention proxies. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). MAX_{news} denotes the MAX associated with DJNS news reports. I identify the MAX as being accompanied by DJNS news reports as long as they lie within the [t-1,t+1] window. The proxies for investor attention are natural logarithm of the total volume of DJNS news reports in a month $NEWSCVG$, natural logarithm of the number of analyst coverage in a month $ASTCVG$, and absolute value of the most recent standardized quarterly unexpected earnings $ABSSUE$. CONTROLS are: SIZE, BTM, BETA, MOM, REV, and ILLIQ. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, ***,**,* indicates a coefficient significant at the 1%, 5% or 10% level, respectively.

	Model 1	Model 2	Model 3	Model 4
CONST	1.3821**	1.9673***	1.3855**	2.0038***
	(2.22)	(2.90)	(2.27)	(2.91)
MAX_{news}	0.0140*	0.0149**	0.0149*	0.0132*
	(1.81)	(1.96)	(1.91)	(1.72)
NEWSCVG	0.0107			-0.0001
	(0.43)			(0.00)
ASTCVG		0.0867*		0.0892**
		(1.90)		(1.96)
ABSSUE			0.0109	0.0100
			(0.32)	(0.28)
CONTROLS	YES	YES	YES	YES

Table 2.11: MAX_{nonews}, MAX_{news} and Information Uncertainty

Table 2.11 reports the relationship between MAX stocks, news coverage, and information uncertainty. The Panel A reports the changes of stock return volatility around MAX days in different examining windows. The Panel B reports the coefficients from bivariate portfolio analysis based on information uncertainty (proxied by idiosyncratic volatility IVOL, earnings volatility EVOL, cash flow volatility CFVOL, and analyst forecast dispersion DISP) on both the equally-weighted and value-weighted basis. In each month, stocks are first ranked into terciles based on information uncertainty, and further sorted into deciles based on the MAX. For brevity, I only report the hedged portfolio return differences for low and high information uncertainty groups. MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). I identify the MAX as being accompanied by DJNS news reports as long as the coverage lies within the [t-1,t+1] window. MAX_{nonews}-MAX_{news} is the difference between the returns of the MAX_{nonews} and MAX_{news} portfolios. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighting in bold indicates a coefficient significant at the 10% level or better.

Panel A	[t-5,t-1] to [t+1,t+5]	[t-10,t-1] to [t+1,t+10]	[t-15,t-1] to [t+1,t+15]
MAX_{nonews}	**-0.91%**	**-2.14%**	**-2.30%**
	(-2.67)	(-5.49)	(-4.93)
MAX_{news}	**-1.95%**	**-2.85%**	**-3.14%**
	(-4.82)	(-6.27)	(-5.50)
MAX_{nonews}-MAX_{news}	**1.04%**	**0.71%**	**0.84%**
	(3.78)	(2.71)	(3.19)

Panel B	Low						High					
	MAX_{nonews}		MAX_{news}		MAX_{nonews}-MAX_{news}		MAX_{nonews}		MAX_{news}		MAX_{nonews}-MAX_{news}	
	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW	EW	VW
IVOL	0.27%	-0.23%	-0.09%	-0.04%	**0.36%**	-0.19%	**-2.11%**	**-1.26%**	-0.45%	-0.42%	**-1.66%**	**-0.84%**
	(1.56)	(-1.11)	(-0.48)	(-0.17)	(2.02)	(-0.78)	(-9.87)	(-3.80)	(-1.29)	(-0.94)	(-5.05)	(-1.67)
EVOL	**-1.02%**	**-0.81%**	**-0.77%**	-0.70%	-0.25%	-0.10%	**-0.88%**	**-0.62%**	-0.11%	-0.19%	**-0.78%**	-0.42%
	(-3.95)	(-2.37)	(-2.11)	(-1.57)	(-0.81)	(-0.24)	(-3.23)	(-1.69)	(-0.30)	(-0.47)	(-2.78)	(-1.08)
CFVOL	**-0.82%**	-0.81%	-0.20%	-0.02%	**-0.68%**	-0.87%	**-1.15%**	**-0.86%**	-0.36%	-0.34%	**-1.02%**	-0.74%
	(-2.01)	(-1.46)	(-0.43)	(-0.04)	(-1.95)	(-1.56)	(-2.87)	(-1.68)	(-0.80)	(-0.63)	(-2.92)	(-1.55)
DISP	-0.29%	0.07%	0.69%	0.66%	**-0.99%**	-0.59%	**-1.65%**	**-1.29%**	-0.83%	-0.50%	-0.82%	-0.78%
	(-0.63)	(0.12)	(1.28)	(1.11)	(-3.05)	(-1.24)	(-3.13)	(-1.92)	(-1.35)	(-0.67)	(-1.60)	(-1.24)

Table 2.12: MIN_{news} and VaR_{news}

Table 2.12 reports average monthly returns for MAX, MIN and VaR stocks associated with DJNS news reports on both an equally-weighted and a value-weighted basis. All hedged portfolio return differences, Fama-French-Carhart four-factor alpha returns and liquidity-augmented Fama-French-Carhart five-factor alpha returns between the highest and lowest groups are presented. MIN is computed by calculating the highest daily return in the prior month, following Bali et al. (2011). VaR is computed by calculating the first percentile of the daily returns over the previous year (250 trading days) as of the end of month t with at least 200 non-missing return observations, following Atilgan et al. (2020). I identify the MAX, MIN and VaR stocks as accompanied by DJNS news reports as long as the coverage lies within the [t-1,t+1] window. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighting in **bold** indicates a coefficient significant at the 10% level or better.

	MAX_{news}		MIN_{news}		VaR_{news}	
	EW	VW	EW	VW	EW	VW
High - Low	**1.34%**	0.70%	**-1.78%**	**-1.66%**	**-0.63%**	-0.19%
	(3.08)	(1.44)	(-3.79)	(-3.23)	(-1.69)	(-0.43)
FFC4F	**1.13%**	0.35%	**-1.89%**	**-1.76%**	**-0.78%**	-0.29%
	(3.03)	(0.80)	(-4.72)	(-3.78)	(-3.52)	(-0.91)
FFC4F + PS	**1.03%**	0.23%	**-1.91%**	**-1.78%**	**-0.79%**	-0.36%
	(2.76)	(0.52)	(-4.74)	(-3.80)	(-3.55)	(-1.15)

Table 2.13: MAX and Investor Sentiment

Table 2.13 reports average monthly returns for MAX stocks with/without DJNS news reports when investor sentiment is above/below its mean on both an equally-weighted and a value-weighted basis. High indicates that investor sentiment is above the sample average whereas low is below the average. Investor sentiment measures comprise: the BW (Baker and Wurgler Sentiment Index) from Baker and Wurgler (2006), the FEARS index (Financial and Economic Attitudes Revealed by Search) from Da et al. (2014a), the MS (Manager Sentiment) Index from Jiang et al. (2019), the MCSI (the Michigan Consumer Sentiment Index), the CBC (the Conference Board Consumer Confidence Index) and the HJTZ (an optimal BW Sentiment Index) from Huang et al. (2015). MAX is computed by calculating the highest daily return in the prior month following Bali et al. (2011). I identify the MAX as being accompanied by DJNS news reports as long as the coverage lies within the [t-1,t+1] window. The sample runs from August 1979 to December 2016. T-statistics are in parentheses, and highlighting in **bold** indicates a coefficient significant at the 10% level or better.

	MAX_{nonews}				MAX_{news}			
	High		Low		High		Low	
	EW	VW	EW	VW	EW	VW	EW	VW
BW	**-2.57%**	**-2.73%**	**-1.08%**	**-0.74%**	0.85%	-0.37%	**1.80%**	**1.50%**
	(-6.02)	(-4.88)	(-3.79)	(-1.94)	(1.12)	(-0.48)	(3.28)	(2.30)
FEARS	**-2.52%**	**-1.69%**	**-1.91%**	-1.18%	0.50%	-0.24%	0.78%	0.88%
	(-3.92)	(-2.04)	(-2.86)	(-1.33)	(0.77)	(-0.30)	(0.83)	(0.89)
MS	**-2.15%**	**-1.81%**	**-1.37%**	-0.65%	-0.06%	-0.13%	**2.46%**	**1.97%**
	(-5.10)	(-3.16)	(-2.57)	(-0.85)	(-0.13)	(-0.27)	(2.96)	(2.16)
MCSI	**-1.82%**	**-1.84%**	**-1.62%**	**-1.21%**	**1.40%**	0.76%	**1.22%**	0.43%
	(-5.14)	(-3.95)	(-5.19)	(-2.87)	(2.12)	(1.10)	(2.39)	(0.68)
CBC	**-2.19%**	**-2.32%**	**-1.35%**	-0.60%	0.58%	-0.09%	**1.76%**	**1.15%**
	(-5.38)	(-4.30)	(-4.22)	(-1.33)	(0.87)	(-0.12)	(3.02)	(1.73)
HJTZ	**-2.51%**	**-2.28%**	**-1.36%**	**-1.24%**	0.99%	-0.60%	**1.48%**	**1.21%**
	(-4.12)	(-2.94)	(-6.58)	(-4.13)	(0.92)	(-0.54)	(3.63)	(2.48)

Chapter 3

News-based Peers and Stock Returns

"Firms do not exist as independent entities, but are linked to each other through many types of relationships."–ANDREA FRAZZINI

3.1 Introduction

A vast literature in finance has studied lead-lag return momentums among economically linked firms (hereafter referred to as peers). Specifically, a firm's future returns can be predicted from its peers' past return performance. This lead-lag return predictability has been documented for different types of peers, such as industry, supplier-customer chain, product market, and for single- and multiple-segment firms in the same sector.[1]

The grouping schemes for identifying the different types of peers often yield sticky peers that cannot reflect the up-to-date relative importance of the economic

[1]For instance, there are lead-lag return effects among same industry (Moskowitz and Grinblatt, 1999); supplier-customer chain (see Cohen and Frazzini (2008) and Menzly and Ozbas (2010)); firm within same product markets (Hoberg and Phillips, 2018); or single- and multi-segment firms operating in the same sector (Cohen and Lou, 2012).

links. Consequently, it is difficult to identify the important peers of a firm reliably and timely. Consider some traditional industry classification schemes, like the Standard Industry Classification (SIC), as an example. Its limitation is not only restricted to the slow/no recognition of new and emerging industries, such as the high-tech sector, but also lies in classifying firms into mutually exclusive groups without any differentiation within a group. To address this challenge, I propose to use news co-coverage based on journalist-assigned subject tags to identify each firm's *news-based peers* (NBPs). A time-varying firm-centric grouping that aims to augment existing industry classification schemes is then constructed. To examine the usefulness of this approach, I test the lead-lag return momentums under different NBP-augmented industry classification schemes.[2]

The benefits of inferring economic links via news articles are three-fold. First, unlike time-invariant traditional industry classifications or annual classification schemes based on the 10-K updates, the up-to-date economic link between firms can be promptly picked up by journalists and revealed in news articles. Second, the number of news articles covering each link between firms provides a natural objective scheme to weight the relative importance of the links. As such, co-coverage news offers a potential wisdom-of-the-crowd solution to rate the links between firms, unlike the lack of differentiation among the links between firms of a traditional industry (e.g., based on SIC code). Third, the Dow Jones News Archive used in this study can provide a broad coverage of topics for different corporate events.[3] Aggregating information from multiple sources (including firm managers, analysts, regulators, and other market participants), journalists may offer additional information about firm peers through their in-depth coverage and insights. Therefore, news articles are

[2]By "NBP-augmented", I mean the modification of an industry classification scheme by keeping only the industry peers in the scheme that are also identified as NBPs and allowing a differentiation between these peers based on the NBPs' fraction weighting (to be explained further in this chapter).

[3]See Table A2 for the full list of topics selected for this study, which includes quarterly earnings, takeovers, analyst recommendations, insider buying and selling, dividend news, bond and stock registrations, labour, unions and strikes, etc.

likely to contain incremental information over and above what is captured by other industry classification schemes.

I provide four sets of analysis to understand the properties of NBPs: (i) Newsworthy links; (ii) Predictive tests; (iii) NBP-augmented schemes; (iv) Investor inattention. In the first set of analysis, I show that NBPs exhibit a strong newsworthy feature. Table A1 in the appendix presents the evolution of an illustrative base firm, Microsoft Corporation, and its NBPs in three different times: January 1999, May 2009, and August 2016. The variation of Microsoft's NBPs is quite substantial. Its highest-ranked peer changes from Netscape in 1999, to Yahoo! in 2009, and to Linkedin in 2016. This example demonstrates the newsworthy nature of the NBPs, illustrating their capability of capturing the time-varying links between firms.

Second, our predictive tests show that the newsworthy nature of NBPs is highly forward-looking. The return shocks transmitted from the NBPs can strongly predict the future returns of the base firm. I refer to this as the *NBP momentum*. I also document a monthly excess return of 1.06% (0.77%) achieved by forming an equal-weighted (value-weighted) dollar-neutral portfolio that buys stocks of the NBPs ranked in the highest quintile by their return shocks and sells those in the lowest quintile. This suggests that investors of the base firms tend to slowly react to the shocks affecting the NBPs.

Third, the advantage of NBP-augmented schemes over traditional industry classifications is to better capture the up-to-date relative importance of the links between a base firm and its peers picked up by journalists.[4] To show this, I first look at the SIC. I identify each firm's NBPs in the NBP-augmented SIC scheme and also its non-NBPs (i.e., SIC-only peers). In the one-month predictive test, I find a momentum effect more pronounced for the NBPs than the SIC-only peers. Sorting the SIC peers by how often they are also identified as the NBPs of some base firms,

[4]This includes the increase from no importance to a non-negligible level of importance for a previously non-existent link and the decrease to no importance for a previously existent link.

I observe a declined SIC-based momentum from the highest to the lowest quintile portfolio. All in all, the evidence is consistent with the notion that journalists can discover the up-to-date relative importance of the base-peer links.

The findings above can be generalized to other NBP-augmented schemes, including those based on the Fama-French 48-industry classification (FFi48), the text-based network industry classifications (TNIC) by (Hoberg and Phillips, 2010, 2016), the supplier-customer industry classification by Menzly and Ozbas (2010), and the single- and multiple-segment firm scheme by Cohen and Lou (2012). The “horse-race” tests between NBP momentum and the lead-lag return momentums of those classifications largely support the premise that co-coverage news articles have incremental information about firm links not already captured by the other industry classifications.

In the fourth set of analysis, I show that investor inattention is a key driver of NBP momentum. When including only a one-month gap in between the formation and the estimation period, the non-NBP part of the lead-lag return momentum of the SIC scheme virtually loses its predictability and is consistent with prior literature (e.g., Hoberg and Phillips (2018)). By contrast, the NBP momentum exhibits a long-standing pattern, which can last for several months. By focusing on abnormal retail attention measured in terms of the Google Searching Volume Index (SVI) after a shock, I observe a gradually increasing attention for NBPs that lags considerably behind the increasing attention for their SIC-based counterpart. The investor underreaction pattern for NBPs is in sharp contrast to the overreaction pattern for the traditional SIC peers.

Finally, a number of additional tests are performed for robustness checking. For example, to study why the base firms positively respond to the price shock of their NBPs, I show that the NBPs and their base firm are characteristically similar in terms of a variety of fundamental characteristics.[5] Moreover, I show that NBP momentum is not driven by small firms responding to large firms' earnings releases (see Hou (2007)), suggesting that the NBPs indeed capture the up-to-date relative importance of the base-peer links.

This chapter contributes to the literature in several aspects. First, using news co-coverage based on journalist-assigned subject tags, I construct more up-to-date time-varying news-augmented industry classification schemes. In contrast, prior related studies focus on disclosures updated annually (e.g., corporate filings like 10-K) or even longer (e.g., data for deriving the supplier and customer industry classification from the US Bureau of Economic Analysis that updates every five years) (Cohen and Frazzini (2008), Cohen and Lou (2012), Hoberg and Phillips (2010), and Hoberg and Phillips (2016)). The firm links embedded in these disclosures are slow to capture the latest development.

In making the contribution above, I demonstrate that NBP-augmented schemes can effectively capture information over and above what is already captured by existing industry classifications. By bringing together a variety of information sources from managers, analysts, regulators and sophisticated market participants, business press articles tend to be more credible in providing the latest and precise information about firm links (Kothari, Li, and Short, 2009). Other wisdom-of-the-crowd approaches are either aggregating investor perceptions through Internet co-searches (e.g., Lee, Ma, and Wang (2015)) or collecting analyst co-coverage in sell-side reports (e.g., Ali and Hirshleifer (2019)).[6] As investors would not co-search inconspicuous firm links and analyst coverage is heavily biased toward large and well-known firms,

[5]By a "characteristically similar" peer, I mean that the NBPs of a base firm exhibit a positive correlation in stock characteristics and valuation multiples (e.g., price-earnings ratio, return on equity, sales growth, leverage, etc.). As suggested by Ali and Hirshleifer (2019) and Müller (2019), lead-lag return momentums largely exist when the firms of a link share similar fundamental characteristics.

[6]Lee et al. (2015) identify firm peers by analyzing EDGAR users downloading the 10-K files of two firms in chronologically adjacent searches, known as search-based peers (SBPs).

there remains room for identifying peers with other approaches less subject to these limitations. I offer one alternative in this chapter.

I also contribute by showing that news plays an important role in the lead-lag return momentums documented in the literature. Prior studies utilise various implicit and less transparent economic links as natural settings for testing lead-lag return momentums (see e.g., Cohen and Frazzini (2008), Menzly and Ozbas (2010), Cohen and Lou (2012), and Hoberg and Phillips (2018)). There is little research on unifying the seemingly distinct findings.[7] I show that the lead-lag return momentums are partially driven by news that updates the relative importance of firm links, which is nevertheless underreacted by investors. In fact, investors not only overlook new and less visible links but also underreact to the latest and precise relative importance of stable important links. These findings shed new light on the previously documented lead-lag return momentums.

Lastly, I add to the news momentum literature by showing that the return momentum phenomenon is not only due to firm-specific news but also exists as a result of news co-coverage. Many papers find that stock return momentum pattern following news-driven price shocks is due to slow information diffusion (see e.g., Chan (2003), Frank and Sanati (2018) and Jiang, Li, and Wang (2020)). However, these studies mainly restrict their empirical settings to firm-specific news items, without testing the investor attention hypothesis using co-coverage news articles. Consistent with Hong and Stein (1999) and Barberis et al. (1998), I find that investors not only underreact to firm-specific news articles but also overlook the information embedded in news co-coverage.

The rest of this chapter is organized as follows: Section 3.2 discusses the related literature. Section 3.3 describes data and variable constructions. Section 3.4 provides the main findings of this chapter. Section 3.5 provides suggestive evidence

[7] I acknowledge that an exception is Ali and Hirshleifer (2019).

to explain the empirical finding, which is consistent with investor inattention theory. Some robustness checks are provided in Section 3.6. Section 3.7 concludes this chapter.

3.2 Related Literature

This chapter is related to the recent literature on alternative industry classifications to identify firm rivals/peers. The most important research in this literature is conducted by Hoberg and Phillips (2016). They construct the text-based fixed industry classification (FIC) and the text-based network industry classifications (TNIC) based on product descriptions in company 10-K filings. The rationale behind these classifications is that similar firms are likely to be identified as rivals if they have similar products provided for the market. The FIC follows the traditional approach to industry classification that imposes a fixed location property, i.e., a firm can only be assigned to one of the mutually exclusive industry groups. The TNIC relaxes this property to allow a firm to belong to potentially multiple industry groups. Their implementation of the TNIC is calibrated to have the same granularity as the three-digit SIC does.

Lee et al. (2015) define that the two firms as search-based peers (SBPs) if EDGAR users download the 10-K files of these two firms in chronologically adjacent searches. A high fraction of such Internet co-searches in a year suggests that two firms are closely connected. Their empirical tests show that top-ten SBPs dominate Global Industry Classification Standard (GICS) industry peers in explaining the cross-sectional variation in a number of firm characteristics.

Liu and Wu (2019) propose a time-varying Labor Market Competitor Network (LMCN) based on online job postings because firms with similar jobs and recruitment are likely to be labour market rivals. They also use a threshold-based approach to ensure that the LMCN has the same granularity as the three-digit SIC does.

In light of the studies above, I reckon that a convenient way to identify peers is to construct a firm-centric grouping scheme allowing for all sorts of firm links. I achieve this goal by proposing a wisdom-of-the-crowd approach that relies on the collective knowledge of journalists as revealed in news articles discussing multiple firms together. The information from multiple sources aggregated by journalists in these articles can potentially capture the up-to-date relative importance of the economic links that is not obvious to investors and analysts.

Our research is also related to the lead-lag return literature in finance and the information transfer literature in accounting. The former literature documents that under various classification schemes, the future stock returns of base firms are positively associated with their peers' past return performance, while the latter literature shows that earlier earnings surprises positively predict the stock returns of other firms in the same industry. For example, Cohen and Frazzini (2008) first find that a customer firm has a lead-lag return momentum on its base firm. Hoberg and Phillips (2018) find an industry momentum based on the TNIC scheme they develop (Hoberg and Phillips, 2010, 2016).[8] Consistent with these literatures, I also find lead-lag return momentums in various NBP-augmented schemes.

Finally, our work is relevant to the news momentum literature. Many studies find a stock return momentum pattern following news-driven price shocks, interpreted as underreaction to public news information. Chan (2003) first documents that investors tend to underreact to the price shock accompanied by news headlines, causing post-news return drifts. Recently, Jiang et al. (2020) documents a daily news momentum pattern, which is substantially stronger for overnight news, supporting the limited investor attention hypothesis. Extending this literature, I demonstrate that news momentum also exists for co-coverage news.

[8]Other studies in this stream include Menzly and Ozbas (2010), Cohen and Lou (2012), Ali and Hirshleifer (2019), and Liu and Wu (2019)).

3.3 Data and Methodology

3.3.1 NBPs' Identification

I use the Dow Jones Newswire Archive to identify the NBPs of a given base firm. The Dow Jones Archive is a professional newswire package, which provides historical major economic and business news since 1979. All articles are stored in an XML format with several useful tags, including headlines, main text, subject code, timestamp, and tickers. I rely on the tagged tickers to identify NBPs. I primarily extract all the news articles from the Archive between 1995 and 2017 with two to six tagged tickers.[9] To clean the data, I apply various filtering criteria.[10] This yields a sample of 1.4 million news articles for identifying the NBPs.

Collectively, panel A of Table 3.1 reports the descriptive statistics of the final NBPs. From January 1995 to December 2017, the NBP classification scheme covers on average 3,000 firms, of which 1,290 are not only NBPs of some base firms but are themselves eligible base firms.[11] The number of NBPs per base firm ranges from 1 to 284, with an average of 25. In terms of the equal-weighted percentage coverage, the number of base firms in our sample accounts for 25% of the CRSP universe. However, the market capitalization of the base firms in our sample accounts for 83% of the total market value of the CRSP universe. Thus, many of the base firms in our sample are large firms. Moreover, I show that the duration of NBPs on average only lasts for 1.4 years. The majorities only appear once a year (more than 75% percentile), suggesting the time-varying and newsworthiness feature of NBPs.

——— Insert Table 3.1 here ———

[9] The sample ends in 2017 because I have access to the data only up to that year.

[10] See the appendix for more details of the sample selection procedure. In particular, I discuss in subsection A.3 how our article-level approach is different from Schwenkler and Zheng (2019) sentence-level approach and the implications of this difference.

[11] Our procedure of identifying NBPs excludes a number of unqualified base firms, details of which can be found in the appendix.

3.3.2 Asset Pricing Variables

In regards to testing stock return predictability, our main variable of interest, *NbpRet*, is computed based on the one-month fraction-weighted stock return using all the NBPs of a base firm. The *NbpRet* for the base firm i in month t is defined as:

$$NbpRet_{it} = \frac{\sum_{j \neq i} Frac_{ijk} * Ret_{jt}}{\sum_{j \neq i} Frac_{ijk}} \tag{3.1}$$

where Ret_{jt} is the return of base firm i in month t, and $Frac_{ijk}$ is the fraction of news articles with co-coverage between firm i and firm j in the year k (i.e., running from January to December of the prior calendar year). I primarily use a calendar year-end cut-off for different NBP-augmented industry classification schemes, such as that based on the SIC code or on annual corporate filings. In later sections, I also apply a monthly-rolling scheme with different trailing windows to study NBPs' newsworthiness feature. By using the pairwise fraction as the weighting scheme, the base-peer links more frequently appearing in the news articles will be given higher weights.

