COMOVEMENT AND THE NEWS

by

Travis Box

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MANAGEMENT

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

WITH A MAJOR IN FINANCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2013

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Travis Box, titled Comovement and the News and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date: April 12, 2013
Eric Kelley

_____ Date: April 12, 2013
Richard Sias

_____ Date: April 12, 2013
Scott Cederburg

_____ Date: April 12, 2013
Ronald L. Oaxaca

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: April 12, 2013
Dissertation Director: Eric Kelley

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Travis Box

ACKNOWLEDGEMENTS

I thank the University of Arizona for their research support. I am grateful to my thesis advisor Eric Kelley and the members of my dissertation committee Richard Sias, Scott Cederburg, and Ronald Oaxaca. I appreciate the helpful comments of Wayne Ferson, Paul Irvine, Angela Box and all the seminar participants at the University of Arizona Finance Department, the Financial Management Association Annual Meeting, and the Southern Finance Association Annual Meeting. I thank Paul Tetlock from Columbia University and Richard Brown and Maciek Pomalecki from Thomson Reuters Machine Readable News for data considerations. All mistakes in this article are my own.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

I introduce a novel approach for the empirical analysis of asset price comovement that relates the inter-firm textual similarity of news reports to their equity return correlation. I find that this measure of news similarity is just as important for predicting future cross-firm comovement as contemporaneous return correlation. This predictability remains after controlling for industry correlation, size, book-to-market, momentum, and price-decile correlation, index membership, and headquarters location, as well as institutional holding and analyst coverage. These results contribute to the growing literature examining the role of the media in financial markets, and provide empirical support for an alternative description of return comovement that does not depend on friction-based explanations such as "category," "habitat," or "information diffusion."

CHAPTER 1 INTRODUCTION

Understanding how the prices of various financial securities evolve in relation to each other has long been a goal of asset pricing researchers and practitioners alike. From simple linear factor models to complex arbitrage strategies, the returns of a security are commonly explained in the context of its comovement with other assets. While empirical research documents the existence of comovement in asset returns, it has provided little explanation for how the underlying cross-firm relationships present themselves to market participants and evolve over time. Such insight is needed because even when historical patterns in comovement can be identified, minor innovations in the determination of individual asset prices can transform the covariance structure of the entire market. Incorporating knowledge of these innovations is essential to forming predictions of future comovement that do not rely on these tenuous historical patterns.

The financial media serves as an information conduit by compressing a vast array of firm-specific material into a digestible news sequence investors can use to make real-time trading decisions. This paper considers the usefulness of the qualitative information in news text circulated on the Reuters Integrated Data Network for the discovery of cross-firm relationships. By analyzing the linguistic similarity in news stories written about pairs of firms, I develop a proxy for the similarity of their information environments and use this proxy to predict the firms' future stock return comovement. I hypothesize that this measure of commonality from firm-specific news text can predict future comovement that is not captured by standard asset pricing models. Consistent with my hypothesis, my main result and key contribution is that the textual similarity of the news stories written about two

firms is positively related to their future stock return correlation. Furthermore, I find this new measure of similarity is as important for predicting future price comovement between firms as contemporaneous return correlation. This predictability remains after controlling for industry, size, momentum, book-to-market, price, index membership, headquarters location as well as similarity in institutional holdings and analyst coverage.

A second contribution of this paper is the reduction in time series data requirements for estimating stock return correlations. Both in research and in practice, estimating the correlation structure has sensibly relied on a lengthy historical times series. However, the time series of a firm's stock returns are the single-dimensional output of a pricing function that contains a broad range of inputs. As this function evolves, the influence of inputs relevant to future prices may not be present in the historical return series. These inputs form the information environment of a firm, and quantifying them provides an opportunity to predict the comovement implied by the contemporaneous pricing function. Thus, the methodology introduced here can produce estimates of future comovement that do not require an abundant price history. The text of a news article written about a firm can inform multiple inputs of the pricing function; and the depth of this information can be used to amend the shortcomings of historical prices for predicting comovement.

This paper also highlights a previously unexplored role of the media in financial markets in explaining the covariance structure of the market. Earlier research has provided evidence that the word content of news publications can predict the returns of individual stocks (Loeffler, 1993) and the broader market (Tetlock, 2007). The qualitative similarity

of this content between firms, as documented in this study, is consistent with a relationship between economic forces influencing their cash flows. While human beings are often capable of recognizing these relationships, representing these associations as quantitative measures is often difficult. Many approaches to firm taxonomy, organizing firms by size, industry, book-to-market style, etc., were designed with the intent of representing this particular type of similarity. The field of linguistics provides an alternative class of methodologies for quantifying the information contained in firm news stories for use in econometric models. With many of these techniques, it is possible to measure the pairwise similarity of firm information environments and predict comovement. I also investigate which of these linguistic approaches yields the most useful description of the information environment for predicting stock return correlation.

Existing studies of stock return comovement focus on the trading frictions triggered by observable barriers and behavioral biases. Barberis, Shleifer, and Wurgler (2005) offer three such explanations—the category," "habitat," and "information diffusion" views—that could delink realized comovement from comovement in fundamental values. I consider each of these alternatives alongside the theory of proposed in this study and find that the stock return correlation resulting from similarity in news content does not simply proxy for another one of these frictions. The evidence presented below provides support for a number of these alternatives, so the theories proposed by Barberis et al. (2005) should not be cast in conflict with the predictions of this article.

Aside from the direct contribution to the understanding of return comovement, my results motivate future work in two other areas of investment research. First, the exploration of portfolio choice has undergone a revival in recent years, and the applications for these findings in that domain are profoundly evident. Box (2012) analyzes the performance of minimum variance portfolios that are constructed from predicted correlations. By dynamically adjusting to changes in firms' information environments, portfolios constructed with predicted correlations realize lower out-of-sample variance and higher return than portfolios constructed from covariances estimated from past returns. Second, the primarily theoretical study of information markets should benefit from an empirical analysis describing the link between the financial media and return comovement. If the output of the financial press is viewed as a commodity, with supply and demand determined by the potential gains from trading on information, the similarity of firms' information environments will be endogenous. Because the production of detailed idiosyncratic information is expensive, the media may focus on the areas of commonality between firms and produce qualitative information that is highly correlated. The framework described below is a starting point from which to study this endogeneity empirically.

CHAPTER 2 LITERATURE REVIEW

The related literature described below can be divided into three distinct research areas: excess comovement, information markets, and the influence of media in asset pricing. Previous analysis of the comovement in asset prices has identified a measurable gap between fundamental and return correlation, and this project attempts to reduce that disparity. Despite an extensive history, the study of information markets has remained primarily in the realm of theoretical research. This study lays the groundwork for an empirical examination of this market that is fundamental to price discovery. Finally, the impact of the financial media in asset prices has been pondered for decades, but this article is the first to focus specifically on the media's relation to return comovement.

2.1 EXCESS COMOVEMENT

The concept of excess comovement was introduced by Pindyck and Rotemburg (1990) to describe the persistent tendency of raw commodity prices to move together in excess of common effects such as inflation, or changes in aggregate demand, interest rates, and exchange rates. In subsequent work, Pindyck and Rotemburg (1993) document unexplained correlation in the equity returns of firms in unrelated lines of business after accounting for changes in current or expected future macroeconomic conditions. The authors first draw individual firms from dissimilar industries that have no vertical relations to each other, and then confirm that their normalized earnings are not significantly correlated. These criteria still allow for return correlations among the chosen firms, but only to the extent that the relationships are mediated through correlations with economy-wide earnings or discount rates. The residual correlation found after controlling for current

and expected macroeconomic conditions suggests that security prices also depend on variables unrelated to aggregate conditions. The authors offer two firm characteristics, company size and institutional ownership that seem to explain some portion of the documented excess comovement.

More recent work by Barberis, Shleifer, and Wurgler (2005) uses additions to the S&P 500 to distinguish between two broad theories of return comovement. The traditional theory, derived from economies without frictions and with rational investors, holds that comovement in prices reflects comovement in fundamental values. In the alternative view with frictions or irrational investors, and in which there are limits to arbitrage, correlation in returns is delinked from comovement in fundamentals. Within this alternative friction- or sentiment-based explanation of comovement, the authors discuss three specific views of comovement that can be described in these terms. The category view, first introduced by Barberis and Shleifer (2003), argues that investors group assets into categories to simplify portfolio decisions. If the investors using categories are noise traders with correlated sentiment, then as they move from one category to another, their coordinated demand induces common factors in the returns of assets in the same category, even when these assets' cash flows are uncorrelated. Next, the habitat view starts from the observation that many investors choose to trade only a subset of all available securities. As these investors' risk aversion, sentiment, or liquidity needs change, they alter their exposure to the securities in their habitat. Such habitats could arise because of transaction costs, international trading restrictions, or lack of information. Finally, the information diffusion view argues that information is incorporated more quickly into the prices of some stocks

than others due to some market friction. In this view, there will be a common factor in the returns of stocks that incorporate information at similar rates. Ultimately, the authors find that a stock's beta with the S&P 500 goes up after being added to the index lending support to the alternative friction- or sentiment-based view.

The strategy advanced by Barberis, Shleifer, and Wurgler (2005) has inspired an upsurge in research focusing on the relationship between stock return comovement and firm characteristics. Pirinsky and Wang (2006) document strong comovement in the stock returns of firms headquartered in the same geographic area that is not explained by proxies for economic fundamentals. Green and Hwang (2009) find that stocks undergoing splits experience an increase in comovement with low-priced stocks and decrease in the comovement with high-priced stocks. This price-based comovement is not explained by firm size, or changes in liquidity or information diffusion, lending support for the role of category investing proposed by Barberis and Shleifer (2003). Finally, Boyer (2011) finds that economically meaningless index labels, such as Value and Growth, cause stock returns to covary in excess of implied fundamentals. I examine each of these alternatives alongside my proposed hypothesis to isolate the effects of comovement explained by similarity in firm's information environment from comovement that is driven by market frictions.

The friction-based sources of unexpalined comovement described above, S&P 500 membership, geographic proximity, stock price, and book-to-market style, are all documented using a similar methodology. After an arbitrary change in a firm characteristic, the stock return of a candidate firm is regressed on an index created from all firms sharing

that characteristic. A signficant change in the index coefficient, before and after the event, suggests a shift in comovement resulting from a change in firm characterisitcs that are unrelated to economic fundamentals. If all other potential changes in fundamentals are accounted for, these findings support the friction- or sentiment-based view of excess comovement. However, if details of the firm's information environment also change during the event, the increase in comovement might actually be the result of a change in fundamental correlation. Furthermore, my methodology focuses on explaining the return correlation coefficient between firms, not the individual firm return series; so the potential for inferences to be affected by omitted systematic factors is reduced. A similar approach, concentrating on the return correlation coefficient, is also used by Israelsen (2010) who finds support for the role of correlated information in explaining excess comovement.

2.2 INFORMATION MARKETS

Understanding the role of correlated information in stock prices has benefited from two recent theoretical projects that focus on the workings of information markets. Veldkamp (2006a) builds a model from the observation that information is fundamentally distinct from other goods because of a high fixed cost and a near-zero cost of replication. This information production technology, coupled with free entry in the information market, results in information prices that decline as demand rises. By extending the model of Grossman and Stglitz (1980), Veldkamp is able to explain media frenzies, or an abundance of information about a single asset, that cause increases in price and price volatility. Related work (Veldkamp 2006b) introduces information markets that generate high price covariance within a rational expectations framework. When information is costly, rational

investors only buy information about a subset of the assets. If investors price assets using a common subset of information, news about one assset affects the other assets; causing price comovement. At a minimum, these models motivate the use of a firm's information environment, as in the methodology described below, to predict return correlation.

2.3 MEDIA AND ASSET PRICING

The role of financial media in determining asset prices has been of interest to researchers for some time. Using analyst predictions published in the popular *Wall Street Journal* "Dartboard" column, Barber and Loeffler (1993) provide some of the earliest evidence that the financial press can have a signficant effect on subsequent stock returns and trading volume. Tetlock (2007) constructs a pessimism measure using principal components analysis from the text of another *Wall Street Journal* column, "Abreast of the Market," that is capable of predicting future aggregate market prices. Both studies however, find evidence of subsequent return reversals, especially in smaller stocks; which is interpreted as support for a sentiment-based theory of the relationship between media and the financial markets. More recently, the research in this area has expanded its focus to a much broader universe of published news. Using an archive similar to the one studied in this article, Tetlock, Saar-Tsechansky, and Macskassy (2008) find that the fraction of negative words in firm-specific news stories can, in fact, predict changes in future earnings. Fang and Peress (2009) find that stocks with no media coverage earn higher returns than stocks with high media coverage even after controlling for well-know risk factors suggesting that the breadth of information dissemination affects stock returns. Engelberg and Parsons (2011) analyze all earnings announcements of S&P 500 Index firms and find

that local media coverage strongly predicts local trading, after controlling for earnings, investor, and newspaper characteristics. They also find that local trading is strongly related to the timing of local reporting, implying a causal impact of the media on the financial market outcomes. Finally, Tetlock (2011) shows that while stock returns may be less responsive to stale news, individual investors may trade more aggressively on stale news leading to subsequent return reversals.