NonNbpRet is the non-NBP-weighted peer returns as a control variable. To calculate this in the value-weighted NBP-augmented SIC scheme, I begin by focusing on the value-weighted return of the three-digit SIC peers of each base firm and then the fraction-weighted return of those SIC peers that are identified as NBPs. To ensure that our results do not critically depend on the choice of the scheme to augment, I also consider alternative value-weighted NBP-augmented schemes, including those based on the Fama-French 48-industry classification, the TNIC, the supplier-customer industry classification, etc. To identify a firm's product market rivals, I utilize the TNIC scheme following Hoberg and Phillips (2010, 2016). The supplier-customer industry momentum is calculated based on the annual input-output import matrix data from the US Bureau of Economic Analysis (BEA) following Menzly and Ozbas (2010). I also include conglomerate industry momentum following Cohen and Lou (2012). Specifically, the conglomerate firm return is calculated by constructing a "pseudo-conglomerate" using its different single-segment's industry returns in a sales-weighted scheme. I do not employ the method provided by Cohen and Frazzini (2008), primarily due to the small sample issue.[12]

For control variables, I construct size and book-to-market ratio variables following the same procedure described by Hoberg and Phillips (2018). Firm size is measured by the natural log of the market capitalization from CRSP. The yearly size variable for each firm is defined for the twelve months from July to the next June. The book-to-market ratio is computed following Hoberg and Phillips (2018) as well. Specifically, I use the book value of equity at each fiscal year end divided by the market capitalization from CRSP at the nearest calendar year end and then take the natural log of the ratio. This variable is also defined yearly like the firm size variable. Similarly, the variable is updated at the beginning of each July until next June. To control for momentum effects, I compute the own-firm momentum variable based on the stock return from the month $t-11$ to $t-1$ period. As standard in the momentum literature, I insert a one-month gap between holding periods and estimation periods in order to avoid one-month short-term return reversals.

The summary statistics of the above-mentioned variables are reported in panel B of Table 3.1.

[12] The data on customer-supplier links from Andrea Frazzini's web is updated until 2005 only and merely cover large firms.

3.4 NBP-augmented Schemes

3.4.1 Portfolio Analysis

I test whether NBPs are largely forward-looking in terms of lead-lag return momentum, i.e., whether the return shock from the NBPs can positively predict the future return of the base firm. A typical assumption in the lead-lag return momentum literature is that the firms involved are characteristically similar and thus are likely to be economically linked (see Menzly and Ozbas (2010), Ali and Hirshleifer (2019) and Müller (2019)). In a later section, I confirm that the NBPs and their base firm are indeed characteristically similar in a range of fundamental characteristics (e.g., price-earnings ratio, return on equity, sales growth, leverage, etc).

I proceed with calendar-time portfolio analysis. The primary aim of this exercise is to examine whether base firms' future returns monotonically increase with the portfolio returns of the NBPs when they are sorted into portfolios by their returns. Most importantly, I test whether different risk factors can explain the abnormal future returns generated by the portfolios.

To form the portfolios, I first sort all firms into quintiles based on the NBPs' returns at the end of the prior month. I then calculate the equal-weighted excess return of the base firms of a quintile portfolio at the end of the current month. I focus on equal-weighted excess return to increase the power of the test because small firms are more likely to be subject to stronger investor underreaction. These portfolios are rebalanced every month; so I eventually obtain the time-series return performance for each quintile across the whole sample. Besides the future returns of each quintile portfolio, I also calculate the hedge portfolio return between the highest and the lowest quintile and examine its risk-adjusted return. I consider five different risk factor models: CAPM, Fama-French 3-factor model (3F), Fama-

French-Carhart 4-factor model (4F), a liquidity-augmented Fama-French-Carhart 5-factor model (4F + PS), and Fama-French 5-factor model (5F)[13].

Table 3.2 reports the results of different portfolios. Specifically, the hedged portfolio between highest and lowest quintile yields on average 1.06% monthly excess returns (XRET), which is significant at the 1% level. The next five columns display the risk-adjusted return after controlling for well-known risk factors. The long-short portfolio reports an average CAPM-adjusted monthly return of 1.16%. The same strategy generates a monthly risk-adjusted return of 1.14% and 1.12% when using the Fama-French 3-factor and the Fama-French-Carhart 4-factor model, respectively. A liquidity-augmented Fama-French-Carhart 5-factor model delivers a 1.20% risk-adjusted return per month. Lastly, the Fama-French 5-factor adjusted alphas are 1.44% per month. Alternatively, I also employ a value-weighted scheme to test whether the NBPs trading strategy is economically meaningful. Moving to the value-weighted portfolio, results are quantitatively similar albeit exhibiting a smaller magnitude. For example, the hedge return difference between lowest and highest quintile achieves 0.77%, which is significant at 5% level. I conclude that these risk-adjusted models cannot explain the abnormal return. Collectively, these results confirm the NBP momentum phenomenon, even after controlling for common risk factors.

——————————— Insert Table 3.2 here ———————————

3.4.2 Cross-sectional Regressions

I compare the return performance of NBPs and non-NBPs in various NBP-augmented industry classification schemes using Fama and MacBeth (1973) regres-

[13]The Fama-French risk factors can be downloaded from Kenneth French's data library `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. The liquidity factor I use is Pástor and Stambaugh (2003), which can be accessed from `https://faculty.chicagobooth.edu/lubos-pastor/data`

sions. Suppose that compared to traditional industry peers not identified as NBPs, the NBPs can better reflect the up-to-date relative importance of the base-peer links underreacted by investors. Then I should expect to see a more pronounced momentum effect for the NBPs than for the non-NBPs.

Our analysis proceeds as follows: I regress the one-month fraction-weighted return of the NBPs' return of a base firm on the one-month lagged return. For ease of comparison, I also consider the specification including the non-NBPs return as an independent variable:

$$Ret_{i,t} = \alpha + \beta_1 * NbpRet_{i,t-1} + \beta_2 * NonNbpRet_{i,t-1} + \beta_2 * X_{i,t-1} + \epsilon_{i,t} \tag{3.2}$$

The dependent variable is the current own-return of each base firm i in month t. The main independent variables are its NBPs' returns and Non-NBPs' returns in month $t-1$, respectively. In the vector X, I consider other firm characteristics. In particular, past one-month return, which is associated with the short-term return reversal and momentum calculated between month *t-2* to *t-12*, which skips the most recent month to avoid microstructure effects; these two variables might contribute to the high future return of a base firm. In all regression specifications, I include basic control variables, namely $Size$, BTM, Rev, Mom. Following the initial hypothesis, a greater positive and significant β_1 coefficient than the β_2 coefficient of non-NBPs implies that investors are more likely to underreact to the newsworthy firm links rather than stable important ones.

Table 3.3 reports the coefficients, T-statistics, and adjusted R^2 under various specifications. In columns (1) and (2), where the NBPs' and Non-NBPs' returns are the sole explanatory variable in the regressions, respectively, in the NBP-augmented SIC scheme, the former shows a larger price impact than the latter (0.034 versus

0.023). To see which of these has a stronger incremental impact than the other, column (3) reports the results when both variables are present. A t-test shows that the coefficient of the former (0.027) is greater than the latter's (0.016). Changing to the coarser FFi48 scheme does not qualitatively alter this finding, as I observe that the coefficient of $NonNbpRet$ is only half of the coefficient of $NbpRet$ in magnitude. In the NBP-augmented TNIC scheme, the $NbpRet$ continues to outperform than the $NonNbpRet$, albeit with a smaller gap. The NBP-augmented IO and NBP-augmented Conglomerate schemes also witness similar findings.

Overall, I find that the lagged NBPs can explain monthly return better than the Non-NBPs in various industry classifications.

——————— Insert Table 3.3 here ———————

3.4.3 Disparity Test

Next, I study how NBPs play a role in different lead-lag return momentums. If NBPs indeed reflect the up-to-date relative importance of the base-peer links that is underreacted by investors, I should see that lead-lag return momentums documented in the literature will become more pronounced when many traditional industry peers are also NBPs. In contrast, the momentum will be weaker if few traditional industry peers are identified as NBPs.

I design a double-sorted exercise to examine this premise. Specifically, in each month, I calculate the "disparity" level between NBPs and each industry peers. The disparity is defined as 1 minus the total number of firms in the intersection of NBPs and industry peers divided by the total number of firms in the union of NBPs and industry peers. All firms are assigned to quintiles based on the disparity level in the prior month. The lowest quintile contains the firms with a small number of NBPs whereas the highest quintile has a large number of NBPs. After that, I regress the one-month lagged peers' returns on the own-firm return in the current month within each disparity quintile. For brevity, I only report the regression results from the highest and lowest quintile for each industry. All the explanatory variables in the regressions are standardized. In all regression specifications, I include basic control variables namely $Size$, BTM, Rev, Mom. Following the hypothesis, a greater positive and significant coefficient estimate of the low disparity group than that of the high disparity group indicates that lead-lag return momentum is more likely to happen when more industry peers are identified as NBPs.

As can be seen in Table 3.4, I first report the regression results for three-digit SIC-based momentum. Regression specification (1) represents the lowest disparity portfolio between NBPs and three-digit SIC peers while (2) is the highest group. The coefficient estimate of the former substantially outperforms the latter (0.036 versus 0.010). A t-test shows that the coefficient of the former is statistically significant at 1% whereas the latter is insignificant. This suggests that the NBPs can fully digest SIC-based momentum. Moreover, I continue to observe the similar finding in the coarser FFi48 industry momentum and TNIC industry momentum, though the t-test shows that both coefficients in the highest disparity group remain statistically significant at 5% (t-statistics are 2.08 and 2.52, respectively). This indicates that our NBPs can partially digest FFi48 and TNIC momentum. Lastly, I observe that IO industry momentum and conglomerate industry momentum becomes insignificant in the highest disparity group. Interestingly, the adjusted R^2 also suggests that the regression specification in the low disparity portfolio consistently has higher explanatory power compared to that of the high disparity portfolio.

Overall, the exercise shows that different industry momentums with more NBPs consistently outperform than those with few ones. The results highlight the role of NBPs in unifying each seemingly distinct finding.

Insert Table 3.4 here

3.5 Investor Attention Theory

3.5.1 Long-run Performance

To test the investor attention hypothesis, prior studies have looked at the return performance over lagged returns in various horizons. If prices indeed slowly incorporate new information, one would expect that the coefficient of the lagged return would consistently show positive statistical significance for the different horizons consistently.

I carry out the exercise as follows: I hold the dependent variable unchanged but lag the *NbpRet* up to four months. To easily compare the long-run performance between the NBPs and Non-NBPs in the SIC scheme, I report both coefficients. Selecting the NBP-augmented SIC as a benchmark is due to its highly visible properties to investors. As such, one might observe that the two groups of firm links exhibit very different long-run return performance. Another benefit of using the three-digit SIC is that I can further distinguish NBP momentum from SIC-based industry momentum (Moskowitz and Grinblatt, 1999). Grundy and Martin (2001) document that SIC-based industry momentum tends to be short-lived. Thus, if I do observe that NBP momentum shows a long-standing pattern, it then should be clear that two groups are very different from each other.

As seen in Table 3.5, when lagging two months, it can be seen that the *NbpRet* remains a positive and significant coefficient, whereas the $IndSic3Ret$ generally loses its predictability power. This result appears to be consistent with Grundy and Martin (2001) who complain that industry momentum for SIC peers is not robust to lagging the portfolio formation periods by more than one-month periods. Moreover, the results also support the investor attention hypothesis, as the coefficient of *NbpRet* virtually drops its magnitudes and statistical level along with longer horizons.

In addition, I also perform a long-run calendar-time portfolio analysis. This exercise aims to study whether the stock return predictability of NBPs reverses in the long run. The investor attention hypothesis predicts that investors' slow reaction to the information transmitted from the NBPs to their base firms' prices should not reverse in the future. Specifically, I employ two six-month holding periods: Month $t+2$ to $t+7$ and Month $t+8$ to $t+13$, respectively. For brevity, I only report the excess returns and Fama-French five-factor alphas on average for the entire sample. In Table 3.5, it is evident that the hedge return difference between highest and lowest quintiles gradually declines, until close to statistically insignificance.

Overall, it seems plausible to conclude that the NBPs are indeed less visible to investors compared to the non-NBPs in the three-digit SIC scheme predicted by the investor attention hypothesis. The results also highlight the unique feature of NBP momentum in terms of its long-standing return predictability.

——————————— Insert Table 3.5 here ———————————

3.5.2 Bivariate Portfolio Analysis

Our next exercise to examine investor attention and NBP momentum is in a bivariate portfolio analysis. This exercise aims to study how these investor attention indicators are associated with the NBP momentum. Following the literature, conventional firm-level investor attention indicators include firm size, analyst coverage, and institutional ownership. I utilize all these attention measures and hypothesize that the NBP momentum will be more profound among firms receiving low investor attention.

The exercise proceeds as follows: Specifically, in each month, I first sort all firm into two groups conditional upon the current month's attention measure above or below the median level. Within each group, I further rank firms into quintiles based on the NBPs' returns in the same month. In Table 3.6, I only tabulate the average time-series portfolio holding period returns across the entire sample for the highest quintile, the lowest quintile and hedge return difference between highest and lowest quintiles due to space constraints. Risk-adjusted returns are also reported after applying CAPM, 3F, 4F, 4F + PS, and 5F models, respectively.

The results appear to be consistent with the investor attention hypothesis. When a firm receives low investor attention (i.e. small size, low analyst coverage, low institutional ownership), the hedge return difference is substantially significant. As a result, the test shows a stronger NBP momentum in the below-median subsample and further supports the initial hypothesis.

——————————— Insert Table 3.6 here ———————————

3.5.3 Industry Peer Returns and Abnormal Retail Attention

To further examine whether investor underreaction is the key driver of NBP momentum, I provide direct evidence to support this premise. Motivated by the abnormal Google Search Volume Index (denoted as ASVI hereafter) - a direct proxy of retail investors' attention (See, e.g., Da, Engelberg, and Gao (2011)) - I study the relation between ASVI and industry peer returns. Figure 3.1 Graph A plots aggregated monthly ASVI from either SIC peers or NBPs experiencing the highest quintile of return shocks across different months. Although one can observe that the ASVI from NBPs is lower than the SIC peers in the first three months, the difference between these two is modest. Presumably, the small gap is due to the high correlation between NBPs returns and three-digit SIC industry returns, as unconditional results are likely to be driven by the same underlying set of firms that exists in both classifications.

To mitigate the bias induced by the high potential correlation between the two schemes, I plot the conditional ASVI for both classifications where NBPs returns are in the highest quintile and SIC peer returns are in the middle quintile (i.e. 40th and 60th percentile so the return shocks are around zero) and vice versa. In other words, I perform an independent double-sorted portfolio exercise where the firms are from the highest NBPs(SIC peers) return quintile and the middle SIC peers (NBPs) return quintile. As such, this conditional sort allows me to examine how ASVI develops when one group of stocks experiences shocks but not the other. Figure 3.1 Graph B this time shows that the SIC has much higher ASVI than the NBPs during month 0. When moving to the following months, the retail attention from SIC gradually decreases, whereas those of NBPs are relatively stable. Collectively, these results suggest that investors, especially individual investors, initially underreact to the return shock from NBPs more than the SIC peers, indicating the key driver of NBP momentum.

——————————— Insert Figure 3.1 here ———————————

3.6 Robustness Checks

3.6.1 Fundamental Characteristics

I investigate whether the NBPs are characteristically similar to the base firm in a variety of firm characteristics. As a typical assumption of testing lead-lag return momentum is that firms are economically linked or characteristically similar. As such, I follow the research design of (Lee et al., 2015) and consider examining the degree of the association on multiple firm dimensions including stock characteristics, financial ratios, as well as valuation multiples. The primary reason of picking up each of those accounting variables aims at capturing different aspects of a firm's fundamentals.

I design the exercise proceeding as follows: First of all, I collect a wide range of accounting ratios from the Compustat universe following Lee et al. (2015). These financial ratios include PB (Price-to-book ratio), EVS (Enterprise value-to-sales ratio), RNOA (Return on net operating assets), LEV (Leverage), ROE (Return on equity), AT (Inverse of Asset turnover), PM (Profit margin), SALESGROWTH (One-year-ahead realized sales growth), RD (Expense-to-sales ratio), and SUE (Standardized unexpected earnings). The details of each variable construction can be found in Table A3 in the Appendix. For the entire firm year-quarter fundamentals' sample, I follow the same procedure: if there are any missing values on total assets, long-term debt, net income before extraordinary items, debt in current liabilities, or operating income after depreciation, I drop entire rows. I require non-negative common or total equity and retain net sales larger than \$100 million. Stock prices at the end of each quarter are at least higher than \$3 per share. To mitigate the influence of outliers, I truncate observations at the 1% and 99% percentiles for each variable. When computing RNOA, I impose non-negative net income before extraordinary items and non-missing values for current liabilities, current assets, and property, plants, and equipment. Next, I utilize these different firms' financial ratios in a simultaneous quarterly cross-sectional regression as follows:

$$Var_{i,t} = \alpha + \beta * Var_{p,t} + \epsilon_{i,t} \tag{3.3}$$

For each base firm i in year quarter t, I construct its main variable of interest using a fraction-weighted scheme for NBPs after dropping the base firm itself. To mitigate the influence of outliers, I truncate observations at the 1% and 99% level for each regression specification. To facilitate comparisons, all variables are standardized to have zero mean and unit standard deviation before performing the regressions.

As reported in Table 3.7, I comprehensively evaluate the characteristic association between base firms and their NBPs. Each column represents one regression specification. I predict that if the NBPs are indeed pairwise economic links between base-peer firms, the β coefficient of NBPs should be statistically significant.

——————————— Insert Table 3.7 here ———————————

In all regression specifications, I find that the NBPs achieve positive coefficients. For example, the regression specification reports the ability to explain the cross-sectional variation of SUE. One standard deviation of the NBPs will lead to a 5.30% impact on the base firm in terms of SUE, which is significant at the 1% level. Other earnings-related measures report similar results including RNOA, ROE, and SALESGROWTH.

In addition, I also evaluate to what extent the NBPs are economically linked to a base firm by lagging independent variables one period. By doing so, I can disentangle the directional information flow from the NBPs to a base firm, not vice versa. The exercise proceeds using the one-quarter lagged cross-sectional regression as follows:

$$Var_{i,t} = \alpha + \beta * Var_{p,t-1} + \epsilon_{i,t} \tag{3.4}$$

For each base firm i in year quarter t, I construct its lagged main variable of interest using both NBPs in year quarter $t-1$ for different firm fundamentals.

The tabulated results show a similar pattern as observed in the prior exercise. One quarter lagged NBPs significantly predict a base firm in all regression specifications.

Overall, it can be concluded that the NBPs and their base firm are characteristically similar in a variety of financial performance measures. Such evidence suggests that the NBPs are highly informative and are likely to capture pairwise economically linked base-peer firms.

3.6.2 NBPs with Shorter Windows

I further examine the newsworthiness feature of NBPs. Specifically, I study its robustness when NBPs are constructed by alternative rolling schemes with different trailing windows. Specifically, a monthly-rolling-forward scheme with a trailing window of 3, 6, and 9 months is employed. It is expected that the NBP momentum will consistently outperform than the non-NBPs for alternative constructions.

In Table 3.8, I report all coefficients of one-month lagged NBPs' returns with different trailing windows in the NBP-augmented SIC scheme. All regression specifications include $Size$, BTM, Rev, Mom. As can be seen, all coefficients of NBPs are statistically significant at the 1% level. The number of base stocks on average declines along with the decrease in trailing windows as a shorter window will include fewer recent observations. This evidence largely confirms that the NBPs' newsworthiness feature exists for a different choice of windows.

Insert Table 3.8 here

3.6.3 NBPs with Abnormal News Co-coverage

One potential concern is how to reconcile our finding with Hou (2007), who documents that intra-industry momentum is essentially post-announcement drift of small firms following the earnings release of big firms within the industry. Indeed, a large firm tends to have news coverage every day whereas a small firm tends to have spare news observations. As such, the top-ranking NBPs are likely to capture large firms rather than the up-to-date relative importance of peers. To show that our finding is not merely picking up a size effect, I construct NBPs using abnormal news co-coverage. Specifically, for all co-coverage news of each pairwise base-peer firm in year t, I calculate median values. By subtracting the median from the raw co-coverage data, I obtain a refined yearly co-coverage metric. The refined NBP-augmented SIC scheme is then constructed to examine whether NBPs indeed capture the up-to-date relative importance of firm links.[14] As can be seen in Table 3.8 Panel B, the coefficient of NBPs remains higher than that of Non-NBPs.

3.6.4 NBPs with no Earnings News Co-coverage

Earnings announcements are times when major news is released. The prior literature shows that the earnings announcements of peers will lead to the subsequent response of a base firm. Hence, I must account for this phenomenon to examine whether the responses to the earnings announcement of peers drive our result or not.

I first find that earnings announcement NBPs only constitute about 30% of all NBPs each month. To see whether earnings NBPs drive NBP momentum, I repeat the NBP-augmented scheme test but exclude all NBPs that have a known earnings announcement in the month when *NbpRet* is calculated. Even after excluding earnings NBPs, the results remain similar compared to those from Table 3.3. I show that a one standard deviation increase of *NbpRet* leads to a positive response of 2.50% per month in the NBP-augmented SIC. In conclusion, post-earnings announcements drifts are not an important driver of NBP momentum.

[14] I only retain abnormally high news co-coverage to construct NBPs as the primary aim is to mitigate size effects. In unreported tests, NBPs with abnormally low news co-coverage achieve similar results.

3.7 Summary

The literature has documented different types of firm links as well as the associated lead-lag return momentums. However, the existing schemes for identifying firm links often yield sticky peers that cannot timely reflect changes in the firm links. Using co-coverage news articles, I address this challenge by constructing a time-varying firm-centric grouping scheme aimed to augment existing industry classification schemes. I show that the proposed NBP-augmented schemes are better in capturing the up-to-date relative importance of the firms links.

NBPs exhibit a lead-lag return momentum. When a price shock is observed for the NBPs, the base firm's return positively responds and persists for several months, showing an NBP momentum. In various NBP-augmented schemes, the NBP momentum performs reasonably well and is stronger than the momentum based on traditional industry peers that are not NBPs. This suggests that news co-coverage plays an important role in the lead-lag return momentums documented in the literature.

NBP momentum appears to be consistent with the investor attention hypothesis. Several empirical tests show that investors are not immediately aware of the up-to-date relative importance of the firm links captured by co-coverage news articles. These results suggest that investors not only overlook new and less visible links but also underreact to the up-to-date relative importance of existing firm links.

Given the newsworthy nature of NBPs, the proposed NBP-augmented schemes have implications for both academics and practitioners. For academics, NBPs provide a potential avenue for different asset pricing and corporate finance applications - for example, revisiting peer effects using different NBP-augmented schemes. As for practitioners, the takeaway is that monitoring news happening about economic links between firms is a key to understand lead-lag return momentums.

Figure 3.1: Event Study of Unconditional and Conditional ASVI

Graph A is the event-study of the base firm's ASVI (Abnormal Search Volume Index) surrounding months when the firm's SIC3 or NBP3 have returns in the highest quintile. ASVI is constructed following Da, Engelberg, and Gao (2011), defined as the daily Google Search Volume Index score calculated on a scale of 0 to 100 minus its past 12-month mean and then divided by the mean. Graph B plots the conditional ASVI for the base firms surrounding months when the firm's SIC3 have returns in the highest quintile while NBP has returns between 40th and 60th percentiles. Analogously, I plot results when NBPs have high-quintile returns and three-digit SIC peers have returns in the 40th to 60th percentiles.