While the aforementioned articles have focused on return predictability, my analysis will expand the financial media literature by focusing on the second moment of stock returns. Much of the previous research has examined the direct relationship between firm-specific news and own-firm stock returns, I will focus on how the fundamental relationships between different companies are represented in the media. This perspective may allow for some alternative interpretations of the results highlighted above that do not depend on sentiment-based or purely behavioral theories. In any case where we would expect trading to be dominated by individual investors (smaller firms, low institutional holdings), we should also expect the flow of information to be slower and less detailed. The propositions tested in this paper would suggest that investors rely on signals about firms with similar information environments whenever detailed news about the firm of interest is not available. When quality company-specific news finally does enter the market, investors have the opportunity to reevaluate that firm's fundamental correlations with other firms, however the arrival rate of such quality information may be very low.

CHAPTER 3 DATA AND METHODOLOGY

The firm universe for this study consists of all domestic common stocks trading on the NYSE, NASDAQ, and Amex exchanges with CRSP share codes 10 or 11. Firm price and shares outstanding are used to calculate market values on the last trading day of each year from 2002 to 2010, and the 4,000 largest firms are included in the sample for the following year. This constraint excludes 1,178 firms from the sample in 2003, but is not binding in recent years due to gradual decline in the number of publicly traded firms.

3.1 NEWS DATABASE

The news text comes from the Thomson Reuters NewsScope Archive, a historical database of Reuters and select third party news stories. The Archive is derived from the Reuters Integrated Data Network (IDN) news feed and consists of the IDN message stream which communicates news to client workstations. News stories are transmitted across the IDN in smaller pieces called "takes." Each observation in the archive represents a take, and multiple takes with a common id number can be combined to recreate a story. In addition to the raw story text, each observation contains a field listing all of the tickers for the firms mentioned in the take. Because ticker changes are not uncommon, a list of active tickers is created for the universe of firms each day of the sample period using the CRSP Names History file. This ticker list is then used to extract the relevant takes from the Archive.

A variety of additional filters are necessary for the construction of an appropriate firm-specific news corpus. The process described above results in a collection of news stories that mention a firm from the universe at least once in the text. However, just because a firm

is mentioned in a particular take does not mean that the majority of the text is relevant. Thomson Reuters also provides a related product known as News Analytics that contains proprietary scores for, among other things, the relevance of a particular take to each of the firms mentioned. This relevance measure is a real valued number bounded between 0 and 1 indicating the relevance of the take for the firm in question. It is calculated by comparing the relative number of occurrences for the firm with the number of occurrences for other organizations and commodities within the text of the take. A take is only retained for a firm if the relevance score is at least 0.5. Any firm whose name is mentioned in the title of the article is automatically given a relevance score of 1. For stories with multiple companies mentioned, the company with the most mentions will have the highest relevance, but stories with more than two firms receiving relevance scores higher than 0.5 are rare. Also, the news archive draws on stories written from all over the world in many different languages, but only stories written in English are retained. Finally, at the end of each trading day, a particular type of story is frequently broadcast over the IDN that is only related to exchange order imbalances and contains no fundamental information, and barely any text, about the firms. All of these stories are identified in the News Analytics database with the genre type "IMBALANCE," and are also filtered from the sample.

Table 1: Summary statistics for the news archive

Panel A describes the daily distribution of sample news takes, stories, and firms taken from the Thomson Reuters NewsScope Archive. The firm universe consists of all domestic common stocks trading on the NYSE, NASDAQ, and Amex exchanges with CRSP share codes 10 or 11. All takes with a relevance score, taken from Thomson Reuters News Analytics, above 0.5 for at least one sample firm are included in the distribution. Panel B describes the distribution of total words that are relevant to each sample firm across all six month periods from 2003 to 2011.

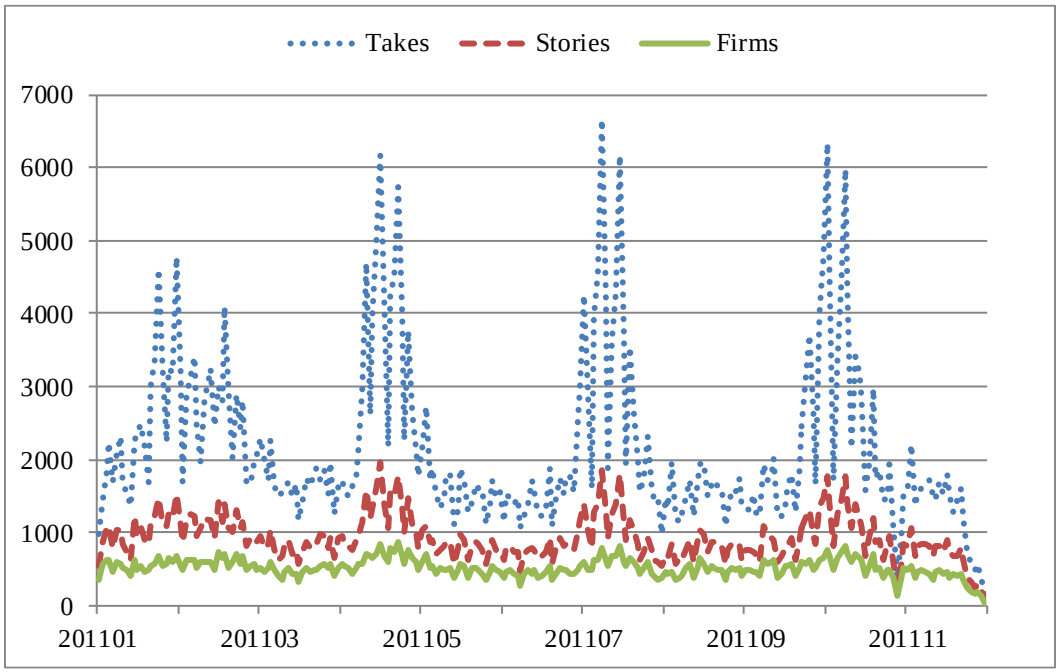
Panel A: Daily news frequency distribution								
	Mean	P5	P10	P25	P50	P75	P90	P95
Takes	1,744.5	618	808	1,118	1,507	2,013	3,145	3,898
Stories	852.7	407	526	660	825	1,003	1,251	1,417
Firms	490.5	259	335	410	492	572	659	710
Panel B: News sample word distribution by firm								
Ending Year/Month	Mean	P5	P10	P25	P50	P75	P90	P95
200306	7,712.6	0	0	0	2,848	9,113	18,437	30,049
200312	7,472.9	0	0	0	3,016	9,072	18,305	29,381
200406	7,795.6	0	0	0	3,798	9,701	18,614	27,769
200412	7,567.0	0	0	0	3,852	9,395	17,504	27,343
200506	10,065.0	0	0	0	5,369	12,340	22,891	35,736
200512	9,950.2	0	0	0	5,727	12,015	21,992	34,104
200606	10,772.6	0	0	0	6,178	13,231	23,445	36,621
200612	10,281.1	0	0	0	6,069	12,478	22,056	35,401
200706	11,132.7	0	0	268	6,596	13,071	23,756	36,842
200712	10,584.2	0	0	0	6,364	12,474	22,501	35,889
200806	11,593.8	0	0	554	6,864	13,195	23,177	37,205
200812	10,915.4	0	0	383	6,481	12,212	21,685	33,716
200906	10,174.6	0	0	196	5,533	11,104	20,590	32,761
200912	11,004.8	0	0	108	6,113	12,448	23,528	35,702
201006	11,319.7	0	0	533	6,592	12,594	23,654	36,524
201012	10,644.8	0	0	268	6,439	12,435	22,626	32,228
201106	12,530.4	0	0	202	7,135	14,208	26,850	40,445