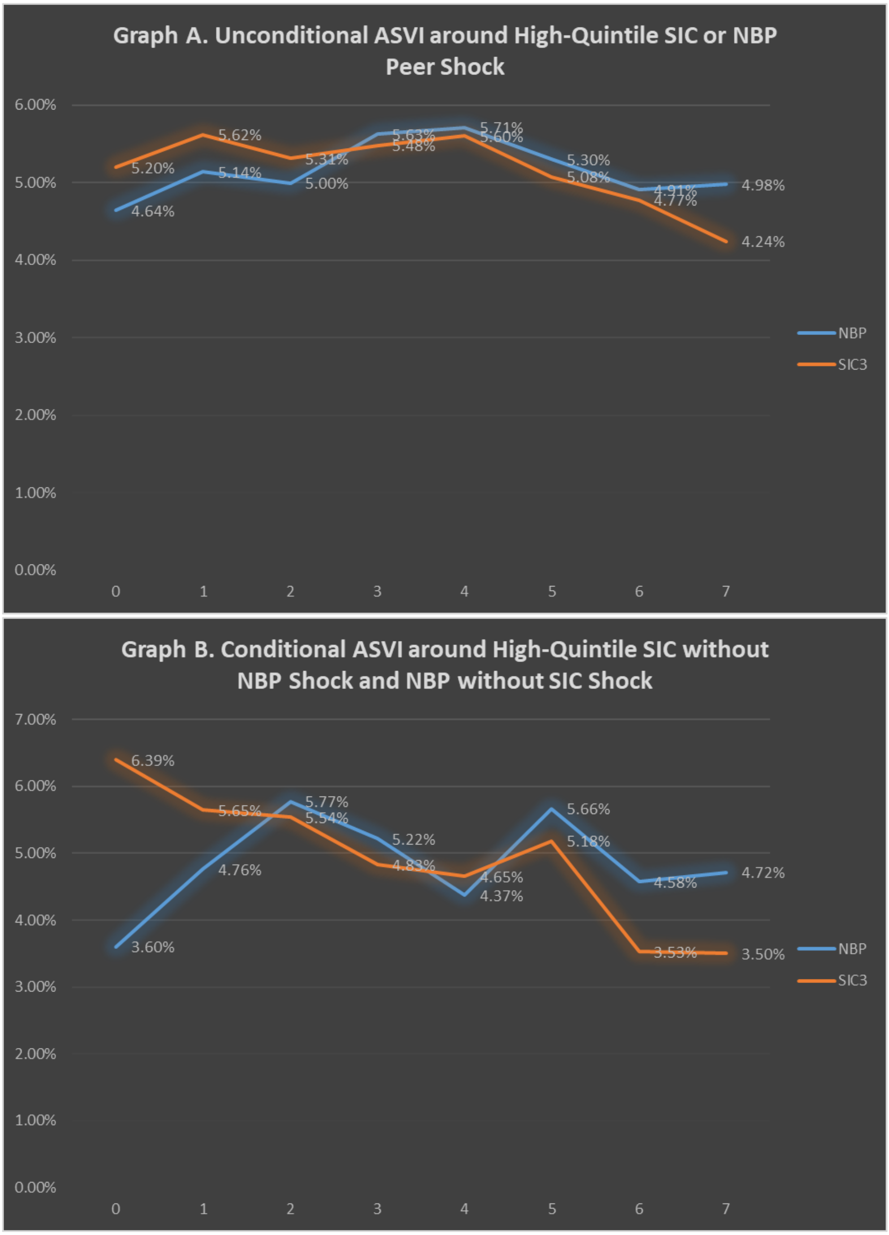

Table 3.1: Summary Statistics

Table 3.1 Panel A shows the summary statistics of NBPs as of every end of December. Percentage coverage stock universe (EW) is the number of base firms with identified NBPs divided by the total number of CRSP stocks. The percentage coverage of stock universe (VW) is the total market capitalization of base firm with identified NBPs, divided by the total market capitalization of the CRSP stock universe. Panel B reports summary statistics for asset pricing variables. One observation is 1 firm in 1 month. $NbpRet$ is computed based on the one-month fraction-weighted stock return using all NBPs of a base firm. $IndSic3Ret$ is the value-weighted return of the three-digit SIC peers excluding itself. $FFi48Ret$ is the value-weighted return of Fama-French 48-industry peers excluding itself. $Tnic3Ret$ is the similarity-weighted return of the Text-Based Industry Classification peers excluding itself. $IORet$ is the import-weighted supplier industry value-weighted returns based on the BEA Input-Output import matrix. $ConigloRet$ is calculated for each conglomerate using the weighted standalone firm returns of its each segment in the same industry. $Size$ is computed by taking the natural log of the CRSP market capitalization at the end of each June. BTM is computed by taking the book value of equity from each fiscal year end and divided by the CRSP market capitalisation at the end of each December in the calendar year. Mom is a firm's own momentum return from t-11 to t-1.

Panel A: Time Series (Jan 1995 - Dec 2017)

	Mean	**Std**	**Min**	**0.25**	**0.50**	**0.75**	**Max**
# of NBP scheme firms	3006	1364	1265	1979	2413	3694	5484
# of base firms	1290	596	508	817	1028	1675	2591
# of peer firms per base firm	25	22	1	12	17	27	284
Base firm % coverage of stock universe (EW)	25%	8%	10%	21%	24%	31%	36%
Base firm % coverage of stock universe (VW)	83%	9%	53%	80%	85%	88%	94%
duration (year)	1.4	1.5	1.0	1.0	1.0	1.0	23.0

Panel B: Asset Pricing Variable

	Mean	**Std**	**Min**	**0.25**	**0.50**	**0.75**	**Max**
Ret	0.01	0.18	-0.98	-0.07	0.01	0.07	11.80
$NbpRet$	0.01	0.09	-0.68	-0.04	0.01	0.05	2.59
$IndSic3Ret$	0.01	0.09	-0.66	-0.04	0.01	0.06	1.32
$FFi48Ret$	0.02	0.07	-0.47	-0.02	0.02	0.05	1.09
$Tnic3Ret$	0.01	0.11	-0.86	-0.04	0.01	0.06	6.27
$IORet$	0.01	0.07	-0.36	-0.02	0.01	0.05	0.39
$CongloRet$	0.01	0.07	-0.79	-0.02	0.01	0.05	3.67
$Size$	14.51	1.90	7.76	13.21	14.48	15.86	20.61
BTM	-0.98	0.97	-9.87	-1.50	-0.90	-0.36	3.63
Mom	0.14	0.79	-1.00	-0.21	0.06	0.33	49.98

Table 3.2: Calendar-time Portfolio Analysis

Table 3.2 reports the average excess returns and alphas of Calendar-time portfolio analysis for one-month predictive returns. To form the portfolio, I first sort all firms into quintiles based on the NBPs' returns at the end of each month. I then calculate the equal-weighted and value-weighted excess return of all based firms within each quintile in the next month. These portfolios are rebalanced every month. XRET is the excess return over the risk-free rate. I include five different risk factor models to compute alphas: CAPM, Fama-French 3-factor model (3F), Fama-French-Carhart 4-factor model (4F), a liquidity-augmented Fama-French-Carhart 5-factor model (4F + PS), and Fama-French 5-factor model (5F). T-statistics are in parentheses.

	Panel A: Equally-weighted Portfolio Analysis							**Panel B: Value-weighted Portfolio Analysis**					
	XRET	**CAPM**	**3F**	**4F**	**4F+PS**	**5F**		**XRET**	**CAPM**	**3F**	**4F**	**4F+PS**	**5F**
Port1	0.13%	-0.73%	-0.79%	-0.58%	-0.67%	-0.60%	**Port1**	0.24%	-0.51%	-0.52%	-0.40%	-0.38%	-0.35%
Low	(0.29)	(-2.94)	(-3.59)	(-2.86)	(-3.15)	(-2.66)	**Low**	(0.58)	(-2.24)	(-2.27)	(-1.77)	(-1.67)	(-1.46)
Port2	0.61%	-0.17%	-0.21%	-0.05%	-0.06%	-0.14%	**Port2**	0.37%	-0.28%	-0.27%	-0.21%	-0.24%	-0.24%
	(1.57)	(-0.98)	(-1.53)	(-0.41)	(-0.53)	(-0.99)		(1.13)	(-1.99)	(-1.89)	(-1.48)	(-1.72)	(-1.61)
Port3	0.83%	0.09%	0.05%	0.22%	0.21%	0.09%	**Port3**	0.78%	0.18%	0.18%	0.24%	0.24%	0.13%
	(2.29)	(0.64)	(0.42)	(2.58)	(2.52)	(0.81)		(2.72)	(1.85)	(1.94)	(2.56)	(2.53)	(1.31)
Port4	0.94%	0.21%	0.15%	0.32%	0.31%	0.25%	**Port4**	0.59%	0.00%	0.02%	0.05%	0.07%	0.09%
	(2.56)	(1.29)	(1.17)	(3.05)	(2.86)	(1.88)		(2.02)	(-0.04)	(0.17)	(0.47)	(0.65)	(0.78)
Port5	1.19%	0.43%	0.35%	0.53%	0.53%	0.54%	**Port5**	1.01%	0.39%	0.38%	0.47%	0.47%	0.40%
High	(2.89)	(1.87)	(1.80)	(2.98)	(2.85)	(2.77)	**High**	(3.03)	(2.13)	(2.07)	(2.63)	(2.58)	(2.09)
High - Low	**1.06%**	**1.16%**	**1.14%**	**1.12%**	**1.20%**	**1.14%**	**High - Low**	**0.77%**	**0.90%**	**0.90%**	**0.88%**	**0.85%**	**0.75%**
	(3.22)	**(3.50)**	**(3.44)**	**(3.32)**	**(3.40)**	**(3.34)**		**(2.16)**	**(2.53)**	**(2.52)**	**(2.42)**	**(2.33)**	**(2.01)**

Table 3.3: Cross-sectional Regressions

Table 3.3 reports the monthly cross-sectional regression results for one-month predictive returns between base and peer firms based on the formula as follows in various NBP-augmented schemes:

$$Ret_{i,t} = \alpha + \beta_1 * NbpRet_{i,t-1} + \beta_2 * NonNbpRet_{i,t-1} + \beta_3 * X_{i,t-1} + \epsilon_{i,t} \tag{3.5}$$

The dependent variable is the current own-return of each base firm. The main independent variables are the NBPs' returns and Non-NBPs' returns in month $t-1$, respectively. Observations are required to be in the Center for Research in Security Prices (CRSP), Compustat, the NBPs datatbase, and each industry classification. One observation is 1 firm in 1 month. *NbpRet* is computed based on the one-month fraction-weighted stock return using all NBPs of a base firm. $NonNbpRet$ is calculated for each base firm using its different industry peers' returns excluding itself and these identified as NBPs. Specifically, $NonNbpRet$ in the NBP-augmented SIC scheme is the value-weighted return of those SIC peers that are not identified as NBPs. In the NBP-augmented FFi48 scheme, it is the value-weighted return of those Fama-French 48-industry peers that are not identified as NBPs. In the NBP-augmented TNIC scheme, it is the similarity-weighted return of those Text-Based Industry Classification peers that are not identified as NBPs. In the NBP-augmented IO scheme, it is the import-weighted supplier industry value-weighted returns based on the BEA Input-Output import matrix. In the NBP-augmented Conglo scheme, it is calculated for each conglomerate using the weighted standalone firm returns of its each segment in the same industry. All the explanatory variables in the regressions are standardized. Control variables include $Size$, BTM, Rev, and Mom. T-statistics are in parentheses.

Cross-sectional Regressions: One-month Lagged

	NBP-augmented SIC			NBP-augmented FFi48			NBP-augmented TNIC			NBP-augmented IO			NBP-augmented Conglo		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
NbpRet	0.032		0.027	0.032		0.028	0.032		0.025	0.029		0.026	0.038		0.032
	(5.86)		(5.55)	(5.86)		(5.89)	(5.60)		(5.07)	(4.43)		(4.41)	(5.01)		(4.37)
$NonNbpRet$		0.023	0.016		0.022	0.014		0.029	0.022		0.025	0.017		0.026	0.018
		(4.38)	(3.31)		(3.05)	(2.03)		(4.39)	(3.46)		(1.88)	(1.35)		(3.15)	(2.21)
Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$Adj.R^2$	6.05%	6.16%	6.41%	6.06%	6.23%	6.48%	6.10%	6.27%	6.52%	6.85%	7.00%	7.38%	7.48%	7.58%	8.04%
$Avg\,base\,stocks$	1116	1116	1116	1126	1126	1126	1097	1097	1097	828	828	828	334	334	334

Table 3.4: Disparity Tests

Table 3.4 reports the monthly cross-sectional regression results for one-month predictive returns conditional upon the disparity level between NBPs and other industry peers. For each industry scheme, the disparity is defined as 1 minus the total number of firms in the intersection of NBPs and industry peers divided by the total number of firms in the union of NBPs and industry peers. For brevity, I only report the regression results from the highest and lowest quintile. $IndSic3Ret$ is the value-weighted return of the three-digit SIC peers excluding itself. $FFi48Ret$ is the value-weighted return of Fama-French 48-industry peers excluding itself. $Tnic3Ret$ is the similarity-weighted return of the Text-Based Industry Classification peers excluding itself. $IORet$ is the import-weighted supplier industry value-weighted returns based on the BEA Input-Output import matrix. $CongloRet$ is calculated for each conglomerate using the weighted standalone firm returns of its each segment in the same industry. All the explanatory variables in the regressions are standardized. Control variables include $Size$, BTM, Rev, and Mom. T-statistics are in parentheses.

	Cross-sectional regression: One-month Lagged									
	(1) Low	(2) High	(3) Low	(4) High	(5) Low	(6) High	(7) Low	(8) High	(9) Low	(10) High
$IndSic3Ret$	0.036	0.010								
	(3.55)	(1.12)								
$FFi48Ret$			0.033	0.018						
			(2.64)	(2.08)						
$Tnic3Ret$					0.050	0.021				
					(4.22)	(2.52)				
$IORet$							0.050	0.009		
							(2.39)	(0.51)		
$CongloRet$									0.043	0.015
									(2.17)	(0.91)
Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$Adj.R^2$	9.69%	5.65%	11.21%	5.66%	9.13%	6.76%	10.76%	7.17%	14.51%	9.01%
$Avg\,base\,stocks$	1100	1100	1104	1104	1157	1157	807	807	318	318

Table 3.5: Long-run Performance of NBPs in Different Horizons

Panel A reports the coefficients of cross-sectional regressions by various horizons. $NbpRet$ is computed based on the one-month fraction-weighted stock return using all NBPs of a base firm. $NonNbpRet$ is the value-weighted return of those SIC peers that are not identified as NBPs in NBP-augmented SIC scheme. All the explanatory variables in the regressions are standardized. Control variables include $Size$, BTM, Rev, and Mom. Panel B reports the average excess returns (denoted XRET) and Fama-French five factor alphas (denoted as 5F) of Calendar-time portfolio analysis over different holding periods. Specifically, I employ two six-month holding periods: Month $t+2$ to $t+7$ and Month $t+8$ to $t+13$. T-statistics are in parentheses.

Panel A: Cross-sectional Regressions: Different Holding Periods					
	$NbpRet$	$NonNbpRet$	**Control**	$Adj.R^2$	$Avg\,base\,stocks$
Two months	0.014 (2.90)	0.010 (1.71)	Yes	5.97%	1090
Three months	0.013 (2.96)	-0.000 (-0.06)	Yes	5.74%	1075
Four months	0.008 (1.65)	-0.002 (-0.28)	Yes	5.65%	1059

	Panel B: Calendar-Time Portfolio Analysis					
	Month t+1		**Month t+2 to t+7**		**Month t+8 to t+13**	
	XRET	**5F**	**XRET**	**5F**	**XRET**	**5F**
Port1 **Low**	0.13% (0.29)	-0.60% (-2.66)	-0.45% (-2.00)	-0.43% (-1.98)	-0.27% (-1.28)	-0.13% (-0.57)
Port5 **High**	1.19% (2.89)	0.54% (2.77)	-0.07% (-0.36)	-0.11% (-0.58)	-0.08% (-0.41)	0.09% (0.45)
High - Low	1.06% (3.22)	1.14% (3.34)	0.39% (2.53)	0.32% (2.02)	0.19% (1.33)	0.22% (1.47)

Table 3.6: Bivariate Portfolio Analysis

Table 3.6 reports the one-month predictive return of lowest and highest NBP return quintile conditional upon high/low investor attention. NBPs' return is computed by aggregating all NBPs' lagged monthly returns for each base firm on a fraction-weighted scheme. The three investor attention proxies are firm size *Size*, number of analyst coverage *AstCvg*, and institutional ownership *InstOwn*. In each month, I sort all stocks into two groups based on median size/analyst coverage of the prior month. In each group, I report the highest/the lowest quintile returns. XRET is the excess return over the risk-free rate. I include five different risk factor models to compute alphas: CAPM, Fama-French 3-factor model (3F), Fama-French-Carhart 4-factor model (4F), a liquidity-augmented Fama-French-Carhart 5-factor model (4F + PS), and Fama-French 5-factor model (5F). T-statistics are in parentheses.

		Predictive Abnormal Return ($ARet_t$) of NBP (Ret_{t-1})						
		Quintile	**XRET**	**CAPM**	**3F**	**4F**	**4F+PS**	**5F**
Size	**<Median**	**Port1**	0.15%	-0.76%	-0.86%	-0.59%	-0.68%	-0.61%
		Low	(0.29)	(-2.51)	(-3.57)	(-2.77)	(-3.06)	(-2.46)
		Port5	1.32%	0.44%	0.32%	0.57%	0.58%	0.59%
		High	(2.66)	(1.47)	(1.42)	(2.83)	(2.77)	(2.62)
		High - Low	1.17%	1.20%	1.18%	1.16%	1.26%	1.20%
			(3.68)	(3.76)	(3.68)	(3.57)	(3.69)	(3.62)
	>Median	**Port1**	0.21%	-0.58%	-0.61%	-0.42%	-0.49%	-0.48%
		Low	(0.51)	(-2.70)	(-2.86)	(-2.11)	(-2.34)	(-2.16)
		Port5	1.03%	0.37%	0.35%	0.44%	0.45%	0.43%
		High	(3.00)	(2.14)	(2.03)	(2.59)	(2.52)	(2.42)
		High - Low	0.81%	0.95%	0.96%	0.86%	0.93%	0.91%
			(2.50)	(2.95)	(2.96)	(2.64)	(2.74)	(2.70)
AstCvg	**<Median**	**Port1**	0.05%	-0.92%	-1.02%	-0.78%	-0.79%	-0.77%
		Low	(0.10)	(-3.32)	(-4.30)	(-3.62)	(-3.64)	(-3.20)
		Port5	1.32%	0.47%	0.36%	0.55%	0.54%	0.58%
		High	(2.98)	(1.82)	(1.69)	(2.80)	(2.73)	(2.76)
		High - Low	1.27%	1.39%	1.38%	1.33%	1.33%	1.36%
			(3.73)	(4.08)	(4.02)	(3.84)	(3.81)	(3.85)
	>Median	**Port1**	0.39%	-0.43%	-0.47%	-0.31%	-0.31%	-0.42%
		Low	(0.95)	(-1.96)	(-2.16)	(-1.49)	(-1.47)	(-1.83)
		Port5	1.25%	0.52%	0.47%	0.62%	0.60%	0.52%
		High	(3.34)	(2.47)	(2.27)	(3.15)	(3.05)	(2.43)
		High - Low	0.86%	0.95%	0.94%	0.93%	0.91%	0.94%
			(2.53)	(2.78)	(2.73)	(2.68)	(2.60)	(2.62)
InstOwn	**<Median**	**Port1**	-0.01%	-0.89%	-0.94%	-0.69%	-0.66%	-0.63%
		Low	(-0.03)	(-3.03)	(-3.59)	(-2.86)	(-2.75)	(-2.36)
		Port5	1.23%	0.42%	0.35%	0.60%	0.60%	0.68%
		High	(2.69)	(1.51)	(1.43)	(2.78)	(2.75)	(2.89)
		High - Low	1.25%	1.31%	1.28%	1.29%	1.26%	1.30%
			(3.49)	(3.65)	(3.56)	(3.54)	(3.44)	(3.51)
	>Median	**Port1**	0.35%	-0.47%	-0.57%	-0.42%	-0.45%	-0.60%
		Low	(0.82)	(-2.23)	(-2.95)	(0.00)	(-2.44)	(-2.99)
		Port5	1.11%	0.39%	0.31%	0.41%	0.41%	0.34%
		High	(2.92)	(1.96)	(1.80)	(2.46)	(2.42)	(1.89)
		High - Low	0.76%	0.87%	0.88%	0.83%	0.86%	0.93%
			(2.66)	(3.05)	(3.07)	(2.90)	(2.96)	(3.17)

Table 3.7: Economic Links

Table 3.7 presents the results from quarterly simultaneous and one-month lagged cross-sectional regressions as follows:

$$Var_{i,t} = \alpha + \beta * Var_{p,t} + \epsilon_{i,t} \quad (3.6)$$

$$Var_{i,t} = \alpha + \beta * Var_{p,t-1} + \epsilon_{i,t} \quad (3.7)$$

The dependent variable is the own-financial ratios of each base firm in year quarter t. The main independent variable is the NBPs' weighted financial ratios in year quarter t (or year quarter $t-1$) for two different regressions. All financial ratios are from the CRSP/COMPUSTAT universe matched with the sample. PB is price-to-book ratio, EVS is enterprise value-to-sales ratio, PE is prices-to-earnings ratio, $RNOA$ is return on net operating assets, ROE is return on equity, AT is Inverse of Asset turnover, LEV is Leverage ratio, $SALESGROWTH$ is One-year-ahead realized sales growth, $RDPERSALES$ is expense-to-sales ratio, and SUE is standardized unexpected earnings. Each column represents the type of financial ratios ultilised in the regression, details of which can be found in the Table A3. All variables are truncated at 1% and 99% percentiles and standardized before running the regression. Coefficient estimates, T-statistics, and the adjusted R-squared are reported.

Panel A: Simultaneous Cross-sectional Regressions: Valuation Multiples

	PB	**EVS**	**PE**	**RNOA**	**ROE**	**AT**	**LEV**	**SALESGROWTH**	**RDPERSALES**	**SUE**
NBP	0.262	0.522	0.100	0.316	0.122	0.600	0.255	0.177	0.603	0.053
	(21.00)	(47.50)	(8.09)	(41.70)	(18.15)	(62.84)	(20.75)	(16.70)	(60.45)	(7.15)
Adj.R^2	6.91%	26.23%	1.24%	10.11%	1.42%	36.06%	5.97%	3.27%	35.51%	0.51%
Avg base stocks	564	564	564	564	563	566	510	475	171	534

Panel B: One-quarter lagged Cross-sectional Regressions: Valuation Multiples

	PB	**EVS**	**PE**	**RNOA**	**ROE**	**AT**	**LEV**	**SALESGROWTH**	**RDPERSALES**	**SUE**
NBP	0.266	0.528	0.127	0.334	0.120	0.605	0.258	0.214	0.618	0.059
	(19.67)	(60.35)	(10.54)	(31.43)	(5.37)	(64.98)	(21.24)	(19.55)	(70.91)	(7.26)
Adj.R^2	7.91%	27.59%	1.77%	11.69%	2.57%	37.01%	6.25%	4.58%	36.71%	0.54%
Avg base stocks	486	485	487	487	485	488	439	411	152	463

Table 3.8: NBPs with Alternative Constructions

Table 3.8 reports the monthly cross-sectional regression results for one-month predictive returns between base and peer firms using the NBP-augmented SIC scheme with alternative constructions based on the formula as follows. In Panel A, NBPs are constructed using monthly-rolling-forward schemes with a trailing window of 3, 6, and 9 months, respectively. In Panel B, NBPs are constructed using abnormal news co-coverage.

$$Ret_{i,t} = \alpha + \beta_1 * NbpRet_{i,t-1} + \beta_2 * NonNbpRet_{i,t-1} + \beta_3 * X_{i,t-1} + \epsilon_{i,t} \tag{3.8}$$

The dependent variable is the current own-return of each base firm. The main independent variables are the NBPs' returns and Non-NBPs' returns in month $t-1$, respectively. Observations are required to be in the Center for Research in Security Prices (CRSP), Compustat, the NBPs datatbase. One observation is 1 firm in 1 month. *NbpRet* is computed based on the one-month fraction-weighted stock return using all NBPs of a base firm. *NonNbpRet* is the value-weighted return of those SIC peers that are not identified as NBPs in the NBP-augmented SIC scheme. All the explanatory variables in the regressions are standardized. Control variables include *Size*, *BTM*, *Rev*, and *Mom*. T-statistics are in parentheses.

Panel A: NBPs with Different Windows

Windows	*NbpRet*	*NonNbpRet*	Control	$Adj.R^2$	*Avg base stocks*
3	0.017 (1.96)	0.011 (0.84)	Yes	12.90%	186
6	0.031 (4.03)	0.007 (0.61)	Yes	10.32%	353
9	0.030 (4.38)	0.007 (0.93)	Yes	8.88%	497

Panel B: NBPs with Abnormal News Co-coverage

	NbpRet	*NonNbpRet*	Control	$Adj.R^2$	*Avg base stocks*
	0.026 (4.92)	0.017 (2.34)	Yes	8.37%	496

Panel C: NBPs with No Earnings Announcements

	NbpRet	*NonNbpRet*	Control	$Adj.R^2$	*Avg base stocks*
	0.025 (5.23)	0.017 (3.38)	Yes	6.35%	1102

Chapter 4

Tomorrow's Fish and Chip Paper? Slowly incorporated News and the Cross-section of Stock Returns[1]

4.1 Introduction

News articles usually have time-limited effects due to rapid information transmission. This intuition can be summed up by the well-known idiom that "today's news is tomorrow's fish and chip paper".[2] Although financial markets build on a continuous flow of news, the analogy does not appear to apply if I follow the existing body of news predictability studies. So far, the evidence has revealed that returns following news arrivals are somewhat predictable. The literature attributes such a predictable pattern to either investor underreaction or overreaction based

[1]An edited version of this chapter appears in the European Journal of Finance (Tao, Brooks, and Bell, 2020a).

[2]In the UK, fish and chips (a takeaway treat) was traditionally wrapped in newspaper in order to absorb grease. This demonstrated that a newspaper was only valuable for the news it carried on the day of publication.

on *ex post* measures.[3] By contrast and to the best of my knowledge, research into the *ex ante* interaction between the information embedded in news articles and the contemporaneous stock reaction has not yet been presented in the literature.

According to Calvet and Fisher (2007), stock returns are a function of news with heterogeneous degrees of persistence ranging from a few days to several months. One negative news event could impact on the stock return gradually and persistently whereas the influence of another might be quick and prominent. Consistent with this phenomenon, Da et al. (2014b) find that news flowing continuously in small amounts is likely to cause strong return momentum, whereas infrequent, dramatic news announcements do not. In this chapter, I take a unique perspective by studying *ex ante* instances when the news tone does not match with contemporaneous stock reactions.[4] The research question that is addressed in this chapter is: What is the impact on future stock returns when contemporaneous stock price reactions mismatch the tone of news flows?

My empirical design is inspired by previous studies. Lochstoer and Tetlock (2020) provide empirical results showing that cashflow news decomposed from well-known anomalies (e.g., value, size, profitability, investment and momentum) is the main determinant of firm-level returns. By using the Dow Jones Newswire Archive dataset, Wang et al. (2018) document a news momentum phenomenon whereby a firm's monthly aggregated news tone predicts its one-month-ahead firm-level returns. This suggests that previous returns did not capture news signals in a one-month window. To measure how quickly information is incorporated *ex ante* into stock prices, I design a sequential, double-sorted approach based on monthly stock returns

[3]For example, Chan (2003) concludes that investors underreact to public information and overreact to private information by measuring long-run stock return performance; Tetlock (2007) documents an investor overreaction pattern by constructing a VAR (Vector Autoregression) model. Other related studies include Steeley (2004), Tetlock et al. (2008), Garcia (2013), Ahmad, Han, Hutson, Kearney, and Liu (2016), Jiang et al. (2017) and Kräussl and Mirgorodskaya (2017).

[4]News tone refers to the sentiment score of news content assessed using computational tools. A news article with positive news tone (or a high news sentiment score) tends to contain good news. The terms "news tone" and "news sentiment score" are used interchangeably hereafter.

and aggregate news sentiment scores. If there is a mismatch between the two, it is likely that the market has not yet fully priced the news flow and a persistent shock would be expected.

I classify news items in terms of whether they convey information quickly or slowly to the market. For example, a stock which has positive stock-related news stories accompanied by modest or negative returns might suggest that its price has incorporated information slowly. Thus, a central hypothesis in this chapter is that future return predictability tends to be stronger among slowly incorporated (SI hereafter) news stocks than those for which news is quickly incorporated (QI) in a given time period.

I perform the following analysis: a sentiment score is assigned to each news article by employing the Loughran and McDonald (2011) dictionary method.[5] SI and QI news are defined in two ways. Specifically, I construct monthly double-sorted portfolios by arranging stocks into terciles based on their current returns and further sorting each return group into another tercile based on their aggregate news sentiment scores. Among all nine groups, the four corner portfolios are studied as my primary research question focuses on those stocks where past returns and news tone are matched and those where they are not matched. High news sentiment scores accompanying low stock returns (LRHS) and low news sentiment scores accompanying high stock returns (HRLS) are defined as SI news, whereas news sentiment scores with matching stock returns (i.e., HRHS and LRLS) are referred to as QI news.

Next, I examine the post-formation return performance of these different types of portfolios. To do so, I compute equally-weighted average portfolio returns in the following month. A long-short portfolio consisting of buying LRHS stocks and

[5]Under this method, each word in a news item is labelled as belonging to a pre-specified category, examples of which include "positive", "negative", "model-weak" and "litigious". This approach provides a means of quantifying business news tone.

selling HRLS stocks generates abnormal future returns even after controlling for well-known risk factors. In sharp contrast, the other strategy of buying HRHS stocks and selling LRLS stocks earns negative future returns. The abnormal returns are robust to various alternative specifications. By including quintile portfolios, splitting into different sample periods, using a weekly frequency of analysis and employing a different measure of news sentiment scores, the results are quantitatively similar, suggesting that these SI news effects are unlikely to arise from data mining or biased measures.

To further understand what drives SI news effects, I propose two hypotheses. First, one explanation follows limited attention theory, suggesting that investors tend to narrow their focus to a few stocks instead of dispersing their selection evenly throughout the entire stock universe (see Hirshleifer et al. (2009)). To be more specific, these limited-attention investors, especially retail investors, tend to concentrate on certain stocks such as large, high media coverage, high trading volume stocks and firms with a high level of analyst coverage. Presumably, these firms have a better information environment where investors face less information asymmetry and therefore it is more efficient for them to react to news. In contrast, small and less information-rich firms are less likely to be under the spotlight, causing delayed reactions to news by investors.

Second, the speed of incorporation could result from the different nature of the news item, namely, dependent on its complexity and informativeness. Prior literature has studied the relationship between textual complexity and corresponding market reactions (Loughran and McDonald, 2014; Lawrence, 2013). For example, Loughran and McDonald (2014) document that the readability of financial disclosures has a significant impact on post-filing date returns, analyst dispersion and standardized unexpected earnings (SUE). Umar (2020) finds that long news headlines lead to a 40-basis point return underreaction on the *Seeking Alpha* forum.

Thus, it is likely that SI news conveys information that is challenging to interpret for less sophisticated investors. Moreover, the literature also argues that different news articles exhibit different features regarding the information they contain. You, Zhang, and Zhang (2017) show that market-oriented media tend to be more comprehensive in reporting corporate events compared to state-controlled media. A plausible explanation for the differential cross-sectional returns between SI and QI news might be that it is driven by the complexity of the news context. Investors tend to react quickly to more informative news articles but under-react to those that they find harder to interpret (see Dougal, Engelberg, Garcia, and Parsons (2012)).

To test the two hypotheses, I examine whether SI news is concentrated among firms with less attention, proxied by small size, low media coverage and low trading volume. I find consistent evidence that SI news has stronger effects among low-attention firms. Furthermore, I investigate limited-attention theory by applying two direct attention proxies: the Google Search Volume Index (SVI) and Bloomberg News Reading Activity Index (Abnormal Institutional Attention, AIA). As conventional attention proxies cannot guarantee that investors are actually reading firm-specific news (albeit they have high coverage), the literature utilises these alternative measures in order to better capture investor attention (Engelberg and Gao, 2011). My results suggest that SI news with low investor attention, particularly without retail investor attention, predicts stronger stock returns. In contrast, I find mixed evidence for cross-sectional differences between the effects of news as measured by textual complexity and by informativeness features. While SI news is not exclusively earnings-irrelevant and too hard to read (i.e. low readability), I do find that these news articles tend to be less accurate (proxied by business uncertainty tone) and less comprehensive (measured by article length). Collectively, these results do not systemically support the notion that SI news is more complex and less informative.

This chapter contributes to several aspects of the literature: First, I propose a novel *ex ante* approach to categorize news articles as to whether they are slowly incorporated or quickly incorporated into prices. Predictable return patterns are more likely to emerge when *ex ante* news sentiment scores and contemporaneous returns are mismatched. Thus, my research complements studies showing that a strong return momentum exists only when news flow arrives continuously in small amounts rather than discretely in large amounts (see Da et al. (2014b)). Second, the results support limited-attention theory, according to which certain stocks remain below the horizon on the investor's radar and therefore they react to stock-related news slowly (e.g., Da et al. (2014b) and Fang and Peress (2009)). By using a variety of investor attention indicators, I find consistent evidence that slowly incorporated news has stronger effects among low-attention firms. Third, the chapter also contributes to research on textual complexity and informativeness. The results do not, however, support the notion that SI news is more complex and less informative, based on a wide range of analysis that I conduct using readability and business uncertainty measures. Finally, the findings also have implications for practitioners. The refined predictive model based on both news and contemporaneous returns provides a potential avenue for a news-based trading strategy, which is viable even after allowing for estimated transaction costs.

The rest of the chapter is organized as follows. The related literature and hypotheses development will be discussed in Section 4.2. Section 4.3 presents the news data collection process and how I define SI and QI news. Section 4.4 reports the empirical results and empirical designs for two potential explanations of the findings. Three Appendices cover the details of the data construction and additional empirical results at the end of this thesis. In the final section, I present the conclusions and outline further directions for research.

4.2 Related Literature and Hypotheses Development

This chapter is related to several existing strands of the literature. First, this research is relevant to the topic of investor attention. The literature argues that it is either limited-attention theory or prominence theory that determines the impact of news on stock prices. Limited-attention theory, as suggested by Fang and Peress (2009), proposes that a firm with a lack of media coverage will earn higher returns than others, indicating that no attention can lead to stock price rises. In contrast, prominence theory embodies the value of visibility – in other words, that more media coverage leads to higher asset returns – for example, Hillert et al. (2014) find that the media makes momentum profitability. They attribute this to an overconfidence-driven overreaction where investors tend to be attracted by a news story with high coverage and therefore become excessively optimistic. Consequently, I am able to test the implication of limited-attention theory that a firm within an inferior information environment and with less attention would tend to have news that is reacted to slowly and thus achieves better stock returns.

Second, the readability literature is also relevant to my study. Accounting research documents that the complexity of corporate releases has certain impacts on post-filing stock returns (Lawrence, 2013; Dougal et al., 2012). A major measure of document complexity is the "Fog Index" – which is defined as a linear combination of average sentence length and the percentage of complex words. Loughran and McDonald (2014) argue that the complexity of 10-Ks certainly affects investor reaction, as simple and concise materials take investors and analysts much less time to digest and to determine the valuation-relevant information.[6] Naturally, news materials can exhibit similar effects (Umar, 2020). Underpinned by this strand of the literature,

[6] A 10-K is a comprehensive corporate filing by a US publicly-listed company about its annual financial performance and is required by the Securities and Exchange Commission (SEC).

I investigate whether the slow information incorporation that I identify is largely attributable to document complexity.

This research is also inspired by the link between textual analysis and investor decision making – see (Tetlock, 2007; Tetlock et al., 2008; Garcia, 2013; Ahmad et al., 2016; Jiang et al., 2017; Caporale, Spagnolo, and Spagnolo, 2018; Tao et al., 2020b). Tetlock et al. (2008) documents that tone can predict earnings surprises and daily stock returns. Tao et al. (2020b) show that investors' gambling behaviour is attenuated when maximum daily return events are associated with public news announcements. Ahmad et al. (2016) select 20 big firms with consecutive daily news and categorizes this news into being either informative or noise. They show that informative news can predict persistent future returns whereas "noise news" forecasts a subsequent reversal. Moving to the intra-day frequency, the effects of media tone become fairly straightforward (i.e., good news predicts positive asset returns while bad news predicts negative returns) (Jiang et al., 2017). My study adds to this literature, as I show that SI news can forecast asset returns.

Based on the studies discussed above, I develop three hypotheses that are then tested in this chapter. It has been widely accepted that both investor underreaction and overreaction can cause predictable return patterns following measurable news. Chan (2003) documents that investors underreact to public information (i.e., measurable DJNS news) and overreact to private information while Tetlock et al. (2008) find that stock prices will have positive drifts after good news and become negative when the news is bad. Wang et al. (2018) show that stocks in the highest news sentiment score portfolio will typically continue to have high news sentiment scores in the next month and vice versa. Their results further show that future returns will be high (low) for those stocks from the highest (lowest) news sentiment score portfolio. On the other hand, Tetlock (2007) documents an investor overreaction pattern: the tone of the *Abreast Of The Market Column* predicts positive stock index price movements on the next trading day and predicts them to be negative on subsequent days, suggesting a return reversal. Applying these findings I infer that investors will underreact to these measurable news stories when contemporaneous stock reactions do not rapidly match with the news tone. As a result, one can expect a delayed reaction in the following months. On the other hand, investors might overreact to these news items when both news sentiment scores and stock returns are in the top (bottom) bins.

Collectively, the three hypotheses tested in this chapter are as follows:

Hypothesis 1: SI news tends to lead to high future returns and QI news tends to produce low future returns.

Hypothesis 2: The differential cross-sectional returns between SI and QI news portfolios are driven by investor underreaction to news.

Hypothesis 3: The differential cross-sectional returns between SI and QI news portfolios are driven by the complexity of the news.

4.3 Data and Methodology

4.3.1 News Collection and Sentiment Analysis

My primary data comprise news taken from the Dow Jones Newswire Archive.[7] It contains all of the news from the Dow Jones newswire and *The Wall Street Journal* newspapers. To examine SI and QI news and their cross-sectional return patterns,

[7]The Dow Jones Newswire is a global real-time news product and is used in many research papers – e.g., (Tetlock, 2010, 2011; Engelberg et al., 2012, 2018).

I construct a sample of firm-level news released by the common US stocks listed on the NYSE, AMEX and NASDAQ between 1979 and 2016. The details of the data collection procedure can be found in Appendix B.2.

As can be seen in Panel A of Table 4.1, firms have a wide range of numbers of news items in each month: the lowest figure is only one whereas the highest is 1568. On average, each firm has around 13 news items every month. This suggests the distribution of the number of news items is highly skewed. I next examine this issue from both time-series and cross-sectional perspectives.

Concerning the time-series pattern of firm-level news, I first examine the ratio using the number of matched Dow Jones News firms against the total number of firms from the CRSP universe in each year. During the early period (1979 to 1995), the percentage of firms covered by measurable Dow Jones news reports is between 20% and 30%. However, this statistic increases rapidly to over 95% after 1995. As a result, it reveals that some missing firms are clustered, probably due to the underdeveloped and incomplete archive in the early years. To further allay readers' potential concerns that the baseline results might be significantly influenced by these unmatched firms in the early years, I perform a subsample analysis and the results remain qualitatively unaltered across the four subperiods in Appendix E.

To gain an understanding of the cross-sectional patterns in the data, I also plot a histogram of the news observations. The distribution of the number of news items is highly skewed. As can be seen in Figure 4.1, the percentage of firms with larger numbers of news items is small. To be specific, there are 4492 firms with 1 to 100 news items, accounting for 32% of the total sample. This statistic declines to 15% for firms with 100-200 news items and only 8% for those with 200-300 news items. Collectively, it can be concluded that most news items belong to a small number of firms. This pattern is consistent with Tetlock et al. (2008), who find that large firms tend to have news coverage every day whereas small firms only have sparse observations. One implication of this property is that the baseline results might be driven by the small size effect. However, further robustness checks eliminate this concern, as I remove stocks whose prices are below $5 per share and the results remain positive. I also incorporate the size factor into the Fama-French factor models and control for firm size in the Fama-MacBeth regression, respectively.

In addition, I perform robustness checks by collecting firm-level news with more strict conditions to increase its relevance. For example, there might be a difference in terms of the likely strength of any stock price reaction between an article analysing the cell phone industry and a news article addressing iPhone products specifically. In Appendix B.2, I show that the use of a stricter firm-level news criterion does not qualitatively alter the results.

——————————— Insert Figure 4.1 here ———————————

To quantify the sentiment from firm-specific news, I employ the Loughran and McDonald dictionary method as this approach builds on the work of Henry (2008).[8] This sentiment analysis method is a reliable and popular technique as it is tailored specifically to finance applications (e.g., see Loughran and McDonald (2011) and Loughran and McDonald (2015)). For instance, the word "abandoned" is pre-assigned as a negative sentiment word in this dictionary. Any document which contains the word "abandoned" will be counted and increases the negative percentage. To show that the results are not sensitive to different sentiment tools, I later employ the Google Natural Language API as an alternative.[9]

[8]Although Henry (2008) is the first work that contributes to sentiment analysis, I do not employ his word list developed by therein for two reasons: First, Henry's list only has a very limited number of sentiment words (e.g., 85 negative words) whereas that of Loughran and McDonald (2011) includes 2329 such words. The full list can be downloaded in `https://drive.google.com/file/d/15UPaF2xJLSVz8DYuphierz67trCxFLcl/view`. More importantly, the most frequent Loughran and McDonald negative words based on 10-K annual reports such as *loss, losses, claims, impairment, against, adverse, restated, adversely, restructuring and litigation*, do not appear in Henry's list. This suggests that Henry's dictionary may not sufficiently capture all potential negative tone from the text.

[9]The Google Natural Language API is a newly-built natural language processing tool by Google Cloud, the details of which can be found at https://cloud.google.com/natural-language/.

The first step is to construct an article-level sentiment score. For each news story, I adapt the method proposed by Garcia (2013), constructing such a score by taking the number of positive words minus negative words divided by the total number of words. In the second step, I standardize these news sentiment scores at the stock-level by subtracting the rolling mean of the previous 12 months and dividing by the standard deviation of the same period for each firm. In the following paragraphs, NSS denotes the standardized news sentiment score. The final step is aggregation. To fit the score into the empirical setting (i.e., monthly portfolio analysis), I aggregate these NSS every month. $NSS_{i,t}$ represents the aggregate news sentiment score standardized in a time-series manner for firm i in month t:

$$nss = \frac{No.\,of\,pos - No.\,of\,neg}{total\,number\,of\,words} \tag{4.1}$$

$$NSS_{i,t} = \frac{nss_{i,t} - \mu_{nss,i,t}}{\sigma_{nss,i,t}} \tag{4.2}$$

where $\mu_{nss,i,t}$ is the rolling mean of the previous 12 months' news sentiment scores and $\sigma_{nss,i,t}$ is the rolling standard deviation of NSS during the same period for each firm. As can be seen in Panel A of Table 4.1, $NSS_{i,t}$ exhibits a fairly even distribution. The average values are close to zero with no extreme outliers, suggesting that the procedure for constructing $NSS_{i,t}$ is effective in capturing unexpected components.[10]

4.3.2 Slowly and Quickly Incorporated News

The finding in Wang et al. (2018) suggests that previous returns did not capture news signals in a one-month window. For the data studied in this chapter,

[10] In an unreported test, I confirm its even distribution by performing the Jarque-Bera test suggested by the Editor.

the correlation between contemporaneous monthly returns and the news sentiment score is also very low (0.066) on a monthly basis. It is perhaps surprising but consistent with Wang et al. (2018) that initial market responses to the tone of news articles are insufficiently large. Otherwise, I should observe that the market responds favourably to the news tone (i.e., a highly positive correlation between the two sorting variables). Collectively, I use a sequential double-sorted approach based on monthly stock returns and news sentiment scores to measure how quickly information is incorporated. If returns cannot match news sentiment, it can be accepted that stock prices did not catch the news signal promptly and a SI news item is defined.

To ensure that the stock returns I utilise are firm-specific, I calculate DGTW-adjusted returns following Daniel, Grinblatt, Titman, and Wermers (1997) to remove the expected return components of common risk factors including SMB (Small-Minus-Big), HML (High-Minus-Low), UMD (Up-Minus-Down). Specifically, I independently sort the entire stock universe into quintile portfolios based on firm size, industry-adjusted book-to-market ratio and industry-adjusted momentum. Next, I compute these 125 (i.e. $5 \times 5 \times 5$) value-weighted benchmark returns. The cleaned return is each stock's raw return minus the corresponding portfolio benchmark return, which can be interpreted as an abnormal return.

The double-sorted portfolios are constructed as follows: in each month, all stocks are divided into tercile portfolios based on contemporaneous DGTW-adjusted returns. Low return, medium return and high return subsets are defined as LR, MR and HR, respectively. I then form another three subsets in each return group individually based on stock-level news sentiment scores (i.e., LS, MS and HS refer to low, medium and high news sentiment subsets), resulting in nine portfolios having different stock return and news sentiment characteristics. Next, I label each portfolio given its news return characteristics. For example, the $HRLS$ portfolio comprises stocks with high contemporaneous stock returns and low news sentiment scores. Similarly, I define the $LRHS$ and $HRLS$ portfolios as SI news groups and the $HRHS$ and $LRLS$ portfolios as QI news groups.

In Panel B of Table 4.1, I report univariate analysis for the four types of news. The $LRHS$, $HRLS$, $LRLS$ and $HRHS$ portfolios are presented in each column. Specifically, both news sentiment scores and stock returns statistics exhibit considerable differences across the groups. For example, the $LRHS$ portfolio has an average stock return of -12% per month and an average news sentiment score of 0.94, which suggests that stocks in this portfolio tend to have negative returns and positive news. In other words, $LRHS$ stocks are slowly incorporating good news. Consistent with this notion, I then find that the $HRLS$ portfolio exhibits negative news (the news sentiment score is -0.96) and positive return performance (average 15% per month), while the figures are: -1.17 (news sentiment) and -15% (monthly return) for the $LRLS$ portfolio and those of $HRHS$ are 1.02 and 15% respectively.

Moving to news volume, $LRHS$ stocks release relatively fewer news articles compared to the other three counterparts (an average of 8.68 for $LRHS$ stocks, 11.75 for $HRLS$ stocks, 10.87 for $HRHS$ stocks and 11.71 for $LRLS$ stocks). The increasing number of news items in the following month can be observed for all four types of stock. This can be interpreted as suggesting that news stocks are likely to have a continuous information flow, especially SI news stocks. Readers may be concerned that these next month news items could more heavily drive their contemporaneous stock returns than the news released in previous periods. However, in later robustness checks, I show that this is not the case by dropping all news observations in the portfolio holding periods. Lastly, the average number of stocks in each month for all four portfolios is reported in Panel B. As can be seen, each portfolio contains roughly 240 stocks per month, accounting for around 11% of all stocks. If all of these portfolios are aggregated, the total percentage of stocks covered is 44%.

Overall, I conclude that the sequential double-sorted approach effectively separates stocks into portfolios based on different types of news.

——————————— Insert Table 4.1 here ———————————

4.4 Empirical Results

4.4.1 Calendar-time Portfolio Analysis

I use a sequential double-sorted calendar-time portfolio approach to examine SI and QI news return predictability. In each month from July 1979 to December 2016, I divide all stocks into tercile portfolios based on their contemporaneous monthly DGTW-adjusted stock returns. The LR, MR and HR groups contain stocks with low returns, medium returns and high returns, respectively. For each return portfolio, I further rank stocks into another three portfolios based on news sentiment scores. LS, MS and HS therefore collect stocks with low, medium and high new sentiment scores. Subsequently, I track the performance of each subset over the following month by computing their equally-weighted stock returns. By rolling this monthly window through the entire sample period, I obtain a time-series of return performance for all nine portfolios. In Panel A of Table 4.2, it can be observed that news predictability decreases across different return groups. For example, the LR and HS portfolios predict a 1.45% return per month in the post-formation period whereas the LR and LS conjunction only achieves 1.05% every month. Similarly, it is also evident that the predictive power of return performance decreases for each news group from LR to HR.

I further study SI and QI news effects in Panel B. If investors do indeed react slowly to arriving news, I should observe that slowly incorporated news predicts stronger post-formation stock returns than quickly incorporated news.

Panel B of Table 4.2 supports this hypothesis: the SI news portfolio (i.e., the spread between $LRHS$ and $HRLS$) earns 101 basis points per month (t-statistic: 5.85) whereas the QI news portfolio (i.e., the spread between $HRHS$ and $LRLS$) has negative return predictability. The difference between the two, which I term SMQ (Slow-Minus-Quick), gains 139 bps per month, which is significant at the 1% level.

Further, I perform a Fama-French regression analysis of these results and I report the alpha for the two news groups. I include five different risk factor models: the Fama-French 3-factor model (FF3F), Fama-French-Carhart 4-factor model (FF4F), Fama-French 5-factor model (FF5F), a liquidity-augmented Fama-French-Carhart 5-factor model (FF4F + Liq) and a short-term reversal-augmented Fama-French-Carhart 5-factor model (FF4F + Rev).[11]. The reason for including liquidity and short-term reversal risk factors is to examine whether the constructed portfolio picks up these effects. However, the risk-adjusted returns of the SI news portfolios remain consistently significant across all models, even when the short-term return reversal factor is included. This result suggests that the SI news effect cannot simply be interpreted as a short-term return reversal. In contrast, the QI news portfolio return becomes insignificant after controlling for the Fama-French three- and five-factors. The alpha of the SMQ portfolio is significant across risk-adjusted models, suggesting that SI news indeed has stronger return predictability compared to QI news. Since most of the long-short portfolio returns come from SI news rather than QI news, I next focus on SI news predictability instead of SMQ in the regression analysis.

[11]The Fama-French risk factors can be downloaded from Kenneth French's Data library `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/datalibrary:html` The liquidity factor I utilise is Pástor and Stambaugh (2003), which can be accessed from `https://faculty.chicagobooth.edu/lubos-pastor/data`.