Summary statistics for the sample of news stories are listed in Panel A of Table 1. On average, the sample contains 1,745 takes representing 853 unique stories and 491 unique firms each trading day. To conform with the price history in CRSP, all takes broadcast on the IDN after the market close will be considered part of the news flow for the next trading day. Figure 1 graphs the number of takes, stories, and firms each included in the sample

each trading day of the year 2011, though all years have a similar pattern. The most obvious feature of the three time series is the effect of earnings season on the flow of company specific news, recognizable by the four distinct peaks throughout the year.

Figure 1: Daily frequency of takes, stories, and firms for 2011

The daily frequency of individual news takes appearing on the Reuters Integrated Data containing information relevant to a particular firm is pictured in blue. Multiple takes with the same matching identification numbers are used to form stories, and the daily frequency of unique stories is pictured in red. The number of individual firms mentioned in these stories each day is pictured in green.



3.2 TERM-DOCUMENT MATRIX

The collection of news stories described above contains a large quantity of qualitative information about the firms in the sample. The basic object of this analysis is the term-document matrix, a mathematical representation of the frequency of terms that occur in a

collection of documents. The intuition behind this methodology is as follows: if the frequencies of words used in the news about different firms is similar, then the qualitative information contained in those stories is also similar. As an example, if the news about two firms frequently uses words like interest, debt, and default, it may be the case that both firms are having some difficulty in the debt markets. Even if these firms are in entirely different industries and have entirely different market capitalizations, a newswire subscriber might expect some additional comovement between these firms' equity price innovations relative to firms whose newswires do not mention these words. Other firms might be susceptible to rising energy prices and have high frequencies of words like oil, energy, and China. Depending on the article, however, the word China might be used in a discussion of manufacturing costs that have little to do with energy demand. Thus it remains an empirical question as to whether the frequencies of different words would accurately capture the qualitative information that is relevant to equity price comovement.

In a term-document matrix, columns correspond to the documents (firms) in the collection and rows correspond to the terms (words). For each period of interest, all takes related to a specific firm are aggregated into one master firm document. The frequencies with which terms appear in this document are recorded as integers in a firm's term-document vector. Combining these vectors for all sample firms produces the term-document matrix for the period. When constructing this matrix, all letters are changed to lower case, summary information about the authors is removed, and all tickers and numbers are deleted. The punctuation is removed with the exception of dashes between words and apostrophes between conjunctions. This should preserve the appropriate interpretation for

tokens like “on-the-run” and “aren’t.” Finally, the individual words in own firm names, as listed in the CRSP Names History file, are also removed from each firms document to avoid arbitrary associations caused only by these words. For instance, the Eaton Corporation, a diversified power management company, might be related to Eaton Vance Investment Managers, a mutual fund manager, for no other reason than the coincidence of a shared name.

After these filters are complete, a grand dictionary is created that contains every word used in at least two firm documents, and the term-document matrix is created with the columns corresponding to the firms and the rows corresponding to the elements of the grand dictionary. Each element of the matrix is the integer valued number of times a particular word appears in a firm’s document. This basic methodology is commonly used by search engines to measure the similarity between search queries and web pages on the internet (Langville and Meyer 2006). Many firms are rarely mentioned in the newswires, however, so making inferences about the economic signals relevant to their payoffs is difficult without a significant volume of text. To choose the appropriate formation period, term-document matrices are constructed using 1, 3, 6, and 12 month spans. All news reported after the close of trading on the last day of the formation period will be included in the following period’s term-document matrix to better align the news data with CRSP.

Figure 2: Number of firms with relevant news takes over different formation periods

After compiling all relevant takes from the Thomson Reuters NewsScope Archive, the number of unique firms appearing in term-document matrices formed over 1, 3, 6, and 12 month horizons from 2003 to 2012 are pictured below.

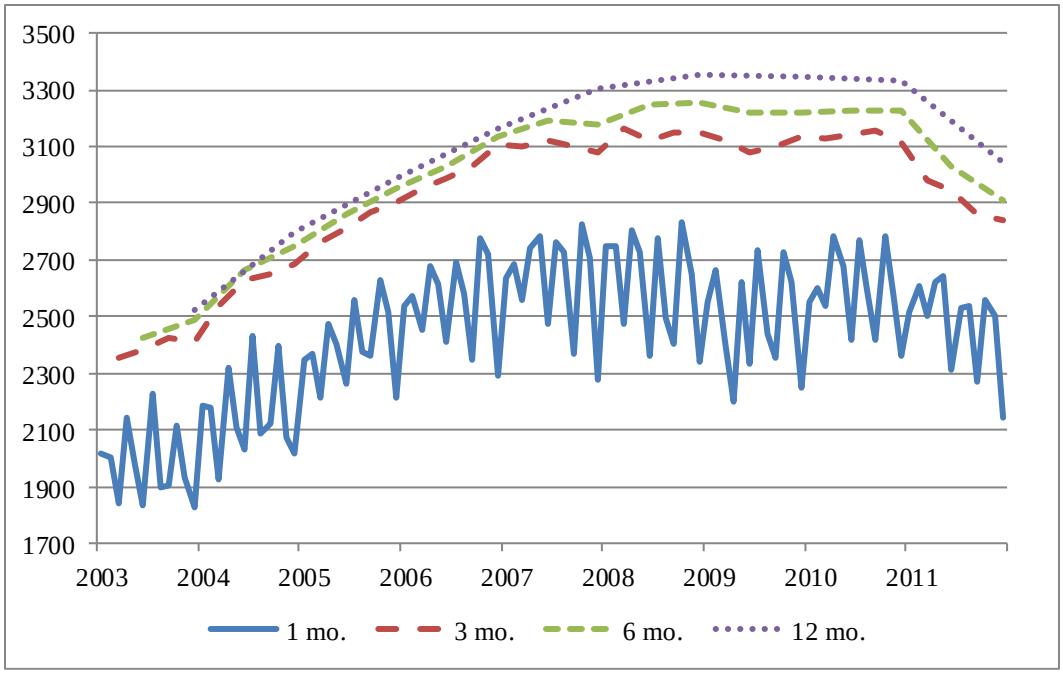


Figure 2 shows the number of firms that would be included in the sample if the formation period ended on the date listed on the horizontal axis. For 1 month term-document matrix formation periods, the number of firms in the matrix is greatly affected by the earnings season. The figure implies that a sizeable contingent of firms are only mentioned in the news around earnings releases, so any formation period that did not span these events would have an extremely volatile sample size. The 3, 6, and 12 month formation periods remove the effect of earnings season from the data, so the 6 month interval is chosen to strike a balance that would allow for observing discrete changes in the information environment while still including newswires pertaining to the broadest universe of firms. All subsequent analysis is also performed, but not tabulated, using 3

month formation periods. Despite the shorter formation period, all results are qualitatively similar, though the economic impacts are somewhat attenuated.