——————————— Insert Table 4.2 here ———————————

4.4.2 Fama-MacBeth Regressions

The asset pricing literature identifies that a number of stock characteristics have stock return predictability. To ensure that the results I documented are not driven by those characteristics, I use Fama and MacBeth (1973) regression models to study the predictability of SI and QI news on the following months' stock returns after controlling for additional stock characteristics. The choice of this regression follows the convention of the asset pricing literature. Specifically, each month I perform an OLS regression of the dependent variable of interest. All regression coefficients are then averaged across different months and determine whether they differ from zero. The regression is as follows:

$$\begin{aligned} EXRet_{i,t+1} = \alpha + \beta_1 * NSS_{i,t} + \beta_2 * (NSS * SINws)_{i,t} + \beta_3 * (NSS * QINws)_{i,t} \\ + \beta_4 * SIGdNws_{i,t} + \beta_5 * SIBdNws_{i,t} + \beta_6 * QIGdNws_{i,t} \\ + \beta_7 * QIBdNws_{i,t} + X_{i,t} + \epsilon_{i,t} \end{aligned} \tag{4.3}$$

where $EXRet_{i,t+1}$ is the excess return of stock i at time *t+1* and $\epsilon_{i,t}$ is an error term that is assumed to be independently and identically distributed with zero mean and constant variances along the diagonal elements and zero elsewhere. NSS is the news sentiment scores of stock i at time *t*. $SIGdNws$, $SIBdNws$, $QIGdNws$ and $QIBdNws$ are dummy variables for whether a stock has a news article identified as slowly incorporated good news (LRHS), slowly incorporated bad news (HRLS), quickly incorporated good news (HRHS) or quickly incorporated bad news (LRLS). $NSS * SINws$ and $NSS * QINws$ are the interaction terms where news sentiment scores are multiplied by the $SINws$ or $QINws$ dummy variables. The literature and previous sections in this chapter imply positive β_2 and negative β_3 coefficients on the interaction variables $NSS * SINws$ and $NSS * QINws$, respectively. In the X vector, a number of control variables are added to capture various stock characteristics. Specifically, I first include $SIZE$ and BTM to control for the Fama and French (1993) characteristics. $LRET$ is the lagged one-month stock return to control for short-term reversals. Furthermore, MOM is added in order to capture momentum effects, which is computed by taking the past twelve-month cumulative return with at least eight months' valid observations. $BETA$ is included and computed by following Scholes and Williams (1977); Dimson (1979). I include $IVOL$ (Idiosyncratic volatility) in the regression motivated by Fu (2009), who argues that idiosyncratic volatility represents how fast firm-level information is incorporated into stock prices. I also add $ILLIQ$, a firm-level illiquidity proxy by Amihud (2002), which is computed by taking the average value of daily absolute stock returns divided by the dollar trading volume over the previous year.

Table 4.3 reports all primary variable coefficients from the three regression specifications. First, the NSS positively predicts future returns, indicating that the news articles are indeed informative. Second, the $SIBdNws$ and $NSS * SINws$ variables are statistically significant under all regression specifications. In particular, the coefficient of $NSS * SINws$ is 0.2004 (t-statistic=2.21) whereas the $NSS * QINws$ term is insignificant in model 6. The positive $SIBdNws$ coefficient suggests that the return continuation mainly comes from slowly incorporated bad news rather than good news occurring during the formation period. Comparing the magnitudes of NSS and $NSS * SINws$, the $NSS * SINws$ term is more than double that of NSS (0.2004 for $NSS * SINws$ vs 0.0959 for NSS). However, the coefficient of $NSS * QINws$ varies across different model specifications and I therefore cannot infer its stock return predictability. Collectively, these tests confirm that the primary news variable, $NSS * SINws$, has some stock return predictability and is robust to the incorporation of a number of control variables.

——————————— Insert Table 4.3 here ———————————

4.4.3 Performance Evaluation

I evaluate the performance of SI news over different horizons. If the news source contains genuine information about a firm's fundamentals, the return predictability of SI news should not subsequently reverse in the long run. In addition, the performance evaluation could potentially have useful implications for practitioners to develop a news-based trading strategy. To examine the long-run performance, I conduct a calendar-time portfolio approach over four different holding periods as follows: 3 months (month 2 to month 4), 6 months (month 2 to month 7), 9 months (month 2 to month 10) and 12 months (month 2 to month 13). The current month is included as a comparison. I again construct an SI news portfolio that buys stocks with low returns but high news sentiment (LRHS) and sells stocks with high returns but low news sentiment (HRLS) and a QI news portfolio comprising buying HRHS stocks and selling LRLS stocks. These portfolios are then rebalanced monthly and their performance monitored over different holding periods. The results, reported in Table 4.4, show that after the first-month period, SI returns are not significantly different from zero. It is evident that the return predictability of SI news does not reverse in the long run. Similarly, I can also observe that the slow-minus-quick (SMQ) factor gains significantly during the post-formation period but these portfolio returns do not reverse in the following months.

——————————— Insert Table 4.4 here ———————————

Graph A of Figure 4.2 shows a comparison of the indexed value of the portfolio with that of the S&P500 index. Overall, the portfolio generates over \$26, assuming

$1 of initial investment. In contrast, the S&P500 index only achieves $5.60. There are, however, two big recessions during the period covered -– namely, the Internet Bubble and the financial crisis of 2007/8. It can be seen that the long-short portfolio is riskier during times of market turbulence. Graph B of Figure 4.2 shows the 12-month holding period returns for the SI news portfolio and the S&P500 index. It can be observed that overall the portfolio performs well. In contrast, the S&P500 has experienced a number of negative return periods – for example, it dropped by 40% during the financial crisis. The holding period return series indicates that since 2010 profitability has declined for both the portfolio and the benchmark. Nevertheless, the results suggest that a news-based strategy may be attractive to practitioners.

In addition, I evaluate the impact of considering reasonable transaction costs on the news trading strategy's profitability. I calculate the Break-Even Trading Cost (BETC) of the strategy – the trading cost that makes the average realized returns of the strategy become zero following Han, Yang, and Zhou (2013). In other words, the higher the BETC, the more likely it is that the trading strategy will survive and remain profitable after considering actual transaction costs. Since the BETC is associated with the strategy's realized returns, I monitor the turnover of the portfolio and compute the BETC only when each stock drops out at the end of that month (i.e., the position of a stock is closed at the end of the month and is not included in the portfolio in the next month). The estimated BETC includes all necessary expenses such as commissions, exchange fees, bid-ask spreads, market impact costs and transaction taxes. Overall, the BETC of the SI news trading strategy is more than 100 bps per month, suggesting that the actual transaction costs would have to be unrealistically high to eliminate the trading profit. This estimated BETC substantially outperforms the transactions cost levels employed in prior studies that are usually set at a fixed rate or a pre-specified range (e.g., 25 bps – see Lynch and Balduzzi (2000) and 1 to 10 bps – see Tetlock et al. (2008)). Graphs

C and D of Figure 4.2 cross-validate that the trading strategy remains profitable after considering 25 and 50 bps transaction costs, respectively.

Moreover, I also compute an alternative BETC to penalize trading on small stocks that are of limited use for arbitrage due to their illiquidity and high bid-ask spreads. In each month, I sort all stocks into deciles based on their market capitalization so that small firms will be assigned to the lowest decile whereas large firms will be in the highest group. The lowest BETC for trading on the size decile is 58 bps on average per month, which is still higher than the transactions cost values assumed in existing studies (i.e., assuming a fixed cost of 50 bps).

——————————— Insert Figure 4.2 here ———————————

Overall, this subsection confirms that SI news does contain new information and indeed this is incorporated into the market at a later stage. The performance evaluation of the news-based trading strategy confirms its profitability and its robustness even after considering transaction costs.

4.4.4 Alternative Information Incorporation Measure

Another measure of news incorporation would be to examine the beta coefficients between contemporaneous stock returns and news sentiment scores. Inspired by the calculation of market beta in the CAPM model as in Frazzini and Pedersen (2014), I regress the daily news sentiment score on the corresponding stock return in a monthly rolling OLS regression model to obtain the coefficient. As such, it should improve measurement accuracy when the data are of daily frequency. The regression equation is:

$$Ret_{i,t} = \alpha + \beta * NSS_{i,t} + \epsilon_{i,t} \tag{4.4}$$

where $Ret_{i,t}$ is the daily return of stock i on date t and $NSS_{i,t}$ is the daily pessimism value of stock-level news for stock i on date t. This regression is then performed using a rolling monthly window. As a result, I obtain time-varying beta parameters between the stock return and news sentiment for each stock and calendar month. To address the concern of overnight news or weekend news, which may have lagged return impacts and therefore might not be captured by the regression model, I allow a lagged three-trading-day window to capture market lagged reactions. For example, Monday night news will not have return impacts until Tuesday morning. Friday night news will also not influence stock returns until the following Monday when the market opens. It is then believed that my time-varying beta should capture all these delayed news reactions. Moreover, I only include a stock in the sample when it has at least five different news stories on five different trading days to preclude the regression having sparse observations. As a result, a steep sloping coefficient β indicates news that is quickly incorporated, whereas a gentle slope implies slow information incorporation.

To trace the performance of different coefficients between the daily news sentiment score and daily contemporaneous stock returns, I employ a single-sort calendar-time portfolio approach with a one-month rolling window. Specifically, in each month, I divide stocks into three terciles based on the β coefficients (i.e., $LoBETA$, $MeBETA$ and $HiBETA$ portfolios). Naturally, the higher is the coefficient, the quicker the stock return responses. The top portfolio $HiBETA$ then contains high beta coefficients, suggesting that news is incorporated quickly into stock prices. The bottom group, $LoBETA$, comprises those stocks with low beta coefficients, which indicates slow information incorporation. Finally, equally-weighted portfolio returns are computed during the post-formation period for the three portfolios separately.

The result of this alternative definition is reported in Panel A of Table 4.5. The coefficient of the top portfolio (SINws group) is 1.28, which is significant at the

1% level, whereas the QINws portfolio achieves 80 bps per month (*t*-statistic: 2.27). The Slow-Minus-Quick (SMQ) portfolio reports an average 49 bps return in each month, which is significant at 1%. Compared to the SMQ portfolio constructed in the first way (94 bps per month), the overall profitability of SMQ created by news betas is halved. Presumably, the sample size is significantly reduced with a number of small stocks excluded due to the fact that all stocks included in the calculation must have at least five news stories published on five different days in a month. It is unlikely that small firms have more than five different news releases in any five-day period in a month. Although this implies that the SI news effect is partially contributed by the size effect, the SI news theory, on the other hand, seems to be more convincing under these two different designs.

4.4.5 Further Robustness Checks

I examine the robustness of the primary SI and QI news definitions. The sequential double-sorting of SI and QI news leads to a potential issue: since news sentiment scores and stock returns are positively correlated, the second sort on returns may create further variation in the next month's stock performance. As such, I perform an independent double-sort on news sentiment and stock returns in the first robustness test. The SI and QI news defined by the independent double-sort displays the same pattern as those in Table 4.2, with return predictability achieving 100 bps (t-statistic=5.80) for SI news and -40 bps for QI news.

A further potential concern regarding the findings is that the effect is largely driven by penny stocks. To avoid the bid-ask bounce effect and illiquidity, I exclude stocks whose prices are below $ 5 per share at the end of each particular formation month. As Panel B of Table 4.5 shows, the SI news effect still holds, albeit with reduced magnitudes.

One could argue that the predictability of SI news is due to a number of news articles arriving in the subsequent months. This might be true given that news volume increases over time (see Table 4.1). The SI news effect could be driven by news released in the holding period rather than caused by the news in the previous formation period. To address this concern, I eliminate all news observations occurring within the holding periods. As can be seen in Panel C, the SI news effects remain statistically significant.

In addition, I attempt to quantify news sentiment using different sentiment indicators, or even different techniques. To mitigate the concern that some news items contain high volumes of both positive and negative sentiment words and lead to low levels of pessimism (i.e. *Neg* - *Pos*), I also report the results measured by only employing either *Neg* or *Pos* indicators by the Loughran and McDonald (2011) dictionary method. Furthermore, I apply a different sentiment analysis tool. Earlier methodologies have been criticized in three main ways: First, the underlying bag-of-words model only detects single words as the leading judgment of sentiment analysis. It fails to take into account that some sentences contain "negating words" such as "doesn't", "don't", "can't", etc., which can change the sentiment of the entire sentence. Second, this model fails to consider "modifier words" such as "really", "too much", etc., as these words or phrases sometimes enhance the positive and negative tone. Third, without analysing grammatical structures, the Loughran and McDonald (2011) dictionary method finds it difficult to deal with part-of-speech tagging.[12] For example, the firm-specific news of company "Best buy" will be more positive than others given its firm name contains a positive sentiment word "Best". To address these pitfalls and the potential bias caused by them, I employ the Google Natural Language API as an alternative sentiment analysis. Overall, the results remain quantitatively similar.

As for any definition, the measures of SI and QI news are somewhat subjective.

[12]Part-of-speech tagging is a computational linguistic method to label the category of words as noun, adverb, adjective, etc.

For instance, I could examine the SI news effect with a one-week time interval instead of a one-month window with or without matching returns. In my sample, SI news predicts a 53 bps return on a weekly basis (equivalent to 2.21% per month), compared to 1.01% using monthly windows. QI news exhibits similar patterns (-0.79% per month). The rationale behind this finding is that the relevance of news is time-limited and its effects will fade away over longer horizons.

Finally, I compute the standardized unexpected earnings (SUE) and cumulative abnormal returns in a [t-1, t+1] window around each quarterly earnings announcement (CAR[-1,1]). The rationale for doing this is that the CAR effectively reflects the actual market reactions to the corresponding firm-level news (i.e., earnings announcements). As such, a stock with a positive SUE and a positive CAR means that the news tone and stock return are matched and vice versa. For each month, I first sort firms into high CAR and low CAR bins. Within each CAR bin, I classify all firms' quarterly earnings into positive and negative SUE bins. A stock with positive SUE and low CAR (or negative SUE and high CAR), if generating a low future return, could be interpreted as a slowly incorporated (SI) news stock. In contrast, a stock with positive SUE and CAR (or negative SUE and CAR), if followed by a high future return, is likely to be a quickly incorporated (QI) news stock. I repeat the baseline analysis reported in Table 4.2. For those stocks with a positive SUE and a low CAR (or negative SUE and high CAR), I again categorize nine groups and take four corner portfolios to construct the SI and QI portfolios. To evaluate their performance, I report the average excess returns. As can be seen, a 1.59% monthly average return is reported for the SI news portfolio whereas that of the QI news portfolio is close to zero. The combined portfolio of buying SI stocks and selling QI stocks also generates 1.58% per month, with a t-statistic of 2.20. In contrast, stocks with positive SUE and CAR (or negative SUE and CAR) fail to observe any predictable return pattern as expected. All such portfolios' return performances are insignificantly different from zero. Taken together, the results clearly show that the original definition of SI news is robust and can survive in an alternative setting.

Collectively, the above robustness checks of the SI news effect should eliminate concerns of data mining or biased measures. In the next section, I study the time-variation of SI news effects within the sample.

——————————— Insert Table 4.5 here ———————————

4.4.6 Subsample Analysis

Technical advances have facilitated information dissemination, therefore reducing the gap between the times when information released and when it is received by investors. If SI news is caused by information that travels slowly, I should observe that the predictive power of SI news has gradually diminished over time. To test this conjecture, I evenly split the full sample into four different periods, the first two of which might be termed pre-Internet and the later two are post-Internet. Within each period, I perform the same analysis as previously discussed above. In Table 4.6, SI news significantly predicts the next month's stock returns during all subsample periods (93 bps during 1979 to 1987, 108 bps between 1988 and 1997, 132 bps from 1998 to 2007 and 62 bps between 2008 and 2016, which are all significant). Interestingly, the largest magnitude of SI news portfolio returns emerges during the 1998-2007 period where the number of news items dramatically increases. During the crisis period (i.e., 2008-2016), I observe a reduced SI news effect.

Given the relatively stable performance of SI news portfolios across each subsample, it seems implausible to conclude that SI news is caused by the slow information diffusion, at least not for the post-Internet period.

——————————— Insert Table 4.6 here ———————————

4.4.7 Limited-attention Theory

Since these news portfolios exhibit cross-sectional differences in future return performance, I ask what drives this cross-sectional difference of stock returns. One explanation for market-delayed reactions is limited-attention theory. The growing literature of this strand of research includes Engelberg and Gao (2011) on the volume of Google searches for stock names, Da et al. (2014b) on continuous information and momentum effects and Fang and Peress (2009) on non-news stock returns. The theory predicts that investors underreact regarding stocks with low volumes of attention and thus the subsequent return predictability is the result of this delayed reaction. Intuitively, stocks with small size, low media coverage, low turnover and low analyst coverage are generally associated with a poor information environment. In other words, it is information asymmetry that slows investors' perceptions of arriving news. If so, I would expect that SI news effects are concentrated in low attention-grabbing stocks whereas QI news stocks would receive a relatively high volume of attention.

In this section, I employ four different investor attention proxies to study SI news effects following the previous literature (in particular (Da et al., 2014b; Huang, 2018)), which includes size, media coverage, trading volume and analyst coverage. I first examine the cross-sectional difference between small and large firms. To do so, I perform a double-sort and construct calendar-time portfolios. Specifically, I rank the firm's market capitalisation values and divide these firms into small and large size groups before constructing SI and QI news portfolios. Market capitalization ($SIZE$) is in logarithmic form and rebalanced at the end of June. The next-month stock return performance is traced for each month and the results are reported in Panel A of Table 4.7. This shows that the returns from small firms are higher than those for large firms with the difference being significant at the 1% level, suggesting that SI news effects are concentrated in small firms, a finding consistent with limited-attention theory. There is no evidence of cross-sectional differences between small and big firms with QI news, although the SMQ factor also confirms the SI news effect.

Changes of media coverage (Δ media coverage hereafter) are often applied as the representation of investors' attention. Stocks with higher Δ media coverage can naturally diffuse information much quicker than lower Δ media coverage stocks. Thus, SI news is more likely to be observed among low-attention stocks and QI news concentrated in high-attention stocks. Empirically, Panel A supports this limited-attention theory with the monthly Δ media coverage shifting from low to high, SI news predicts next-month returns from 144 basis points per month to only 9 bps. The difference between the two coverage groups is significant at the 1% level. The SMQ factor also suggests that information diffusion is likely to be slower within the low media volume portfolio (2.21% with a t-statistic of 7.56 for the cross-sectional difference between low and high media coverage groups), which is consistent with limited attention theory.

Panel B also tests the hypothesis regarding whether stocks with lower trading volume predict higher SI news profitability. The literature documents that trading volume is a proxy for investor attention (Barber and Odean, 2007). I again construct a double-sorted calendar-time portfolio by employing stock-level trading volume. The variable is calculated as the natural log of average share volume divided by the number of shares outstanding over each month using daily data. I then trace the portfolio performance by a rolling one-month window to observe SI news effects within either high or low trading volume groups. In Table 4.7, I report all coefficients with T-statistics for each subset of stocks, finding that SI news stocks earn 1.36% per month (t-statistic=8.94). The difference in profitability between low and high turnover SI news portfolios is statistically significant at the 1% level. This is consistent with the theory that stocks without active trading activity are more likely to be under the investor's attention radar. Overall, comparing low and high turnover profits, it can be concluded that low trading volume stocks are associated with higher SI news profitability.

Finally, I examine how analyst coverage is associated with cross-sectional differences in SI news predictability. The literature uses analyst coverage as a proxy for investor attention (Hirshleifer and Teoh, 2003). If investor inattention slows information incorporation from news articles, the predictability should be concentrated among firms with low analyst coverage. The analyst coverage data is retrieved from Datastream for one-year-forward earnings per share forecasts for each firm. The sample is then divided into two groups based on the median number of analysts covering the stock, for each of the SI and QI news stock portfolios that have been constructed in each month. In Panel A, the performance of each portfolio is reported. Stocks with low analyst coverage have higher levels of SI news compared to those with high coverage – 1.03% and 0.57% per month, respectively. The *t*-statistic of the low-minus-high spread portfolio is 1.19, which is significant at the 5% level. In contrast, there is no evidence of QI news effects in either the low or high coverage subsamples.

To summarize, four different investor attention proxies consistently support limited-attention theory, which indicates that SI news predictability is positively associated with low investor attention.

——— Insert Table 4.7 here ———

4.4.8 Complexity and Informativeness

SI news effects can also potentially be explained by news characteristics. The literature identifies that variation in news content does affect market reaction (Dougal et al., 2012; Umar, 2020). For instance, (especially retail) investors might find a complex news article challenging, in the sense that they might be uncertain about how to interpret the news content. Likewise, an ambiguous news tone could also fail to give investors a clear trading signal, which therefore discourages them form quickly responding. On the other hand, investors might under-react to a short news article because it means a less comprehensive report about a corporate event. They are also more likely to be interested in earnings-related news since these news items contain more valuable information (Tetlock et al., 2008; Chen, De, Hu, and Hwang, 2014). If SI and QI news is driven by the nature of the news in this setting, one would expect that easily-understandable and more informative news predicts lower future stock returns. In other words, investors will react quickly to news stories with fewer ambiguous words, fewer complex sentences, more comprehensive content and more earnings-related topics.

To capture the textual complexity of the news documents, I apply the fog index and weak modal words measures. The fog index is one of the most commonly applied readability measures for a firm's financial disclosures (Dougal et al., 2012; Lawrence, 2013; Loughran and McDonald, 2014). It captures document readability by computing the average sentence length and the percentage of complex vocabularies. Thus, the higher the fog index value, the more complex the material is. Weak modal words were first introduced by Loughran and McDonald (2011) in their financial dictionary to gauge uncertain business tone. Words such as ”maybe”, ”appears”, ”might” frequently shown in the news suggest that these stories are less likely to attract investors. For example, Ahern and Sosyura (2015) uses this word category to gauge merger and acquisition rumours. These proxies are commonly used in the literature and should be effective to measure complexity in this chapter.

To measure the news informativeness features, I first utilise article length as a proxy for comprehensiveness, as suggested by You et al. (2017). A long news article tends to convey more useful information and helps (especially retail) investors better understand underlying events. Next, I assess news informativeness by categorizing the news topic as earnings-related or non-earnings-related. Specifically, I detect keywords from each news article using the word stem "earn". I identify the topic of a news article related to earnings if at least one word stem "earn" is found. The rationale behind this is that an article mentioning the word stem "earn" contains more information about a firm's fundamentals and is more likely to be value-relevant (Tetlock et al., 2008). Overall, it can be argued that a news story with long length and more "earn" word stems represents greater informativeness.

To test the hypothesis, I examine the cross-sectional difference of news characteristics individually. I first calculate the median level of weak tone at each time point, and divide the stocks with news having weaker tone into an *Ambiguous* group and the others into an *Accurate* group. Similarly, I divide more readable news stocks into a *Concise* group and less readable stocks into a *Complex* group based on the median fog index in each month. For article length, I divide all stocks into *Short* and *Long* groups based on the median word count in each month. Lastly, I split the sample into two groups: *EarningsEx*, where stocks without any news articles containing earnings topics are collected and *EarningsIn*, where stocks with only earnings-related news articles are included. To examine the statistical difference between each group, I construct $Ambiguous - Accurate$, $Complex - Concise$, $Short - Long$ and $EarningsEx - EarningsIn$, respectively. If the complexity and informativeness of news does affect investor trading behaviour, I expect that these four cross-sectional differences would be positive and significant for the SI news portfolio. The equally-weighted portfolio return is then computed for each subset sample to trace its post-formation performance.

Panel A of Table 4.8 reports the cross-sectional differences by textual complexity. The evidence is mixed: although the cross-sectional difference between ambiguous and accurate news stocks is positively and statistically significant (*t*-

statistic=2.29 for SI news), there is no clear evidence showing a significant difference between complex and concise news. This indicates that SI news predictability is not likely to come from hard-to-read news articles. Moving to Panel B, it is evident that the cross-sectional difference between short news statements (less comprehensive) and long news statements (more comprehensive) is significant (t-statistic of 2.70) for SI news stocks. Interestingly, when I split the sample into earnings-related and non-earnings-related topics, the cross-sectional difference between these two is insignificantly different from zero. This suggests that SI news is not exclusively earnings irrelevant and contains at least some valuable information.

Collectively, it can be concluded that SI news tends to use ambiguous tone and reports less comprehensively but such stories are not too complex to read and are not completely earnings-irrelevant. Hence the empirical results do not systemically support complexity and informativeness theory.

4.5 Discussion and Concluding Remarks

In this chapter, I first distinguish firm-level news as slowly incorporated, whereby stock returns do not promptly respond to corresponding news content, or quickly incorporated, where the stock return performance matches the news sentiment score. The feature of SI news, specifically, is having a good (bad) sentiment score but with bad (good) stock returns. A long-short portfolio is then constructed to capture the SI news effect by buying stocks with low returns but a high news sentiment score and selling stocks with high returns but a low news sentiment score. As a result, I find that this news-induced anomaly can achieve average returns of 139 basis points per month (equivalent to 16.68% per year).

The SI news effect can be explained by limited-attention theory. According to this theory, investors tend to focus on a few stocks instead of evenly dispersing their attention throughout the entire stock universe. This then leads to the phenomenon where a group of firm-related news items drop under an investor's radar. The stock returns thus react very slowly. In empirical tests, I proxy less investor attention by small size, low media coverage, low trading volume, low analyst coverage, low Google SVI and low Bloomberg AIA. The results, strikingly, show stronger stock return predictability for less attention-grabbing stocks.

The empirical results, however, do not strongly support complexity and informativeness theory in which the nature of news might cause investors' slow reaction. Generally, the textual complexity of news content should influence investors' trading behaviour, particularly for retail investors, whereby they might have no idea how to interpret complex news items or how to respond to ambiguous news articles. In addition, a short news article tends to be less comprehensive in reporting underlying events and therefore discourages investors from responding quickly. An article that does not contain value-relevant information could also fail to make investors take interest. Yet the evidence does not align with this theory.

In addition, other behavioural finance theories are consistent with the findings. Anchoring bias predicts that investors depend too heavily on initial information, making subsequent adjustments very hard. Consistent with the theory, investors may initially interpret good or bad news from stock returns since it is the easiest and most accessible approach. Contemporaneous news articles released from professional newswire archives may contain additional information over and beyond observed stock returns. The aggregated news tone associated with monthly returns, if mismatched, is less likely to be quickly accepted by investors to adjust their prior biased beliefs. As a result, this will lead to an underreaction. Nevertheless, this chapter provides an interesting avenue for future research in assessing whether the news articles tested in this chapter contain long-term forward-looking content, and as a result, incorporate value-relevant information into stock prices conditional upon the underlying firm's near-term events. It would also be worthwhile to conduct a more detailed topics analysis of the news articles I examine in this study.

Figure 4.1: Distribution of Firm-level News Observations

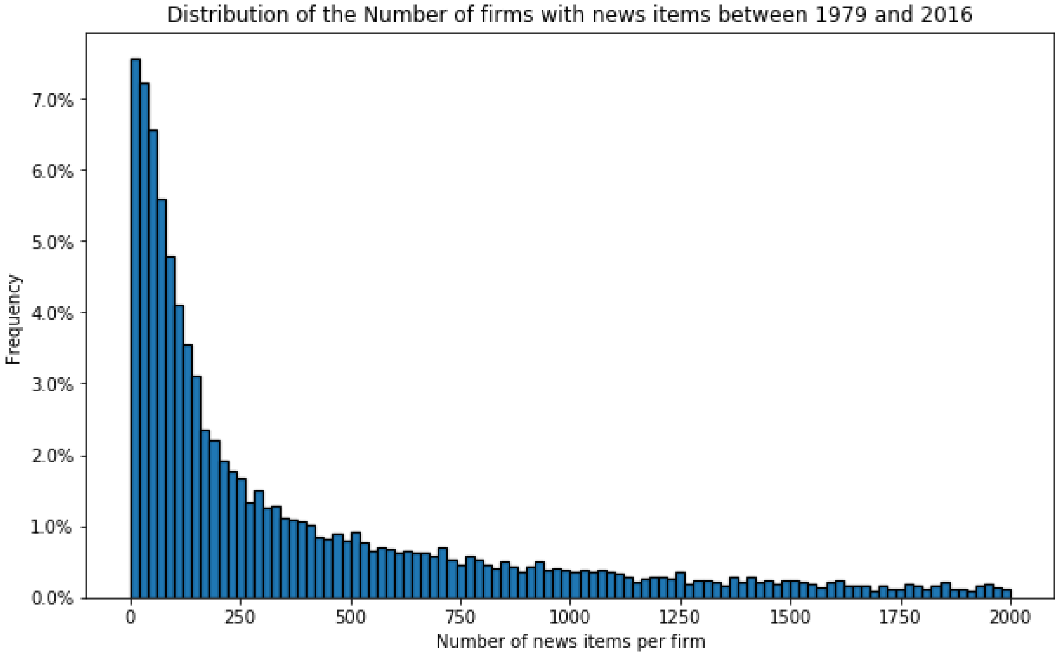

Figure 4.1 reports the frequency of news observations. The X-axis shows the number of news items for each firm while the Y-axis displays the percentage of firms among the total sample. The sample period is 1979 to 2016.

Figure 4.2: Portfolio Performance

Figure 4.2 plots the portfolio performance compared with the S&P500 during 1979 to 2016. Graph A is the portfolio performance index assuming a \$ 1 initial investment and Graph B is the portfolio's 12-month holding period return. Graph C and D are portfolio performance after considering 25 and 50 bps transaction costs.

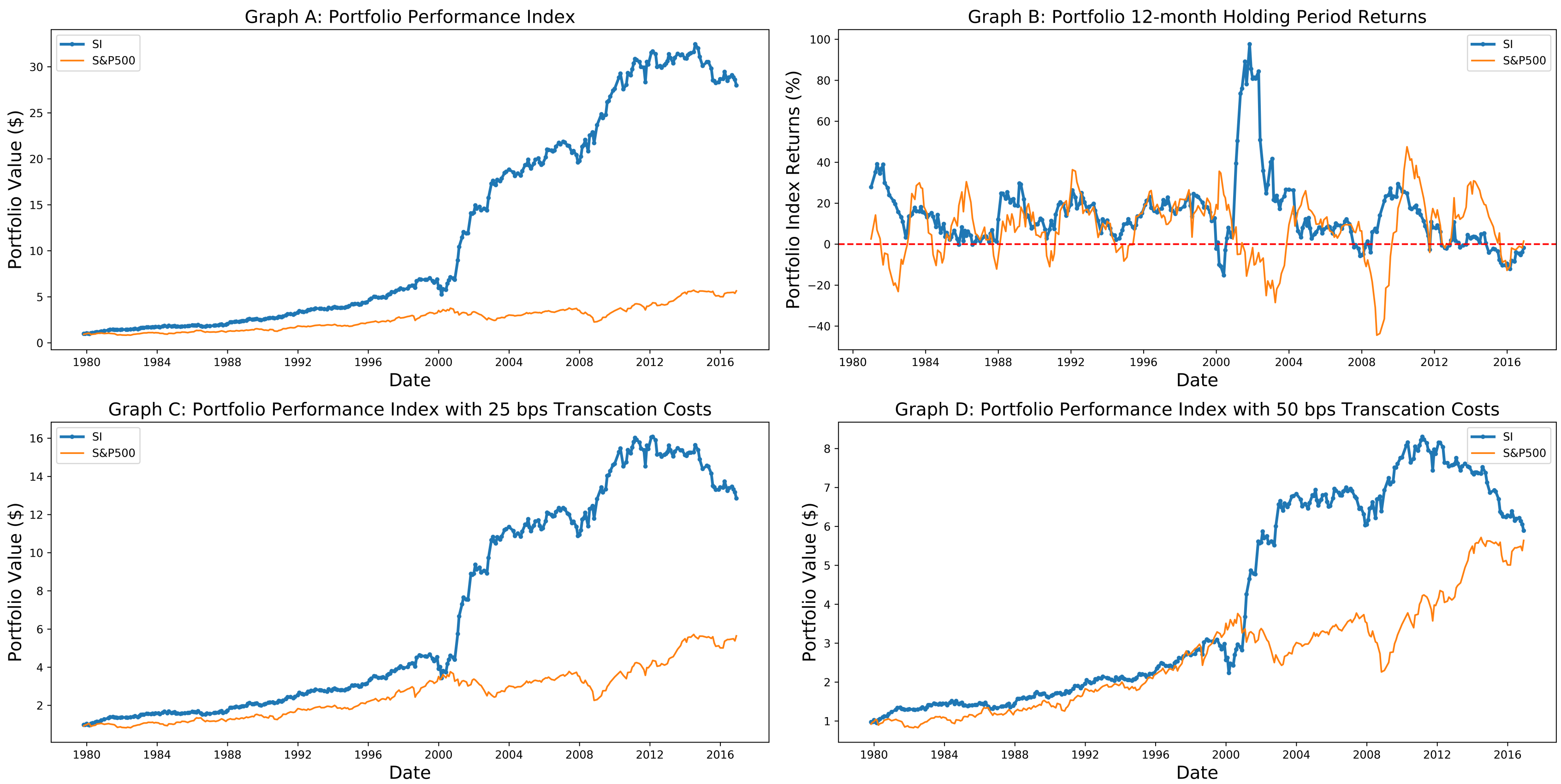

Table 4.1: Summary Statistics

Table 4.1 Panel A presents the summary statistics. *CVG* is the number of news articles. *RET* is the DGTW-adjusted return computed by following Daniel et al. (1997). *NSS* is the aggregated monthly news sentiment score. *SIZE* is computed by taking the natural logarithm of the stock market capitalisation at the end of each June. *BTM* is computed by taking the natural logarithm of the stock market value divided by the firm's book value, adjusted at each end of June. *MOM* is the stock's most recent 12-month cumulative returns. *ILLIQ* is a proxy for stock liquidity based on Amihud (2002). *BETA* is calculated following Scholes and Williams (1977); Dimson (1979). *IVOL* is the idiosyncratic risk computed by the standard deviation of residuals from the Fama-French-Carhart four-factor model over the month using daily returns. Panel B reports the summary statistics across all four types of news portfolios. The monthly sequential double-sorted approach is used. In each month, firms are sorted into three groups based on their past abnormal returns and then within each group stocks are further ranked into three groups based on their news sentiment scores. *SIGoodNews* is defined as those with low current stock returns but good positive news; *SIBadNews* are those with high current stock returns but bad negative news; *QIGoodNews* refers to news portfolios with high current stock returns and good positive news. *QIBadNews* are low stock return and bad news stocks. *Avg # of Stocks* is the average number of stocks in each portfolio in each month.

Panel A	Mean	Std	Min	0.25	0.5	0.75	Max
CVG	12.81	29.12	1.00	2.00	6.00	13.00	1568.00
RET (%)	-0.28	11.83	-37.36	-6.60	-0.58	5.47	49.90
NSS	0.00	0.89	-2.33	-0.61	0.04	0.63	2.19
SIZE (Ln)	5.90	1.98	1.50	4.41	5.85	7.31	10.97
BTM	-0.67	0.79	-3.16	-1.16	-0.60	-0.12	1.33
MOM (%)	-0.74	37.73	-80.16	-22.79	-2.45	14.29	194.50
ILLIQ	0.22	0.81	0.00	0.00	0.01	0.07	9.76
BETA	0.96	1.51	-4.47	0.13	0.88	1.73	6.75
IVOL (%)	2.62	1.93	0.47	1.28	2.02	3.32	12.28

Panel B	LRHS SI Good News	HRLS SI Bad News	HRHS QI Good News	LRLS QI Bad News
NSS_t	0.94	-0.96	1.02	-1.17
RET_t	-0.12	0.15	0.15	-0.14
CVG_t	8.68	11.75	10.87	11.71
CVG_{t+1}	10.70	12.84	13.21	12.37
Avg # of $Stocks_t$	240	239	240	240

Table 4.2: SI and QI News

Table 4.2 Panel A reports the return and news predictability across terciles. The formation period and estimation period are both one month. In each month, firms are sorted into three groups based on their past abnormal returns and then within each group stocks are further ranked into three groups based on their news sentiment scores. Abnormal return is the DGTW-adjusted return computed by following (Daniel et al., 1997). All returns are reported as monthly averages denoted in percentages. The sample period ranges from 1979 to 2016. Panel B reports time-series portfolio returns with risk-adjusted alphas including the Fama and French (1993) three-factor model (FF3F), the Fama-French-Carhart four-factor model (FF4F), the Fama and French (2017) five-factor model (FF5F), a liquidity-augmented Fama-French-Carhart four-factor model (FF4F Liq) and short-term augmented Fama-French-Carhart four-factor model (FF4F Rev). The alpha estimates are obtained by regressing monthly portfolio excess returns on the monthly returns from the risk factors. The definition of SI and QI presents as follows:

$$SI = LRHS - HRLS$$

$$QI = LRLS - HRHS$$

where SI news is a long-short portfolio of buying slowly incorporated good news (LRHS) and selling slowly incorporated bad news (HRLS). QI news is a long-short portfolio made by buying quickly incorporated good news (HRHS) stocks and selling quickly incorporated bad news (LRLS) stocks. T-statistics are reported in parentheses and **, *** refers to 5% and 1% significance levels respectively.

Panel A	HS	Mid	LS
LR	1.45%	1.22%	1.05%
Mid	0.94%	0.77%	0.84%
HR	0.67%	0.53%	0.45%

Panel B	TS Ret	FF3F	FF4F	FF5F	FF4F Liq	FF4F Rev
SI News	1.01%***	0.92%***	1.09%***	0.85%***	1.09%***	0.82%***
t-stat	(5.85)	(3.66)	(3.68)	(3.00)	(4.00)	(4.11)
QI News	-0.39%**	-0.27%	-0.64%***	-0.29%	-0.61%***	-0.38%**
t-stat	(-2.09)	(-1.14)	(-2.70)	(-0.94)	(-2.55)	(-2.42)
Slow-Minus-Quick	1.39%***	1.19%***	1.73%***	1.14%**	1.69%***	1.20%***
t-stat	(4.26)	(2.59)	(3.40)	(2.01)	(3.70)	(3.75)

Table 4.3: Predicting Stock Returns by SI News and QI News

Table 4.3 reports the stock return predictability of slowly incorporated news and quickly incorporated news based on various regression specifications. $SINws$ and $QINws$ are dummy variables if news is slowly incorporated or quickly incorporated, respectively. $NSS{*}SINws$ and $NSS{*}QINws$ are the interaction terms where NSS is a news sentiment score. $CONTROLS$ is a battery of control variables including $LRET$, $SIZE$, BTM, $BETA$, $IVOL$, MOM, $ILLIQ$. $SIZE$ is computed by taking the natural logarithm of stock market values in each previous month. $LRET$ is the lagged one-month stock return. BTM is computed by taking the natural logarithm of stock market values divided by firm book values adjusted at each end of June. $BETA$ is calculated following Scholes and Williams (1977); Dimson (1979). MOM is the stock's most recent 12-month cumulative returns. $IVOL$ is the idiosyncratic risk computed by the standard deviation of residuals from the Fama-French-Carhart four-factor model over the month using daily returns. $ILLIQ$ is a proxy for stock liquidity based on (Amihud, 2002). The sample period ranges from 1979 to 2016. T-statistics are reported in parentheses and **, *** refers to the 5% and 1% significance levels, respectively.

	Dep = EXRet					
	1	2	3	4	5	6
CONST	0.8823***	0.8602***	0.8841***	2.0814***	2.0905***	2.0721***
	(3.38)	(3.36)	(3.45)	(6.22)	(6.24)	(6.22)
NSS	0.1029***	-0.0445	0.2489***	0.0868***	0.1249***	0.0959***
	(3.64)	(-0.97)	(7.05)	(3.09)	(4.44)	(3.11)
NSS*SINws		0.3110***		0.2080**		0.2004**
		(3.24)		(2.30)		(2.21)
NSS*QINws			-0.2764***		-0.1651	-0.1351
			(-2.89)		(-1.63)	(-1.30)
SIGdNws		0.3508**		-0.0555		-0.0488
		(2.54)		(-0.52)		(-0.45)
SIBdNws		-0.1276		0.2653**		0.2768**
		(-1.04)		(2.36)		(2.44)
QIGdNws			-0.2443		0.1339	0.1581
			(-1.81)		(1.10)	(1.28)
QIBdNws			0.1705		-0.1609	-0.1262
			(1.10)		(-1.13)	(-0.87)
LRET				-0.0261***	-0.0265***	-0.0265***
				(-5.96)	(-6.02)	(-5.76)
SIZE				-0.1198***	-0.1195***	-0.1197***
				(-3.12)	(-3.11)	(-3.12)
BTM				0.1635**	0.1625**	0.1622**
				(2.31)	(2.30)	(2.30)
BETA				-0.0671	-0.0658	-0.0661
				(-1.41)	(-1.39)	(-1.39)
IVOL				-0.2075***	-0.2040***	-0.2078***
				(-4.44)	(-4.36)	(-4.45)
MOM				0.0043***	0.0044***	0.0043***
				(3.22)	(3.33)	(3.25)
ILLIQ				0.0982**	0.0994***	0.1021***
				(2.53)	(2.57)	(2.64)
Nobs	965505	965505	965505	794317	794317	794317
Adj_R^2	0.05%	0.36%	0.34%	6.12%	6.14%	6.19%

Table 4.4: Long-run Performance of SI and QI News over Different Horizons

Table 4.4 reports the performance of slowly incorporated news and quickly incorporated news over different holding periods. The portfolio is constructed by sorting stocks into tercile portfolios over five different holding periods: one month, three months, six months nine months and twelve months. In each month, firms are sorted into three groups based on their past abnormal returns and then within each group, stocks are further ranked into three groups based on their news sentiment scores. Abnormal return is the DGTW-adjusted return computed by following (Daniel et al., 1997). The definition of SI and QI is presented as follows:

$$SI = LRHS - HRLS$$
$$QI = LRLS - HRHS$$

where SI news is a long-short portfolio made by buying slowly incorporated good news (LRHS) stocks and selling slowly incorporated bad news (HRLS) stocks. QI news is a long-short portfolio constructed by buying quickly incorporated good news (HRHS) stocks and selling quickly incorporated bad news (LRLS) stocks. The sample period ranges from 1979 to 2016. T-statistics are reported in parentheses and **, *** refers to the 5% and 1% significance levels respectively.

Panel A: Long-run returns of SI and QI news

Months	SI News	t-stat	QI News	t-stat	Slow-Minus-Quick	t-stat
[1,2]	1.01%***	(5.85)	-0.39%**	(-2.09)	1.39%**	(2.43)
[2,4]	-0.03%	(-0.24)	0.20%	(1.54)	-0.22%	(-1.01)
[2,7]	-0.07%	(-0.83)	0.20%	(1.93)	-0.28%	(-1.64)
[2,10]	-0.06%	(-0.83)	0.17%	(1.79)	-0.23%	(-1.53)
[2,13]	-0.09%	(-1.30)	0.14%	(1.76)	-0.22%	(-1.90)

Panel B: Long-run risk-adjusted returns of SI and QI news

Months	SI News	QI News	Slow-Minus-Quick
[1,2]	0.28	-0.12	0.21
[2,4]	0.00	0.07	-0.04
[2,7]	-0.03	0.09	-0.06
[2,10]	-0.03	0.08	-0.06
[2,13]	-0.05	0.08	-0.07

Table 4.5: Robustness Checks

Table 4.5 reports robustness checks of calendar-time portfolio tests. Panel A reports an alternative method of portfolio construction for returns and news sentiment scores. Panel B excludes stocks with prices below than \$ 5 per share. Panel C eliminates stocks with news observations in the holding period. Panel D measures news tone using different sentiment tools including Loughran and McDonald (2011) negative words and positive words and the Google Natural Language Sentiment scores. Panel E reports portfolio performance on a weekly basis. Panel F reports portfolio performance using the alternative definition of slowly incorporated news. T-statistics are reported in parentheses and **, *** refers to the 5% and 1% significance levels respectively.

Panel A	SI News	t-stat	QI News	t-stat	SMQ	t-stat
News Beta	1.28%***	(5.16)	0.80%**	(2.27)	0.49%**	(2.36)
Independent double-sorted	1.00%***	(5.80)	-0.40%**	(-2.22)	1.41%***	(4.33)
Panel B	**SI News**	**t-stat**	**QI News**	**t-stat**	**SMQ**	**t-stat**
Price>= \$5	0.71%***	(4.81)	-0.16%	(-1.08)	0.87%***	(3.22)
Panel C	**SI News**	**t-stat**	**QI News**	**t-stat**	**SMQ**	**t-stat**
Excluded subsequent news	1.45%***	(4.59)	-0.98%***	(-3.07)	2.43%***	(5.07)
Panel D	**SI News**	**t-stat**	**QI News**	**t-stat**	**SMQ**	**t-stat**
LM negative score	0.99%***	(5.47)	-0.49%***	(-2.68)	1.47%***	(4.43)
LM positive score	0.74%***	(4.22)	-0.69%***	(-3.65)	1.43%***	(4.23)
Google NL score	0.78%***	(4.53)	-0.63%***	(-3.32)	1.41%***	(4.30)
Panel E	**SI News**	**t-stat**	**QI News**	**t-stat**	**SMQ**	**t-stat**
Weekly frequency	2.09%***	(12.32)	-0.79%***	(-4.47)	2.88%***	(9.71)
Panel F	**SI News**	**t-stat**	**QI News**	**t-stat**	**SMQ**	**t-stat**
Alternative SI news	1.59%***	(3.63)	0.01%	(0.02)	1.58%**	(2.20)
Alternative QI news	-0.07%	(-0.15)	0.01%	(0.01)	-0.08%	(-0.10)

Table 4.6: Slowly Incorporated and Quickly Incorporated News over Time

Table 4.6 reports the subsample analysis. The sample is split into four different periods: 1979 - 1987, 1988 - 1997, 1998 - 2007 and 2008 - 2016 respectively. All returns are converted into monthly averages in percent. Panel A reports the SI and QI news portfolio returns monthly. The definition of SI and QI are presented as follows:

$$SI = LRHS - HRLS$$

$$QI = LRLS - HRHS$$

where SI news is a long-short portfolio comprising buying slowly incorporated good news (LRHS) stocks and selling slowly incorporated bad news (HRLS) stocks. QI news is a long-short portfolio of buying quickly incorporated good news (HRHS) stocks and selling quickly incorporated bad news (LRLS) stocks. T-statistics are reported in parentheses and **, *** refers to the 5% and 1% significance levels respectively.

	1979 - 1987	1988 - 1997	1998 - 2007	2008 - 2016
SI News	0.93%***	1.08%***	1.32%***	0.62%**
t-stat	(3.06)	(4.75)	(2.77)	(2.20)
QI News	0.14%	-0.53%**	-0.51%	-0.58%
t-stat	(0.47)	(-1.97)	(-1.03)	(-1.65)
Slow-Minus-Quick	0.79%	1.61%***	1.83%**	1.20%**
t-stat	(1.61)	(3.96)	(1.97)	(2.06)

Table 4.7: SI and QI News under Different Information Environment

Table 4.7 reports the performance of slowly incorporated news and quickly incorporated news portfolios in different information environments. I independently sort all stocks into two portfolios based on their most recent market capitalisation (Size), current monthly change of media coverage, turnover ratio and analyst coverage. I report the "Small-Large" SIZE, "Low-High" Δ MEDIA, "Low-High" TURN, "Low-High"AstCvg spread profitability in the post-formation period. T-statistics are reported in parentheses and **, *** refers to the 5% and 1% significance levels respectively.

Panel A	Size Subsamples			MEDIA Subsamples		
	SmallSIZE	BigSIZE	Small-Big	LowMEDIA	HighMEDIA	Low-High
SI News	1.61%***	0.50%***	1.11%***	1.44%***	0.09%	1.35%***
	(6.58)	(3.04)	(4.54)	(7.52)	(0.54)	(7.10)
QI News	-0.61%	-0.18%	-0.43%	-0.64%***	0.22%	-0.86%***
	(-1.92)	(-1.07)	(-1.50)	(-3.18)	(1.41)	(-4.55)
Slow-Minus-Quick	2.22%***	0.67%**	1.55%***	2.08%***	-0.13%***	2.21%***
	(4.67)	(2.28)	(3.65)	(6.00)	(-0.48)	(7.56)

Panel B	TURN Subsamples			AstCvg Subsamples		
	LowTURN	HighTURN	Low-High	LowAstCvg	HighAstCvg	Low-High
SI News	1.36%***	0.56%**	0.80%***	1.03%***	0.57%***	0.46%**
	(8.94)	(2.03)	(2.95)	(4.53)	(3.06)	(1.99)
QI News	-0.34%**	-0.29%	-0.04%	-0.08%	-0.20%	0.12%
	(-2.12)	(-1.01)	(-0.16)	(-0.33)	(-1.04)	(0.50)
Slow-Minus-Quick	1.70%***	0.85%	0.85%	1.11%***	0.77%**	0.34%
	(6.79)	(1.71)	(1.87)	(2.76)	(2.33)	(0.94)

Table 4.8: SI and QI News in Different Nature of News

Table 4.8 reports the performance of slowly incorporated news and quickly incorporated news for different news characteristics. In Panel A, I independently sort all stocks into two portfolios based on textual complexity. In Panel B, I again sort all stocks into two portfolios based on news informativeness. I report the “Ambiguous-Accurate” Tone, “Complex-Concise” Readability, “Short-Long” Length and “EarningsEx-EarningsIn” Topic spread profitability in the post-formation period. T-statistics are reported in parentheses and **, *** refers to the 5%, 1% significance levels respectively.