Panel B of Table 1 provides summary statistics on the number of eligible words written about each of the sample firms for each term-document matrix formation period. In every 6 month span, at least half of the 4,000 sample firms have some news written about them. The total volume of firm-specific news has risen steadily over the sample period, but this seems to be driven primarily by an increase in the depth of coverage for the lower profile firms. For instance, the number of words written about the median firm, by news coverage, has increased 151% while firms in 95th percentile have only experienced a 35% increase since the first half of 2003. Thus, recent newswire subscribers have benefited from a significantly broader scope of coverage relative to previous years.

Overall, the textual analysis used for this study most closely resembles the techniques used in Hoberg and Phillips (2010a and 2010b). The field of linguistics refers to this type of analysis, dissecting a document by examining only word frequencies, as the bag-of-words model (Bilisoly 2008). Because any random permutation of the text produces the same frequencies as the original version, word order is irrelevant. Obviously, this permutation removes information from the text, but it allows for a tractable comparison of the news content for each firm.

This raw term-document matrix may possess some undesirable qualities, however, that hinder a comparison between firms based on information content. High-frequency words are often used for grammatical purposes. For example “the” is used with nouns to

emphasize a particular instance of that noun, as in "the picture." Also, function words like "that," "this," and "is" are frequent, but add little to the information content of the text. The most common method of dealing with these function words is by simply removing them with a stop list[1]. The stop list approach requires a universally accepted list of these function words, and the list used in this study is included in the PERL Lingua module available for download on CPAN.

The term-document matrix itself can be thought of as the raw quantitative data for the empirical analysis. To make comparisons about the information environments of different firms however, the pairwise similarity of firm news must be computed explicitly. The most common measurement of document similarity is the cosine of the angle θ_{ijt} between the term-document vectors $\mathbf{f}_{it}$ and $\mathbf{f}_{jt}$ of firms i and j during period t:

$$\pi_{ijt} = \cos\theta_{ijt} = \frac{\mathbf{f}_{it}^{\mathrm{T}}\mathbf{f}_{jt}}{|\mathbf{f}_{it}||\mathbf{f}_{jt}|} \qquad 1$$

The angle θ_{ijt}, and thus the cosine of the angle, between the term-document vectors of two firms is higher when those vectors share similar proportion of words. If the text about a

[1] An alternative to the standard stop list approach is to use mathematical functions of the frequencies in the term-document matrix that correct for the abundance of these function words (Manning and Schütze 1999). The most common functional approach weights the term (word) frequencies by the inverse document frequency:

$$IDF_w = \log\frac{N}{N_w}$$

where N is the total number of documents (firms) in the collection and N_t is the total number of documents containing word w. If a word appears in nearly every document, its weight IDF_w will be very close to 0 and have little effect on the analysis. This method essentially produces a stop list of function words endogenously, negating the need for a universally accepted list of stop words. All subsequent analysis is performed using both the stop list and inverse document frequency approach. Though both methodologies produced economically meaningful and significant results, the stop list approach was better able to predict future price comovement in all specifications so only those results are reported.

particular firm contains none of the same words, the pairwise cosine similarity π_{ijt} will be 0, and if the documents have identical proportions of words, the cosine similarity will be 1. Cosine similarity is often referred to as the "un-centered correlation" because when the term-document vectors are centered about their firm-specific average word frequencies, the cosine similarity would exactly equal the Pearson correlation between the two vectors. All subsequent analysis is performed using both cosine similarities and correlations, but because the results are nearly identical in both specifications, only the cosine similarity results are reported.

CHAPTER 4 ESTIMATION METHODOLOGY

The ensuing analysis will measure the relationship between the information environment similarity of two firms i and j, measured by cosine similarity π_{ijt}, and their future stock price comovement, measured by pairwise Pearson daily return correlation ρ_{ijt+1}. Before proceeding, there is one additional control variable necessary to account for the text of news takes that mention more than one firm. If Lowes and Home Depot are always mentioned in the same take, their term-document matrices will be identical because all of the text written about them would be from the same source. Though their returns may be highly correlated, the positive relationship observed between the similarity of their information environments and their price comovement would not be useful for predicting a similar relationship between firms that were never mentioned in the same take. The following measure will account for this take correlation:

$$\rho_{ijt}^{take} = {N_{ijt}^{take}} \Big/ \sqrt{N_{it}^{take} N_{jt}^{take}} \qquad 2$$

where N_{ij}^{take} is the number of takes that mention both firms i and j in a period t, and N_{it}^{take} and N_{jt}^{take} are the number of takes mentioning firms i and j respectively. If both firms are mentioned together in every take, ρ_{ijt}^{take} will be 1, and if they are never mentioned in the

same take, ρ_{ijt}^{take} will be zero. All subsequent analysis will rely on the following basic regression model[2]:

$$\rho_{ijt+1} = \beta_0 + \phi_1\rho_{ijt} + \beta_1\omega_{ijt}^{pos} + \beta_2\rho_{ijt}^{take}\omega_{ijt}^{pos} + \beta_3\pi_{ijt}\omega_{ijt}^{pos} + \varepsilon_{ijt+1} \qquad 3$$

where ω_{ijt}^{pos} is a dummy variable indicating that both firms had some positive level of news coverage in period t. This variable is necessary to distinguish when news similarity is 0 because the firm information environments were empirically unrelated or because one of the firms just did not have any news coverage over the entire period.