Panel A	SI and QI News across Accuracy Subsamples			SI and QI News across Readability Subsamples		
	Ambiguous	Accurate	Ambiguous - Accurate	Complex	Concise	Complex - Concise
SI News	1.05%***	0.42%	0.63%**	1.10%***	0.81%***	0.30%
	(3.95)	(1.61)	(2.29)	(5.34)	(3.97)	(1.30)
QI News	-0.51%**	0.02%	-0.53%	-0.67%***	-0.32%	-0.34%
	(-2.08)	(0.07)	(-1.91)	(-2.93)	(-1.54)	(-1.68)
Slow-Minus-Quick	1.55%***	0.40%	1.15%***	1.77%***	1.13%***	0.64%**
	(3.37)	(0.83)	(2.55)	(4.77)	(3.29)	(2.11)

Panel B	SI and QI News across Length Subsamples			SI and QI News across Earnings Subsamples		
	Short	Long	Short - Long	EarningsEx	EarningsIn	EarningsEx - EarningsIn
SI News	1.29%***	0.70%***	0.58%***	0.92%***	0.77%***	0.16%
	(5.78)	(3.62)	(2.70)	(5.33)	(3.61)	(1.04)
QI News	-0.79%***	-0.19%	-0.60%***	-0.45%**	-0.27%	-0.18%
	(-3.53)	(-0.81)	(-2.60)	(-2.35)	(-1.36)	(-1.09)
Slow-Minus-Quick	2.08%***	0.90%**	1.18%***	1.38%***	1.04%***	0.34%
	(5.40)	(2.45)	(3.74)	(4.10)	(2.90)	(1.65)

Chapter 5

Summary of the Thesis

5.1 Concluding Remarks

Recent literature has linked firm-level news items to cross-sectional stock return studies. A large body of literature has documented that news has substantial explanatory power for existing anomalies. For example, Engelberg et al. (2018) show that anomaly returns are 50% higher on corporate news days and six times higher on earnings announcement days using 97 different anomalies. This thesis continues to investigate and advance this strand of literature. Using large and comprehensive firm-level news items from the Dow Jones Newswire Archive, I document the differential return performance in each category of anomalies conditional upon whether news arrives in the market.

In Chapter 2, I first study how the presence of news coverage affects lottery-type stocks, proxied by the maximum daily returns (MAX) in the prior month. There is an augmented negative relationship between MAX stocks without news and expected returns, whereas MAX with news coverage generates return momentum. Furthermore, the differential future return performance, caused by the arrival of news, is best explained by information uncertainty mitigation theory. As the lottery-type stocks exhibit high information uncertainty and equity's option-likeness (Barinov, 2018), the arrival of news attenuates firm-level uncertainty around each MAX event. As a result, the lottery features of those stocks will disappear, with the subsequent stock returns positively priced. Other possible channels such as investor inattention and divergence of opinions fail to explain the empirical phenomenon. By controlling for investor attention indicators, I find that MAX_{news} still generates positive future returns. The MIN_{news}, by contrast, generates negative future returns, rejecting the possibility that different opinions cause positive future returns.

By applying a "co-coverage" concept to news articles, Chapter 3 investigates the lead-lag return momentum literature. I show that news co-coverage is a key to understanding lead-lag return momentum previously documented in the literature. Specifically, the base firm's share price responds more favourably to the shocks to the stock returns of its News-Based Peers (NBPs) than of its traditional industry peers that are not NBPs. I also document a monthly excess return of 1.06%(0.77%) by generating an equal-weighted (value-weighted) long-short portfolio that buys NBPs in the highest quintile by their return shocks and sells those in the lowest quintile, indicating a slow reaction to the shocks transmitted from NBPs. The response can persist for several months and does not reverse, which rejects the investor overreaction theory and suggests that investors are not immediately aware of the importance of the economic links reweighted by NBPs. The abnormal Google Search Volume Index test further provides suggestive evidence that investor attention to the shocks of NBPs initially underperforms than of highly-visible SIC peers with a subsequent upward trend in a few months.

Chapter 4 looks into the short-term return reversals. I show that whether aggregated news sentiment scores match with contemporaneous stock prices is an important indicator of predictable future stock returns. Specifically, stocks are categorized as either news slowly or quickly incorporated into contemporaneous stock prices. The slowly incorporated (SI) news, identified by having a good (bad) news sentiment score but with bad (good) stock returns, can strongly predict future returns. In contrast, the quickly-incorporated (QI) news does not observe a predictable pattern for future returns. The differential cross-sectional return between SI and QI news yields a statistically significant profit of 139 basis points per month and is robust after considering transaction costs. I also perform additional empirical tests to study the underlying driver of slowly incorporated news. I find the evidence is consistent with investor inattention theory. The profitability of slowly incorporated news is more pronounced among small-size firms, low change of media coverage, low trading volume, and low analyst coverage subsamples. The empirical results, however, do not systematically support the alternative explanation that the textual complexity of news content might cause investors' slow reaction.

5.2 Suggestions for Future Research

While a rich set of empirical evidence has revealed the significant role of news in a different category of anomalies, a large amount of work could be further developed in the future research in each chapter of this thesis.

First of all, this thesis only investigates the relation between firm-level news and different anomalies, leaving the possibility of studying other types of news in these empirical settings. For instance, in Chapter 2, I investigate the future returns of MAX stocks conditional upon whether the firm-level news coverage is surrounded by each MAX event. This news type is most likely to cover own-firm's corporate news events. As a result, it opens a potential discussion about the pricing effect of other types of news such as market-wide news announcements, industry news events or overnight news articles. In particular, prior literature documents that market reaction to firms' earnings announcements tends to be more pronounced when another simultaneous news arrives in the same industry, referred as to the attention distraction effect (Hirshleifer et al., 2009). As Bali et al. (2019) find that retail investors' attention to lottery-type stocks induces overvaluation, it seems reasonable to expect that the investor distraction effect will also attenuate lottery stocks' features. This will be an interesting avenue to explore in the future.

There is also a broad range of applications for my findings in Chapter 3 where I construct a time-varying firm-centric grouping scheme using news co-coverage, termed as News-Based Peers (NBPs). The unique feature of the NBP scheme is that it can consistently capture the up-to-date relative importance of existing economic links, such as three-digit SIC, TNIC-3, and supply industry. In the current analysis, I only look at the asset pricing implications of the scheme while other implications might be worth exploring. According to Ahern and Harford (2014), a sectoral shock (e.g., merger activity) can transmit to close industries quickly and to distant industries with a delay in the supply-customer scheme. This helps to explain the formation and propagation of merger waves. If the NBP can consistently reweight these existing industries, one would expect a significant difference when comparing the likelihood of merger deals between NBP and NonNBP schemes. As such, it could be an interesting avenue to explore the validity and real economic consequences of my proposed grouping scheme.

Finally, my empirical setting is also restricted to the US equity market, which has limited implications for international investors. For example, in Chapter 4, I develop a trading strategy using slowly incorporated news that earns 139 basis points per year and is robust after considering transaction costs. As the trading strategy can only be implemented in the US market due to data availability, replicating my findings in different markets will be the best way to examine their robustness. As such, it is worth constructing an alternative news data set using local newspapers.

References

Ahern, K. R., Harford, J., 2014. The importance of industry links in merger waves. Journal of Finance 69, 527–576.

Ahern, K. R., Sosyura, D., 2015. Rumor has it: Sensationalism in financial media. Review of Financial Studies 28, 2050–2093.

Ahmad, K., Han, J., Hutson, E., Kearney, C., Liu, S., 2016. Media-expressed negative tone and firm-level stock returns. Journal of Corporate Finance 37, 152–172.

Ali, U., Hirshleifer, D., 2019. Shared analyst coverage: Unifying momentum spillover effects. Journal of Financial Economics 136, 649–675.

Amihud, Y., 2002. Illiquidity and stock returns: Cross-section and time-series effects. Journal of Financial Markets 5, 31–56.

Atilgan, Y., Bali, T. G., Demirtas, K. O., Gunaydin, A. D., 2020. Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns. Journal of Financial Economics 135, 725–753.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. Journal of Finance 61, 1645–1680.

Bali, T. G., Bodnaruk, A., Scherbina, A., Tang, Y., 2017a. Unusual news flow and the cross section of stock returns. Management Science 64, 4137–4155.

Bali, T. G., Brown, S. J., Murray, S., Tang, Y., 2017b. A lottery-demand-based explanation of the beta anomaly. Journal of Financial and Quantitative Analysis 52, 2369–2397.

Bali, T. G., Cakici, N., Whitelaw, R. F., 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. Journal of Financial Economics 99, 427–446.

Bali, T. G., Hirshleifer, D. A., Peng, L., Tang, Y., 2019. Attention, social interaction, and investor attraction to lottery stocks, working paper, Georgetown University.

Barber, B. M., Odean, T., 2007. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. Review of Financial Studies 21, 785–818.

Barberis, N., Mukherjee, A., Wang, B., 2016. Prospect theory and stock returns: An empirical test. Review of Financial Studies 29, 3068–3107.

Barberis, N., Shleifer, A., Vishny, R., 1998. A model of investor sentiment. Journal of Financial Economics 49, 307–343.

Barinov, A., 2018. Stocks with extreme past returns: Lotteries or insurance? Journal of Financial Economics 129, 458–478.

Ben-Rephael, A., Da, Z., Israelsen, R. D., 2017. It depends on where you search: Institutional investor attention and underreaction to news. Review of Financial Studies 30, 3009–3047.

Bernard, V. L., Thomas, J. K., 1989. Post-earnings-announcement drift: Delayed price response or risk premium? Journal of Accounting Research 27, 1–36.

Bethke, S., Gehde-Trapp, M., Kempf, A., 2017. Investor sentiment, flight-to-quality, and corporate bond comovement. Journal of Banking & Finance 82, 112–132.

Boudoukh, J., Feldman, R., Kogan, S., Richardson, M., 2018. Information, trading, and volatility: Evidence from firm-specific news. Review of Financial Studies 32, 992–1033.

Bushee, B. J., Core, J. E., Guay, W., Hamm, S. J., 2010. The role of the business press as an information intermediary. Journal of Accounting Research 48, 1–19.

Calvet, L. E., Fisher, A. J., 2007. Multifrequency news and stock returns. Journal of Financial Economics 86, 178–212.

Caporale, G. M., Spagnolo, F., Spagnolo, N., 2018. Macro news and bond yield spreads in the euro area. European Journal of Finance 24, 114–134.

Chan, W. S., 2003. Stock price reaction to news and no-news: Drift and reversal after headlines. Journal of Financial Economics 70, 223–260.

Chen, H., De, P., Hu, Y. J., Hwang, B.-H., 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies 27, 1367–1403.

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

Cohen, L., Lou, D., 2012. Complicated firms. Journal of Financial Economics 104, 383–400.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. Journal of Finance 66, 1461–1499.

Da, Z., Engelberg, J., Gao, P., 2014a. The sum of all fears investor sentiment and asset prices. Review of Financial Studies 28, 1–32.

Da, Z., Gurun, U. G., Warachka, M., 2014b. Frog in the pan: Continuous information and momentum. Review of Financial Studies 27, 2171–2218.

Daniel, K., Grinblatt, M., Titman, S., Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. Journal of Finance 52, 1035–1058.

DeLisle, R. J., Mauck, N., Smedema, A. R., 2016. Idiosyncratic volatility and firm-specific news: Beyond limited arbitrage. Financial Management 45, 923–951.

Dimson, E., 1979. Risk measurement when shares are subject to infrequent trading. Journal of Financial Economics 7, 197–226.

Dougal, C., Engelberg, J., Garcia, D., Parsons, C. A., 2012. Journalists and the stock market. Review of Financial Studies 25, 639–679.

Egginton, J., Hur, J., 2018. The robust "maximum daily return effect as demand for lottery" and "idiosyncratic volatility puzzle". Journal of Empirical Finance 47, 229–245.

Engelberg, J., Gao, P., 2011. In search of attention. Journal of Finance 66, 1461–1499.

Engelberg, J., McLean, R. D., Pontiff, J., 2018. Anomalies and news. Journal of Finance 73, 1971–2001.

Engelberg, J. E., Reed, A. V., Ringgenberg, M. C., 2012. How are shorts informed?: Short sellers, news, and information processing. Journal of Financial Economics 105, 260–278.

Fama, E. F., French, K. R., 1992. The cross-section of expected stock returns. Journal of Finance 47, 427–465.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E. F., French, K. R., 2017. International tests of a five-factor asset pricing model. Journal of Financial Economics 123, 441–463.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81, 607–636.

Fang, L., Peress, J., 2009. Media coverage and the cross-section of stock returns. Journal of Finance 64, 2023–2052.

Fong, W. M., Toh, B., 2014. Investor sentiment and the max effect. Journal of Banking & Finance 46, 190–201.

Frank, M. Z., Sanati, A., 2018. How does the stock market absorb shocks? Journal of Financial Economics 129, 136–153.

Frazzini, A., Pedersen, L. H., 2014. Betting against beta. Journal of Financial Economics 111, 1–25.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Review of Financial Studies 33, 2326–2377.

Fu, F., 2009. Idiosyncratic risk and the cross-section of expected stock returns. Journal of Financial Economics 91, 24–37.

Garcia, D., 2013. Sentiment during recessions. Journal of Finance 68, 1267–1300.

Grinblatt, M., Titman, S., 1992. The persistence of mutual fund performance. Journal of Finance 47, 1977–1984.

Grundy, B. D., Martin, J. S. M., 2001. Understanding the nature of the risks and the source of the rewards to momentum investing. Review of Financial Studies 14, 29–78.

Han, Y., Yang, K., Zhou, G., 2013. A new anomaly: The cross-sectional profitability of technical analysis. Journal of Financial and Quantitative Analysis 48, 1433–1461.

Henry, E., 2008. Are investors influenced by how earnings press releases are written? Journal of Business Communication 45, 363–407.

Hillert, A., Jacobs, H., Müller, S., 2014. Media makes momentum. Review of Financial Studies 27, 3467–3501.

Hirshleifer, D., Lim, S. S., Teoh, S. H., 2009. Driven to distraction: Extraneous events and underreaction to earnings news. Journal of Finance 64, 2289–2325.

Hirshleifer, D., Teoh, S. H., 2003. Limited attention, information disclosure, and financial reporting. Journal of Accounting and Economics 36, 337–386.

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

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy 124, 1423–1465.

Hoberg, G., Phillips, G. M., 2018. Text-based industry momentum. Journal of Financial and Quantitative Analysis 53, 2355–2388.

Hong, H., Stein, J. C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. Journal of Finance 54, 2143–2184.

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

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28, 650–705.

Huang, D., Jiang, F., Tu, J., Zhou, G., 2015. Investor sentiment aligned: A powerful predictor of stock returns. Review of Financial Studies 28, 791–837.

Huang, J., 2018. The customer knows best: The investment value of consumer opinions. Journal of Financial Economics 128, 164–182.

Jiang, F., Lee, J., Martin, X., Zhou, G., 2019. Manager sentiment and stock returns. Journal of Financial Economics 132, 126–149.

Jiang, H., Li, S. Z., Wang, H., 2017. News momentum, working paper, Michigan State University.

Jiang, H., Li, S. Z., Wang, H., 2020. Pervasive underreaction: Evidence from high-frequency data, working paper, Michigan State University.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 263–291.

Kaniel, R., Parham, R., 2017. Wsj category kings–the impact of media attention on consumer and mutual fund investment decisions. Journal of Financial Economics 123, 337–356.

Kelly, B. T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics 134, 501–524.

Kothari, S. P., Li, X., Short, J. E., 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. Accounting Review 84, 1639–1670.

Kräussl, R., Mirgorodskaya, E., 2017. Media, sentiment and market performance in the long run. European Journal of Finance 23, 1059–1082.

Kumar, A., 2009a. Hard-to-value stocks, behavioral biases, and informed trading. Journal of Financial and Quantitative Analysis 44, 1375–1401.

Kumar, A., 2009b. Who gambles in the stock market? Journal of Finance 64, 1889–1933.

Lawrence, A., 2013. Individual investors and financial disclosure. Journal of Accounting and Economics 56, 130–147.

Lee, C. M., Ma, P., Wang, C. C., 2015. Search-based peer firms: Aggregating investor perceptions through internet co-searches. Journal of Financial Economics 116, 410–431.

Liu, Y., Wu, X., 2019. Labor links and shock transmissions, working paper, University of Rochester.

Lochstoer, L. A., Tetlock, P. C., 2020. What drives anomaly returns? Journal of Finance 75, 1417–1455.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. Journal of Finance 66, 35–65.

Loughran, T., McDonald, B., 2014. Measuring readability in financial disclosures. Journal of Finance 69, 1643–1671.

Loughran, T., McDonald, B., 2015. The use of word lists in textual analysis. Journal of Behavioral Finance 16, 1–11.

Lynch, A. W., Balduzzi, P., 2000. Predictability and transaction costs: The impact on rebalancing rules and behavior. Journal of Finance 55, 2285–2309.

Menzly, L., Ozbas, O., 2010. Market segmentation and cross-predictability of returns. Journal of Finance 65, 1555–1580.

Moskowitz, T. J., Grinblatt, M., 1999. Do industries explain momentum? Journal of Finance 54, 1249–1290.

Müller, S., 2019. Economic links and cross-predictability of stock returns: Evidence from characteristic-based "styles". Review of Finance 23, 363–395.

Nguyen, H. T., Truong, C., 2018. When are extreme daily returns not lottery? at earnings announcements! Journal of Financial Markets 41, 92–116.

Ohlson, J. A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18, 109–131.

Pástor, L., Stambaugh, R. F., 2003. Liquidity risk and expected stock returns. Journal of Political Economy 111, 642–685.

Patton, A. J., Verardo, M., 2012. Does beta move with news? firm-specific information flows and learning about profitability. Review of Financial Studies 25, 2789–2839.

Scholes, M., Williams, J., 1977. Estimating betas from nonsynchronous data. Journal of Financial Economics 5, 309–327.

Schwenkler, G., Zheng, H., 2019. The network of firms implied by the news, working paper, Boston University.

Shi, Y., Liu, W.-M., Ho, K.-Y., 2016. Public news arrival and the idiosyncratic volatility puzzle. Journal of Empirical Finance 37, 159–172.

Steeley, J. M., 2004. Stock price distributions and news: Evidence from index options. Review of Quantitative Finance and Accounting 23, 229–250.

Tao, R., Brooks, C., Bell, A., 2020a. Tomorrow's fish and chip paper? slowly incorporated news and the cross-section of stock returns. European Journal of Finance (forthcoming).

Tao, R., Brooks, C., Bell, A. R., 2020b. When is a max not the max? how news resolves information uncertainty. Journal of Empirical Finance 57, 33–51.

Tetlock, P. C., 2007. Giving content to investor sentiment: The role of media in the stock market. Journal of Finance 62, 1139–1168.

Tetlock, P. C., 2010. Does public financial news resolve asymmetric information? Review of Financial Studies 23, 3520–3557.

Tetlock, P. C., 2011. All the news that's fit to reprint: Do investors react to stale information? Review of Financial Studies 24, 1481–1512.

Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. Journal of Finance 63, 1437–1467.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5, 297–323.

Umar, T., 2020. Complexity aversion when seeking alpha, working paper, Rice University.

Wang, Y., Zhang, B., Zhu, X., 2018. The momentum of news, working paper, Central University of Finance and Economics.

You, J., Zhang, B., Zhang, L., 2017. Who captures the power of the pen? Review of Financial Studies 31, 43–96.

Zhang, X. F., 2006. Information uncertainty and stock returns. Journal of Finance 61, 105–137.

Appendix A

Dow Jones News Filtering Process for NBPs

A.1 Identification of NBPs

Our primary textual data is collected from Dow Jones Newswire Archive, which contains rich information [1]. It contains all of the news from the Dow Jones Newswire and all *The Wall Street Journal* newspapers. Besides the headline and main text, each news article has tags for, e.g., related tickers, subject codes, etc. Figure A1 shows an example of typical news co-coverage from the Dow Jones news article. One can see this news item attaches four tickers: AAPL, AMZN, GOOG, and GOOGL. In particular, the second leading paragraph mentions these companies: Apple Inc, Amazon.com Inc, and Google LLC with its parent company Alphabet Inc, with all revenues boosted.

To obtain the baseline results, we pre-process the raw Dow Jones Newswire Archive data in four major steps. The pre-step is to map all potential news items from the Dow Jones Newswire Archive to the CRSP universe. To do so, we download

[1]Dow Jones Newswire is a global leading real-time news product and is commonly used in many research papers – e.g., (Tetlock, 2010, 2011; Engelberg et al., 2012, 2018).

the historical name change file. All firm PERMCO (CRSP's permanent company identifier) are changed to their historical tickers according to the period in which they were officially registered. As a result, we construct PERMCO-TICKER linktable and assign news articles to the particular firm if the ticker tagged in the news matches the stock's within the valid period. Following the conventional asset pricing literature, we only retained these common stocks (share code is 10 and 11) and traded in NYSE, AMEX, and NASDAQ stock exchanges.

Through this method, our initial sample has 13,985 Dow Jones news firms between July 1979 and December 2017. One concern is that since the Dow Jones Newswire Archive is underdeveloped and somewhat incomplete during the early stage, it might be the case that a number of firms have news articles in the Newswire but these articles do not have tickers attached. Therefore, our procedure does not capture these firm-specific news items. On the other hand, these missing firms might not be covered by the Newswire at all. In an unreported test, we compare two statistics: the total number of stocks from the CRSP universe and the number of matched Dow Jones News stocks. During the early period (1979 to 1994), the percentage of stocks covered by measurable Dow Jones news articles is around 20% to 30%. However, this statistic increases rapidly to over 95% since 1995. As such, there is strong evidence to believe that most missing firms are clustered in the early years. Consequently, we only retain the sample between 1995 and 2017 to avoid any potential bias and this leaves us 7.7 million news observations.

In the second data processing step, we extract multiple-ticker articles. We request that each article has at least two tickers, which is to apply the news co-coverage setting. This step substantially reduces the sample to only 1.7 million observations, accounting for only 22%. The result also implies that the majority of the news articles are single-ticker attached, which is firm-specific.

To filter these noise news such as morning briefs and market overlooks, we perform a coarse subject analysis. As the Dow Jones Newswire Archive assigns a set of subject code to each article to classify news topics, we only retain the article with clear themes and meaningful subject tags.[2] As Table A2 shows, top-frequent tags include corporate actions, bond-issuing, ADRs, and mergers and acquisitions, which are likely to report economic links between firms. As a result, this filter reduces the initial sample by 25.77%.

Lastly, we apply a simple rule of thumb to further potentially rule out any uninformative news based on the number of tickers under each news article. In Figure A2, we find that 95% Dow Jones news has fewer than six attached tickers. In other words, there are 5% news articles with labelled tickers larger than six, which is unlikely to report clear business links across firms and investors are less likely to pick up these stories. Collectively, the final step cuts the sample size to only 1 million news articles.

Since our primary goal is to identify genuine newsworthiness economic links, we consider several refinements to further tackle the "noisy" link issue. First, we output the relatedness fraction pairwise measures between each base peer firm. Specifically, for peer firm j of a base firm i in month t, $Frac_{ijt}$ is computed as a percentage of co-coverage pairwise firm i and j among all base firm i's co-coverage in that month. A high relatedness between firm i and j suggests that two firms are frequently mentioned within the same news articles, which generally should be considered as highly economically linked. Putting it differently, one can reckon it as a distance metric, reflecting the degree of relatedness between two firms in the natural business events or the perspective of journalists. The intuition behind is analogous to the recommendation systems employed by high-tech companies such as Netflix, Amazon or Youtube. For example, Netflix provides each user with a watch list when he or

[2] We acknowledge that there is no standard process to reckon whether a topic code is meaningful or not. Therefore, our refined subject list remains arbitrary. In an unreported test, we employ an non-arbitrary method to cut upper and lower bounds based on the topics' frequency to obtain these informative news articles and results remain similar.

she chooses a particular video under the "people watch this video also watch these". The Amazon also employs the similar technique – "customers who bought this item also bought".

Second, we require that each base firm has at least ten NBPs. Based on our observation to the intermediate news data, the firm with fewer than ten peers tends to have very few news observations, which is more likely to suffer severely from noisy link issue. For example, if a given pair of firms - firm A (base firm) and firm B (peer firm) - are only in one co-coverage news in a month, then firm B would be the only identified peer for base firm A. The relatedness fraction between A and B then becomes 100%. If they appear in co-coverage news by any chance of randomness rather than true economically linked, a likely outcome is that the relatedness is overestimated and the identification turns to be biased.

Lastly, following Hoberg and Phillips (2016) and Liu and Wu (2019), we impose a simple minimum relatedness threshold by including only the base firm whose pairwise relatedness fractions above a pre-specified minimum threshold. A high threshold will lead to fewer but economically closer NBPs for each base firm, and a low threshold yields larger groups but more low relevant peers. Since the primary aim of the exercise is to augment three-digit SIC scheme, we set up a granularity of the NBPs the same as that of the three-digit SIC. The steps are as follows: First, we calculate the percentage of the true membership pairs for the three-digit SIC, which is 2.05%. In other words, one can consider any two randomly selected firms i and j as a "membership pair" from the three-digit SIC, but only a small fraction of pairs are actual membership pairs. The alternative scheme which generates the same "true membership pair" percentage can be regarded as comparable with the three-digit SIC. Given the 2.05% "true membership pair", we next adjust the minimum relatedness threshold until which it generates industries with the same fraction of membership pairs as the three-digit SIC. With the threshold gradually increases, the peer with lower relatedness fraction will be excluded. The final NBPs with threshold-based approach can be called NBPs-3.

Unlike 10-K based industry schemes such as TNIC which updates every year, the NBPs do not suffer any look-ahead bias and therefore can be constructed promptly. That is, for each base firm in each month, we could identify all firms in a co-coverage news article at least once during a past rolling window in the filter news sample. However, the choice of window length as well as update frequency is a bias-variance trade-off. On the one hand, a long window can include all quarterly and annually earnings-related news observations that help to precisely estimate economically linked NBPs and better augment existing industry schemes. On the other hand, the effect of these older observations will gradually decline and are no longer newsworthy to investors. As a result, we choose a calendar year-end cut-offs to construct NBPs in the baseline setting. Later in the robustness checks, we also consider a monthly rolling forward scheme with a trailing window of 3, 6, and 9 months, respectively to study NBPs' newsworthiness feature. Next, for peer firm j of a base firm i in year k, $Frac_{ijk}$ is computed as a percentage of co-coverage firm pairwise i and j among all base firm i's co-coverage. A high relatedness between firm i and j suggests that two firms frequently appear in the same news articles, which generally should be considered as economically linked.

A.2 Newsworthy Links

We look into the time-varying feature of the NBPs. Table A1 presents the evolution of an example base firm Microsoft Corporation and its NBPs in three different periods: 1999, 2009, and 2016, respectively. As can be seen, the variation of Microsoft's NBPs is substantially large. Its highest-ranked peer consistently changes from *Netscape* as of 1999, to *Yahoo!* in 2009, and to *Linkedin* in 2016. This example demonstrates the time-varying nature of the NBPs, suggesting that a firm's peers can be identified promptly compared to other traditional industry classifications.

——————————— Insert Table A1 here ———————————

A.3 Sentence-level Co-coverage

A necessary discussion is which scheme is better to fit in our setting to identify co-coverage firms. One recent work is noteworthy: Schwenkler and Zheng (2019) employ a Name-Entity-Recognition (NER) technique to identify economic links for each sentence. For example, if General Motors Corp and Ford Motors Co appear in a sentence, two firms are likely to be economically linked. Furthermore, they conduct a series of comparisons to justify the sentence-level link is a robust alternative method compared to the article-level technique. A variety of analysis shows that the majority of the information embedded in news articles about economic links between firms appears in the same sentence. The sentence-level technique also includes 78% of all firms identified by the article-level method. Thus, it seems sensible that the sentence-level method is a robust choice to identify NBPs.

However, we use the article-level method for three reasons: First, Schwenkler and Zheng (2019) state that firms identified by the sentence-level technique tend to be small firms, traded over-the-counter and non-publicly listed. The effort to identify these firms yields no benefit in our research setting given that we focus on publicly-traded common stocks. Second, we show that our article-level NBP momentum is much stronger than the alternative sentence-level scheme using their database in an unreported test. Given our approach based on journalist-assigned subject tags build upon the expert knowledge of journalists and is computationally less burdensome than unsupervised or dictionary-based machine learning approaches, we do not further extract sentence-level firm links using their method. Third, the final output produced by the article-level technique is highly overlapped with the alternative sentence-level scheme. In an unreported test, we examine the degree of overlap between two schemes: The average overlap is 73.58%, which is quantitatively similar to Schwenkler and Zheng (2019). Collectively, our choice of an "article-level" co-coverage approach is unlikely to bias the results.

Appendix B

Tomorrow's Fish and Chip Paper? Additional Analysis

B.1 Name-ticker News Collection

Our primary data comprise news items taken from the Dow Jones Newswire Archive. All news articles are stored in an XML format with tags including headlines, dates, stock tickers and subject codes. We examine the influence of firm-level news in the U.S. equity market and the detailed procedures used to extract and quantify the news are given below.

1. Extract all news items with attached stock tickers in the Dow Jones Newswire Archive.

2. Download a list of company names and tickers from CRSP's historical name change file for all common stocks (i.e., where share code is 10 or 11) traded on the NYSE, AMEX and NASDAQ between 1979 and 2016.

3. Merge each stock ticker PERMCO (CRSP's permanent company identifier) with the company news items within the relevant period.

We now provide firm-level news summary statistics for replication purposes. In total, the procedure collects 10.6 million news observations and involves 14,079 unique firms between July 1979 and December 2016. The Archive on average produces 279,528 news items each year, with the lowest and highest numbers reported as 6,112 and 666,321, respectively. There is a large variation over the years in the number of firms mentioned in the Archive, with a minimum of 860 and a maximum of 7,746. There are, on average, 763 news articles for each firm during the entire lifespan.

To further examine the validity of the PERMCO-ticker linking table, we impose additional matching criteria roughly following Tetlock et al. (2008). Briefly, the procedure is as follows:

1. The firm name string must appear in the first 25 words of a news article including the headline.

2. The firm name must be detected at least twice in the main body of the news story.

3. News reports with fewer than 50 words are excluded.

We "tweaked" the firms' names depending on the search quality we had. This is because CRSP provides very unique and different name strings in comparison to the common names applied in the Dow Jones Newswire Database. In detail, the modifications are:

1. CRSP puts spaces between the letters in abbreviations – e.g., F D X CORP. We delete the spaces in such cases.

2. If the firm name ends with Inc, Ltd, Corp, Co. then the suffix will be removed. Note that we apply this rule to all sample firms including, for example, Apple Inc. This is necessary, but note that the keyword "Apple" will not match

irrelevant news reports related to apple (the fruit) since we are conditioning on company tickers.

3. We replaced the abbreviated words INTL, MFG, CHEM with International, Manufacturing and Chemical, respectively.

4. A name string ending with NEW or OLD to specify the company's status was tweaked to keep only the name before these words.

Overall, after imposing more strict matching criteria, our sample of news items halves to 4,530,243 non-repeated firm-level news stories and the number of sample stocks decreases to 12,072.

B.2 Google Search Volume Index Collection

Unlike Engelberg and Gao (2011), which uses a sample of Russell 3000 stocks, our sample extends to all U.S. common stocks and we also acknowledge the fact that not all stocks have a non-zero SVI. We therefore carefully deal with the potential biases and errors when retrieving the Google SVI from the Google Trends web page. The procedure is as follows:

1. Instead of using firm legal names from the CRSP historical name file as Google censor keywords, we utilise these firm names as search inputs but look into the Google auto-suggestion menu. The optimal keyword would be chosen if the category of a keyword is “company” or “corporation”. The rationale behind this is to convert firm legal names to the names commonly searched by users or mentioned in the financial press.

2. The SVI data has been scaled by Google based on the sample range. In this paper, we set a 90-day period for each data request and Google censors therefore return daily data.

3. To get rid of “noise” attention, the underlying search source is restricted to business news only. Given that investors could come from the outside of the U.S., geographical location is worldwide when downloading.

4. We improve the data quality by manually validating the firm name sought with the original one. If there is a significant difference that can be visualized, the data will be excluded from our sample.

In total, we finally obtain a 4,545-firm sample out of 7,864 targets.

B.3 Google SVI and Bloomberg AIA Attention Proxies

Although the empirical results confirm limited attention theory, we have had to assume that investors would have paid attention to the news if a company's name is mentioned. To address this issue, we further study limited-attention theory by utilising two direct attention proxies: Google Search Volume Index (SVI) and Bloomberg news reading activity (AIA, i.e., Abnormal Institutional Investor Attention). These direct attention proxies have been proposed and studied in the literature and have provided compelling evidence involving certain financial variables (Engelberg and Gao, 2011; Ben-Rephael, Da, and Israelsen, 2017). The Google SVI measures investor attention by using the aggregate web user search frequency based on the Google search engine. Google counts the number of visits for a particular key word during a specific time period. As exemplified in Figure A3, the SVI index of the keyword *Apple Technology company* rises to a peak of 100 on September 12, 2018 when a major launch event was held by Apple. Thus the SVI index brings a direct linkage between the interest of the general public and firm-level news events.

——————————— Insert Figure A3 here ———————————

Each Bloomberg terminal provides a function where the news reading activity of each firm is monitored on daily basis. Due to the financial cost and expertise required to use Bloomberg, most terminal users are from the financial services sector. In other words, Bloomberg news reading activity is highly likely to represent institutional investor attention compared to the Google search engine which tends to represent retail investor attention.

To download the Google SVI data, we use a Python web crawler to automatically send keywords (firm legal names) and retrieve the data for each firm following certain procedures detailed in Section 4.4.4, leading to a total of 7,864 U.S. firms, for which we have evidence for 4,545 firms. As for the Bloomberg AIA data, we follow the procedure by Ben-Rephael et al. (2017), in which they download the Russel 3000 stocks from the year 2010 forwards.[1] We then move to data pre-processing. Specifically, Bloomberg news reading activity data is a time-series measure of the rolling prior 30-days terminal user reading on an hourly basis. The value is 0, 1, 2, 3 or 4 if the rolling average is below 80%, between 80% and 90%, between 90% and 94%, between 94% and 96% and above 96%, respectively. Following the method in Ben-Rephael et al. (2017), the AIA measure is a dummy variable which takes a value of zero when the score is 0, 1, or 2 and a value of one otherwise. To be consistent with the Bloomberg AIA, we also assign the score for the Google SVI data based on the same method so the Google ASVI (Abnormal SVI) value will be one if the corresponding score is 3 or 4 and zero otherwise.

Using these two attention proxies, we ask whether low attention could explain stronger news predictability, particularly for SI news. To examine this, we conduct a calendar-time portfolio approach with a one trading day formation period and use the following ten trading days as a holding period for either the high or the low attention subsample. We change the research setting in this case for two reasons: first, both Google ASVI and Bloomberg AIA are reported on a daily basis and have been scaled in a proprietary manner prior to downloading. We are cautious to keep the natural data frequency in order to capture investor attention precisely. Second, due to the sample range, the period examined covers recent years where both the SI and QI effects become weaker, as discussed in Appendix B.3. The weaker effects might be partially attributed to advanced newswire transmissions such as Twitter, which speed up information dissemination. Overall, daily-basis settings can better capture the SI news effects.

[1]Bloomberg AIA data is missing for the periods 12/6/2010 - 1/7/2011 and 8/17/2011 - 11/2/2011.

In Table A4, it can be observed that Google ASVI predicts a stronger SI news effect. The difference between the low and high Google ASVI is even significant at the 1% level (with t-statistic 4.44). The coefficients for the QI news subgroup are nearly zero (all statistically insignificant). Unfortunately, we do not observe any significant evidence for the Bloomberg AIA (the proxy for institutional investor attention). The results, therefore, suggest that it is retail investors rather than institutional investors for whom genuine news slips under their radar.

——— Insert Table A4 here ———

Appendix C

Appendix Tables and Figures

Figure A1: The Sample Article of the Dow Jones Newswire Archive

Figure A1 shows a news article sample from the Dow Jones Newswire Archive. The red circle highlights the display timestamp, attached tickers, and news headline.

1. Display timestamp

2. Attached tickers

3. Headline


```
<!DOCTYPE doc SYSTEM "djnml-1.0b.dtd">
<doc msize="000012383" md5="d95677a4d902d7bb3a0e1dde1444ff57" sysId="sbknwsarchp01" destination="AW" distId="    " transmission-date="                " >
<djnml publisher="DJN" docdate="20180201" product="DN" seq="15691" xml:lang="en-us" >
<head>
<copyright year="2018" holder="Dow Jones &amp; Company, Inc." ></copyright>
<docdata>
<djn>
<djn-newswires news-source="DJDN" origin="DJ" service-id="CO" >
<djn-press-cutout/>
<djn-urgency>0</djn-urgency>
<djn-mdata brand="DJ" temp-perm="P" retention="N" hot="N" original-source="DJDN" accession-number="20180201015691" page-citation="" display-date="20180202T004542.078Z" >
<djn-coding>
<djn-company>
<c>AAPL</c>
<c>AMZN</c>
<c>GOOG</c>
<c>GOOGL</c>
</djn-company>
<djn-company-sig>
<djn-isin>
<djn-isin-sig>
<djn-industry>
<djn-government>
<djn-subject>
<djn-market>
<djn-product>
<djn-geo>
</djn-coding>
</djn-mdata>
</djn-newswires>
</djn>
</docdata>
</head>
<body>
<headline brand-display="DJ" >
Tech Giants Power to New Heights</headline>
<text>
<pre>
By Laura Stevens, Tripp Mickle and Jack Nicas </pre>
<p>
    Three of the biggest tech companies reported record quarterly financial results on Thursday as they extended their dominance over swaths of the global economy. </p>
<p>
    Apple Inc., Alphabet Inc. and Amazon.com Inc. -- with a combined market value of more than $2 trillion -- all boosted growth by broadening their reach into new areas. </p>
<p>
    Apple&apos;s revenue rose 13% to $88.29 billion, fueled by its move to increase smartphone prices behind its new flagship iPhone X, released in November for $1,000 each. The co
<p>
    Google parent Alphabet has averaged revenue growth of at least 20% over the past 32 quarters, when adjusting for currency movements, according to RBC Capital Markets analyst Ma
<p>
    Meanwhile, Amazon -- long known for prioritizing growth over earnings -- delivered a profit exceeding $1 billion for the first time as its revenue jumped 38% to $60.5 billion.
```

Figure A2: Cumulative Percentage of Total Articles

Figure A2 shows the cumulative percentage of articles with different number of attached tickers. The vertical and horizontal Dash line highlights the 95% percentile, which corresponds to six tickers.

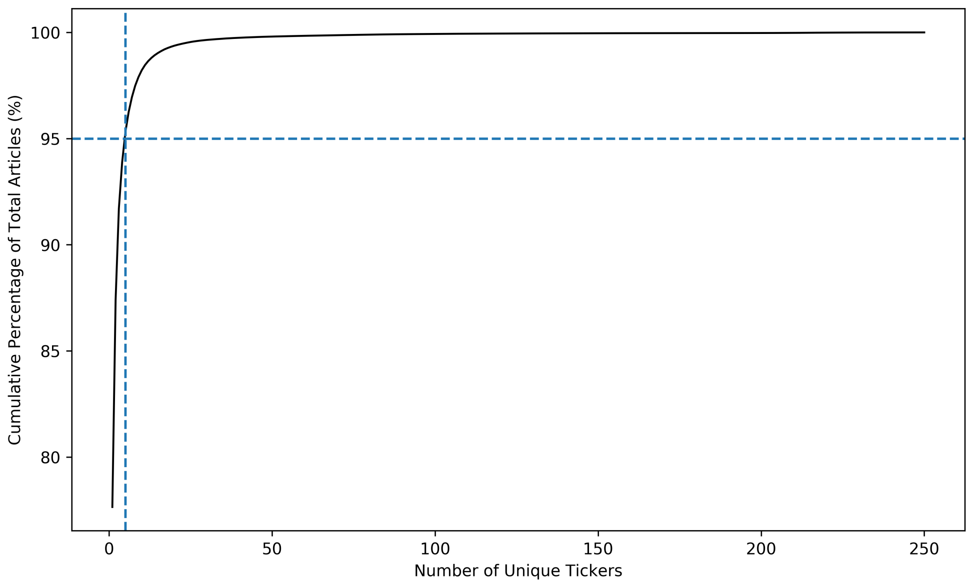

Table A1: Examples of Peer Firms and Variation Across Time

Table A1 illustrates the NBPs of the example base firm Microsoft Corporation in Jan 1999, May 2009, and August 2016 respectively. Frac refers to the fraction between base and peer firms which is computed using a monthly-rolling-forward scheme with a trailing window of 6 months.

Frac	**NBPs as of Jan, 1999**	**Frac**	**NBPs as of May, 2009**	**Frac**	**NBPs as of Aug, 2016**
26.35%	NETSCAPE COMMUNICATIONS CORP	15.79%	YAHOO INC	30.47%	LINKEDIN CORP
14.70%	SUN MICROSYSTEMS INC	15.07%	GOOGLE INC	17.17%	ALPHABET INC
13.10%	APPLE COMPUTER INC	6.22%	INTERNATIONAL BUSINESS MACHS COR		
12.18%	INTEL CORP	5.26%	ORACLE CORP		
8.11%	AMERICA ONLINE INC DEL	5.02%	AMAZON COM INC		
		5.02%	APPLE INC		

Table A2: Topic Analysis of Multiple-Ticker Articles

Table A2 reports the frequency of the Dow Jones news topics selected for this study. Each news article may have multiple subject tags assigned by Dow Jones to indicate the different topics involved.

Subject Code	**Details**	**Percentage**
N/CAC	Corporate Actions	29.05%
N/HIY	High-Yield Issuers	28.59%
N/ADR	American Depository Receipts/Shares	19.49%
N/TNM	Acquisitions, Mergers, Takeovers	15.49%
N/FF	Dow Jones Corporate Filings Alert	9.24%
N/ERN	Earnings	6.87%
N/BON	Bond News	6.72%
N/COB	Corporate Bonds	5.60%
N/ERP	Earnings Projections by Companies or Analysts	5.22%
N/ANL	Analysts' Comments Ratings of Stocks	5.14%
N/ISD	Insider Trading	3.32%
N/LWS	Lawsuits	3.10%
N/PER	Personnel Appointments	3.07%
N/REG	Bond Stock Registrations	2.30%
N/RTG	Bond Ratings Comments	2.04%
N/POV	Point of View	1.93%
N/JVN	Joint Ventures	1.90%
N/DJYY	Test Product	1.80%
N/ISS	Insider Stock Sells	1.79%
N/LAB	Labor Issues	1.79%
N/CNF	Conferences	1.62%
N/DVT	Divestitures or Asset Sales	1.56%
N/COF	Corporate Officers	1.48%
N/INI	Initial Public Offerings	1.46%
N/BCY	Bankruptcy-Related Filings	1.43%
N/MNT	Management Issues	1.42%
N/SLS	Sales Figures	1.40%
N/PCA	Partnerships/Contracts/Business Alliances	1.33%
N/DIV	Dividend News	1.32%
N/ISB	Insider Stock Buys	1.29%
N/ARG	Analysts' Ratings	1.25%
N/BRD	Boards of Directors	1.24%
N/TST	Antitrust News	1.11%
/PBP	Public-Policy Regulatory Issues	1.08%
N/LIC	Licensing Agreements	0.99%
N/MOR	Mortgages	0.99%
N/RND	Research Development	0.99%
N/OWN	Stock Ownership	0.99%
N/RCN	Corporate Restructurings	0.87%
N/FNC	Financing Agreements	0.85%
N/8K	Significant Corporate Events (SEC Form 8-K)	0.69%
N/PAT	Patents	0.67%
N/BPR	Bond Pricings	0.67%
N/SOP	Stock Options	0.65%
N/AST	Asset-Backed Securities	0.62%
N/RGU	Securities Regulations	0.61%
N/BBK	Buybacks	0.51%
N/COGV	Corporate Governance	0.50%
N/IOV	Industry Overview	0.49%
N/144	Form 144 Filings	0.49%
N/SDT	Sovereign Debt	0.48%

Table A3: Variable Definition

Variable	**Description**	**Definition**
$NbpRet$	NBPs' return	fraction-weighted NBPs' returns on a month t (excluding base firm i itself)
$IndSic3Ret$	Three-digit SIC peers' return	value-weighted three-digit SIC peers' returns on a month t (excluding base firm i itself)
$TnicRet$	TNIC peers' return	similarity-weighted TNIC peers' returns on a month t (excluding base firm i itself)
$SupIndRet$	Supplier Industry return	import-weighted supplier industry return on a month t using BEA Input-Output import matrix
$CusIndRet$	Customer Industry return	import-weighted customer industry return on a month t using BEA Input-Output import matrix
$CongloRet$	Conglomerate return	sales-weighted standalone firm constituted industry return on a month t for a multi-segment firm i
$Size$	market capitalization	price [prc] multiplying sharesoutstanding [shrout] and taking natural logarithm
BTM	Book-to-market ratio	book value of equity dividing market capitalisation and taking natural logarithm
Mom	short-term momentum	own-firm return from $t-11$ to $t-1$
PB	Price-to-book ratio	market cap[prc]*[shrout]/total common equity[ceqq]
EVS	Enterprise value-to-sales ratio	(market cap + long-term debt[dlttq])/net sales[saleq]
PE	Price-to-earnings ratio	market cap/net income before extraordinary items[ibq]
$RNOA$	Return on net operating assets	net operating income after depreciation[oiadpq]/(property, plant, and equipment[ppentq] +current assets[actq]-current liabilities[lctq])
ROE	Return on equity	net income before extraordinary items[ibq]/total common equity[ceqq]
AT	(Inverse of) Asset turnover	total assets[atq]/net sales[saleq]
LEV	Leverage	long-term debt[dlttq]/total stockholder's equity[seqq]
$SALESGROWTH$	One-year-ahead realised sales growth	(net sales one year ahead in the future - current year net sales)/current year net sales[saleq]
$RDPERSALES$	R&D expense-to-sales ratio	R&D expense[xrdq]/net sales[saleq]
SUE	Standardized Unexpected Earnings	actual earnings in quarter t minus median analyst estimate divided price in quarter $t-1$

Figure A3: Google Search Volume Index for Apple Technology

Figure A3 plots the search volume index of Apple Technology by the Google search engine between 4 July and 30 September, 2018. The raw search volumes have been scaled and displayed as time-series values between 0 and 100.

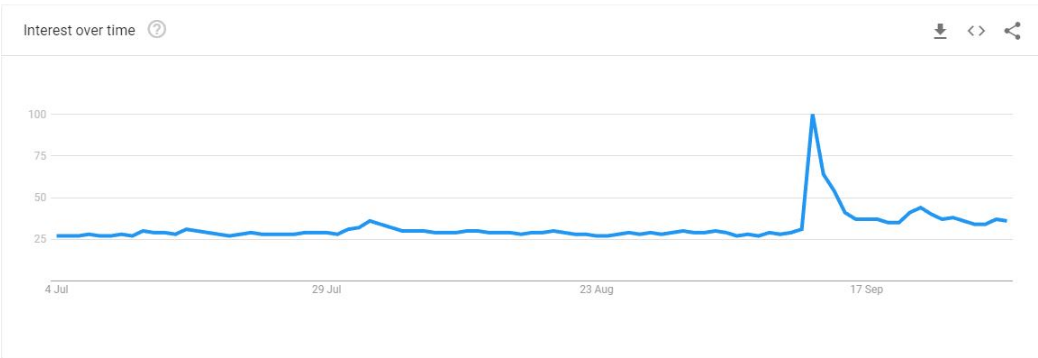

Table A4: SI and QI News under Different Investor Attention

Table A4 reports the performance of slowly incorporated news and quickly incorporated news portfolios in different information environments. We independently sort all stocks into two portfolios based on their current investor attention index. We report the "Low-High"Google ASVI ATTN, "Low-High"Bloomberg AIA ATTN spread profitability in the following ten trading days. The Google ASVI is the Google Search Volume Index, which is a proxy for retail investor attention and Bloomberg AIA measures Bloomberg users' news reading activity, which represents institutional investor attention. Coefficients are ten-trading-day cumulative returns in percentages. *t*-statistics are reported in parentheses and **, *** refers to the 5% and 1% significance levels respectively.

	Google ASVI ATTN Subsamples			Bloomberg AIA ATTN Subsamples		
	LowATTN	HighATTN	Low - High	LowATTN	HighATTN	Low - High
SINws	0.05%***	0.01%**	0.04%***	0.02%**	0.00%	0.01%
	(6.23)	(2.22)	(4.44)	(2.24)	(0.28)	(0.85)
QINws	0.00%	0.01%	-0.01%	0.00%	-0.01%	0.00%
	(-0.15)	(1.22)	(-1.12)	(-0.17)	(-0.83)	(0.35)
Slow-Minus-Quick	0.05%***	0.00%	0.05%***	0.02%	0.01%	0.01%
	(4.22)	(0.42)	(3.92)	(1.72)	(0.74)	(0.40)