As written, the disturbances estimated from Equation 4 would contain some unfavorable structure. Like most panel datasets, all of the observations occurring in time period $t + 1$ should be related to each other because of immeasurable common factors generating the stock returns. Furthermore, the pairwise return correlation ρ_{ijt+1} at time $t + 1$ for firms i and j is almost certainly related to the return correlation of the same firm-pair at all other points in their time series. The addition of firm-pair fixed effects to the specification could correct for any omitted variable bias associated with this particular relationship; however the disturbances are still likely to have structure induced by the firm-specific relationships. The return correlation for firms i and j is mechanically related to the return correlation for firms i and k because the same price information about firm i is included in the correlation estimates ρ_{ijt+1} and ρ_{ikt+1}. Finally, Equation 3 attempts to

[2] Contemporaneous news similarity measures are used to explain future, as opposed to contemporaneous, return correlation to avoid endogeneity caused by news reports that are only responding to previous return correlation in the same period.

measure the change in future return correlation that would result from a hypothetical change in contemporaneous news similarity. It is possible that contemporaneous changes in news similarity are actually responses to changes in return correlation earlier in the same period. To isolate only the effect of news similarity, the specification should also account for the current period's return correlation. This argument is similar to the motivation for the familiar test of Granger (1969) causality, which uses a specification containing both lagged dependent and independent variables to isolate the individual sources of dependence. These relationships motivate the following model with a lagged dependent variable and three different types of fixed effects included[3]:

$$\rho_{ijt+1} = \beta_0 + \phi_1\rho_{ijt} + \beta_1\omega_{ijt}^{pos} + \beta_2\rho_{ijt}^{take}\omega_{ijt}^{pos} + \beta_3\pi_{ijt}\omega_{ijt}^{pos} + \alpha_t + \gamma_{i\wedge j} + \delta_{i\vee j} + \varepsilon_{ijt+1} \quad (4)$$

where α_t is a fixed effect for time, $\gamma_{i\wedge j}$is a fixed effect for a unique pair of firms, and $\delta_{i\vee j}$ is a fixed effect for each individual firm i and j. Unfortunately OLS estimation of Equation 4 would still be biased and inconsistent. Since the variables ρ_{ijt+1} and ρ_{ijt} are both functions of the fixed effects $\gamma_{i\wedge j}$and $\delta_{i\vee j}$, these parameters will be correlated with the disturbances[4].

[3] Empirically, the introduction of a lagged dependent variable may subsume the firm-pair fixed effects.

[4] In time period t, ρ_{ijt} s determined by a linear function with $\gamma_{i\wedge j}$and $\delta_{i\vee j}$ on the right-hand-side. Because ρ_{ijt} also appears in the function generating ρ_{ijt+1} alongside $\gamma_{i\wedge j}$and $\delta_{i\vee j}$, the time independent fixed effects are correlated with the disturbances (Greene, 2008).

Rewriting Equation 4 in terms of first differences will remove these correlated time independent fixed effects:

$$\Delta\rho_{ijt+1} = \phi_1\Delta\rho_{ijt} + \beta_1\Delta\omega_{ijt}^{pos} + \beta_2\Delta\rho_{ijt}^{take}\omega_{ijt}^{pos} + \beta_3\Delta\pi_{ijt}\omega_{ijt}^{pos} + \alpha_t + \Delta\varepsilon_{ijt+1} \quad (5)$$

Notice that the coefficients ϕ_1, β_1, β_2, and β_3 are unchanged by this transformation and that the number of parameters required to estimate has declined drastically with the removal of the fixed effects. Instrumental variables estimation is required because ρ_{ijt} is used to calculate both $\Delta\rho_{ijt+1}$ and $\Delta\rho_{ijt}$, so OLS estimation of Equation 5 would still lead to biased estimates (Anderson and Hsiao, 1981).

Arellano and Bond (1991) argue that this procedure for dynamic panel estimation would produce consistent but not necessarily efficient results. After taking first differences, they propose using all past information about the dependent variable and all exogenous information about the independent variables as instruments, then estimating the model with the generalized method of moments. For instance, when trying to predict the period 4 change in return correlation $\Delta\rho_{ij4}$, the lagged levels ρ_{ij1} and ρ_{ij2} and the lagged difference $\Delta\rho_{ij2}$ are mechanically uncorrelated with $\Delta\varepsilon_{ijt+1}$ and can be included as instruments. Moving forward through time, the number of potential instruments, all historical observations of these levels and differences, continues to grow[5]. The past, present, and

[5] The dimensions of the instrument matrix $\boldsymbol{Z}$ are much larger than typically found in other instrumental variables estimations. Individual matrices $\boldsymbol{Z}_{ij}$,with dimensions $(T-2)\times L$, are constructed for each firm-pair. The number of columns L is a function of total periods T and the number of instruments chosen for

future levels and differences of the independent variables, in this case ω_{ijt}^{pos}, $\pi_{ijt}\omega_{ijt}^{pos}$ and $\Delta\rho_{ijt}^{take}\omega_{ijt}^{pos}$, could potentially be included as instruments if they are also believed to be uncorrelated with $\Delta\varepsilon_{ijt+1}$ (Greene, 2008). The Arellano and Bond (1991) methodology is capable of dealing with unbalanced panels, so firm-pairs do not need a lengthy time series to be included in the estimation.

Not only can this approach help to identify the determinants of future return correlation, but practioners should enjoy the limited data requirements neccesary to generate predictions. To forecast the next period's return correlation between two firms, only the current and previous period's return correlations are required to generate an estimate. Thus, the correlation of a new firm could be included in the development of a trading strategy relatively quicky, instead of waiting years for the data neccesary to estimate a sample covariance matrix.

estimation. The matrix $\boldsymbol{Z}_{ij}$ has the following basic structure if only lagged levels of the dependent variable are used as instruments:

$$\boldsymbol{Z}_{ij} = \begin{bmatrix} \rho_{ij1} & 0 & \cdots & 0 \\ 0 & \rho_{ij1}, \rho_{ij2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \rho_{ij1}, \rho_{ij2}, \rho_{ij3} \cdots \rho_{ij(T-2)} \end{bmatrix}$$

The addition of lagged changes in the dependent variable or exogenous independent variables to the set of instruments is straightforward. When the transpose of $\boldsymbol{Z}_{ij}$ is premultiplied by the matrix of explanatory variables relevant to firms i and j, the explanatory variables only interact with the appropriate set of instruments in each time period.

CHAPTER 5 EMPIRICAL TESTS

The daily Pearson return correlation ρ_{ijt} and the cosine similarities of the term-document vectors π_{ijt} are calculated for each six month period; the first ending in June of 2003 and the last ending in December of 2011. Because the Arellano and Bond (1991) methodology uses first differences, only firm-pairs with at least two consecutive return correlation observations are retained. The resulting sample contains 117,693,376 firm-pair-period observations that include 13,874,350 unique firm-pairs.

The sheer size of this panel makes the estimation of Equation 5 computationally infeasible. For the subsequent regression analysis, 125,000 firm-pairs are randomly selected from the initial sample of 13,874,350, and then all of the time series observations from those firm-pairs included in the estimation of Equation 5. Some firm-pairs might only exist for a few periods in the beginning or end of the time series, and others might have usable observations over the entire sample period. This means that the number of eligible time series observations that a firm-pair may have does not affect the likelihood of its inclusion in the final sample which ultimately contains 1,062,422 firm-pair-period observations. When viewed in terms of individual firm prices and news text, this sampling methodology still makes use of nearly all available firm-specific information in the news text and the CRSP price data. For the results reported below, the final sample includes individual price and news text for all of the possible 5,676 firms that stay in the sample longer than 1 period. Thus firms of all different size, age, and, most importantly, news coverage profile are included in the final estimation.

A series of related projects have studied the determinants of return correlation in a truncated sample of firm-pairs. Israelsen (2012) and Muslu, Rebello, and Xu (2012) examine the effect of correlated analyst coverage on yearly stock-price comovement. Chen, Chen, and Li (2012) use historical and predicted return correlation to construct equity pairs, and then show that a trading stratgey based on the price convergence of these pairs can generate abnormal returns. All three studies analyze a sample truncated by firm size, data availability, or index membership to arrive at a sample size comparable to the one used in this project. Truncation of this sort would be inappropriate for a study such as this one that is intended to measure the relationship between the similarity of a firm's information environment and its comovement with other firms. Not only do these truncated samples consist primarily of firms that are very large, but they are also likely to be extremely well covered by the financial press (Fang and Peress 2009). Abbreviating the sample with a characteristic so highly correlated with news coverage could result in biased estimates of the relationship of interest.

The presentation of the empirical results will utilize the theoretical framework provided by Barberis, Shleifer, and Wurgler (2005) as an organizational tool. In addition to the relationship of interest, future price comovement and the similarity of contemporaneous information environments, a variety of additional relationships, suggested by Barberis et al. (2005) and other related projects, will also be included. Tests of the information diffusion view, explained below, will include explanatory variables related to the amount of news coverage a firm-pair receives. The specifications testing the fundamentals view will include explanatory variables accounting for firm beta, size, book-

to-market, momentum, and industry. The category view will be tested with S&P 500, Value, and Growth Index membership and a variable related to stock price. Finally, the habitat view will be tested with measures of a firm's institutional ownership and headquarters location. Of course, these partitions are not entirely consistent with the theory as laid out by Barberis et al. (2005). For instance, they argue that comovement resulting from a firm's inclusion in the S&P 500 could be explained by the information diffusion, category, or habitat views. The partitions used in this study are adapted to ease the presentation of the results.

5.1 NEWS SIMILARITY AND THE INFORMATION DIFFUSION VIEW

The first tests will examine the degree to which future return correlation is explained by commonality in the contemporaneous information environments between firms. The economic impact of changes in the variable measuring information environment similarity, $\pi_{ijt}\omega_{ijt}^{pos}$, will be positive and significant if the information produced by the financial press can accurately predict future stock price comovement. The alternative explanations of comovement mentioned above will be tested alongside the main hypothesis of this study to ensure that news similarity measures a unique source of comovement. The explanations provided by Barberis, Shleifer, and Wurgler (2005) are not necessarily mutually exclusive alternatives for explaining comovement, but rather controls necessary to identify the specific relationship between news similarity and return correlation.

The information diffusion view states that, due to some market friction, information is incorporated more quickly into the prices of some stocks than others. In this view, there

will be a common factor in the returns of stocks that incorporate information at similar rates. The volume of text printed in the financial press should serve as an adequate proxy for the diffusion of information. According to the information diffusion view, two firms with large volumes of text should comove more with each other than with lower profile firms. This motivates the following specification:

$$\rho_{ijt+1} = \phi_1\rho_{ijt} + \beta_1\omega_{ijt}^{pos} + \beta_2\omega_{ijt}^{P50} + \beta_3\rho_{ijt}^{take}\omega_{ijt}^{pos} + \beta_4\pi_{ijt}\omega_{ijt}^{pos} + \beta_5\pi_{ijt}\omega_{ijt}^{P50} + \varepsilon_{ijt+1} \quad (6)$$

where ω_{ijt}^{P50} is a dummy variable indicating that both firms had a volume of text, measured in total words, above the median for period t, and the difference operator Δ and the time series fixed effects α_t are implied but omitted for brevity. Positive and meaningful coefficients on ω_{ijt}^{pos} and ω_{ijt}^{P50} would lend support to the information diffusion.

At a minimum, the similarity of news flow should be related to the comovement of returns in excess of the risk free rate. Other authors analyzing return correlation have instead focused on some measure of excess comovement, but the field of finance is lacking in a widely accepted definition of what constitutes "excess." Ledoit and Wolf (2003) and Bekaert, Hodrick, and Zhang (2009) analyze covariance in the context of risk based models; thus excess comovement could be approximated by the correlation between traditional factor model residuals. Because the goal of this study is to identify period t news similarity's relationship with period $t+1$ excess comovement, the factor coefficients are estimated in period $t+1$. If the news similarity measure is able to predict

residual correlation even after the factor model coefficients have adjusted to the next period's values, then this variable must contain valuable information about price comovement that systematic factors cannot explain. Thus, residuals from the traditional market model estimated within a particular period are used to calculate that span's "excess" return correlation[6].

[6] If abnormal returns, calculated using factor model coefficients estimated in a previous period, were used instead of within period residuals, it would not be possible to accurately measure the relationship between news similarity and excess comovement. Using historical estimates of the factor model coefficients could induce measurement error into the calculation of excess return correlation. It would be difficult to determine whether a positive and significant coefficient on the news similarity measure indicated that the variable could predict excess price comovement or if it could just predict future changes in the factor model coefficients.

Table 2: Regression summary statistics

This table reports the pooled summary statistics for all regression variables that appear in the subsequent tables. Panel A describes the case when the Pearson correlation ρ_{ijt} is calculated from the daily returns of firms i and j in excess of the risk free rate for each six month period t from 2003 to 2011. The firm universe consists of all domestic common stocks trading on the NYSE, NASDAQ, and Amex exchanges with CRSP share codes 10 or 11. The binary variables ω_{ijt}^{pos} and ω_{ijt}^{P50} are determined by the volume of text, measured in total words, that appear in the term-document matrix each period. The former has a value of 1 whenever both firms have some positive number of total words, and the latter has a value of 1 whenever both firms have a volume of text exeecing the median during that period. The variable ρ_{ijt}^{take} is equal to $N_{ijt}^{take} / \sqrt{N_{it}^{take} N_{jt}^{take}}$ where N_{ij}^{take} is the number of news takes that mention both firms i and j in a period t, and N_{it}^{take} and N_{jt}^{take} are the number of takes mentioning firms i and j respectively. The news similarity variable π_{ijt} is the cosine similarity between the firm vectors i and j in the term-document matrix for period t. Each firm in the sample is assigned to NYSE decile portfolios based on the following individual firm characteristics: market model beta over period t, firm market value from the last trading day of the prior year, book-to-market from the most recent quarterly report before the beginning period t, total return over the previous $t-12$ to $t-2$ months, closing price on the last trading day of the prior year, and institutional holdings during period t. The variables $BetaCorr_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktCorr_{ijt}$, $MomCorr_{ijt}$, $PrcCorr_{ijt}$, and $InstCorr_{ijt}$ are the daily return correlations between the portfolios containing firms i and j during period t. The return correlations between the 49 industry portfolios, as defined on Kenneth French's website, are used to form the variable $IndCorr_{ijt}$. The binary variables $SP500_{ijt}$, $SPVal_{ijt}$, and $SPGrw_{ijt}$ are set 1 if both firms i and j are members of the S&P 500, S&P1500 Value, and S&P 1500 Growth indices, respectively, on the last trading day of period t. The binary variable $SPValGrw_{ijt}$ is set to one if both firms are members of both the S&P1500 Value and S&P 1500 Growth Indices. The binary variable MSA_{ijt} is set to 1 if both firms i and j are headquartered in the same Metropolitan Statistical Area as defined by the Office of Management and Budget. The variable $EPSCorr_{ijt}$ is equal to $N_{ijt}^{an} / \sqrt{N_{it}^{an} N_{jt}^{an}}$ where N_{ij}^{an} is the number of analysts following both firms i and j in a period t, and N_{it}^{an} and N_{jt}^{an} are the number of analysts following firms i and j respectively. The variables $S34Corr_{ijt}$ and $S12Corr_{ijt}$ are defined similarly for institutional and mutual fund holdings respectively. Panel B describes the case when the Pearson correlation ρ_{ijt} is calculated from the market model residuals of firms i and j in each six month period t. The portfolio correlations used for $BetaCorr_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktCorr_{ijt}$, $MomCorr_{ijt}$, and $IndCorr_{ijt}$ are also calculated from residuals after the returns of each portfolio are regressed on the market factor. All the varaibles in Panels C and D are computed in a similar fashion using the Carhart (1997) four-factor model and the Lewellen and Nagel (2006) specification with lagged systematic factors respectively.

Panel A: Correlation calculated from excess returns

	MEAN	STD	P1	P10	P25	P50	P75	P90	P99
ρ_{ijt}	0.17022	0.16854	-0.15	-0.03	0.05	0.15	0.28	0.41	0.61
ω_{ijt}^{pos}	0.61843	0.48577	0.00	0.00	0.00	1.00	1.00	1.00	1.00
ω_{ijt}^{P50}	0.28278	0.45035	0.00	0.00	0.00	0.00	1.00	1.00	1.00
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$	0.00001	0.00091	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\pi_{ijt}\,\omega_{ijt}^{pos}$	0.23071	0.21484	0.00	0.00	0.00	0.26	0.42	0.51	0.63
$\pi_{ijt}\,\omega_{ijt}^{P50}$	0.11622	0.19503	0.00	0.00	0.00	0.00	0.29	0.46	0.60
$BetaCorr_{ijt}$	0.81666	0.16784	0.34	0.55	0.72	0.87	0.94	1.00	1.00
$SizeCorr_{ijt}$	0.94399	0.04876	0.77	0.88	0.92	0.95	0.98	1.00	1.00
$Bk/MktCorr_{ijt}$	0.90263	0.05689	0.75	0.83	0.87	0.90	0.94	0.98	1.00
$MomCorr_{ijt}$	0.88039	0.09158	0.58	0.76	0.84	0.90	0.94	1.00	1.00
$IndCorr_{ijt}$	0.80217	0.09687	0.57	0.68	0.74	0.80	0.87	0.92	1.00
$SP500_{ijt}$	0.01615	0.12606	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$SPVal_{ijt}$	0.07609	0.26514	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$SPGrw_{ijt}$	0.05351	0.22504	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$SPValGrw_{ijt}$	0.01500	0.12154	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$PrcCorr_{ijt}$	0.90577	0.06373	0.72	0.82	0.87	0.91	0.94	1.00	1.00
$InstCorr_{ijt}$	0.91422	0.05706	0.75	0.83	0.88	0.92	0.95	1.00	1.00
MSA_{ijt}	0.02896	0.16769	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$S34Corr_{ijt}$	0.27315	0.18309	0.00	0.00	0.12	0.29	0.43	0.50	0.60
$S12Corr_{ijt}$	0.17389	0.16550	0.00	0.00	0.01	0.13	0.29	0.42	0.60
$EPSCorr_{ijt}$	0.00240	0.02731	0.00	0.00	0.00	0.00	0.00	0.00	0.08

Panel B: Correlation calculated from market model residuals									
	MEAN	STD	P1	P10	P25	P50	P75	P90	P99
ρ_{ijt}	0.01673	0.10401	-0.22	-0.11	-0.05	0.01	0.08	0.15	0.28
$BetaCorr_{ijt}$	0.06871	0.48677	-0.71	-0.50	-0.33	-0.01	0.39	1.00	1.00
$SizeCorr_{ijt}$	0.57744	0.40721	-0.65	0.01	0.44	0.66	0.88	1.00	1.00
$Bk/MktCorr_{ijt}$	0.07203	0.40077	-0.52	-0.36	-0.22	-0.01	0.26	0.69	1.00
$MomCorr_{ijt}$	0.07907	0.45420	-0.66	-0.46	-0.27	0.01	0.33	1.00	1.00
$IndCorr_{ijt}$	0.04468	0.29741	-0.41	-0.25	-0.15	0.00	0.14	0.38	1.00
Panel C: Correlation calculated from residuals of market model with lagged factors									
	MEAN	STD	P1	P10	P25	P50	P75	P90	P99
ρ_{ijt}	0.00154	0.09859	-0.23	-0.12	-0.06	0.00	0.07	0.13	0.24
$BetaCorr_{ijt}$	0.06813	0.41918	-0.50	-0.36	-0.22	-0.04	0.22	1.00	1.00
$SizeCorr_{ijt}$	0.21978	0.44209	-0.47	-0.26	-0.12	0.12	0.40	1.00	1.00
$Bk/MktCorr_{ijt}$	0.06773	0.34438	-0.41	-0.25	-0.14	-0.01	0.12	0.45	1.00
$MomCorr_{ijt}$	0.06833	0.37085	-0.45	-0.27	-0.16	-0.03	0.16	1.00	1.00
$IndCorr_{ijt}$	0.04649	0.27427	-0.39	-0.21	-0.11	0.01	0.11	0.28	1.00
Panel D: Correlation calculated from residuals of four-factor model									
	MEAN	STD	P1	P10	P25	P50	P75	P90	P99
ρ_{ijt}	0.01631	0.10400	-0.22	-0.11	-0.05	0.01	0.08	0.15	0.28
$BetaCorr_{ijt}$	0.06779	0.48604	-0.71	-0.50	-0.33	-0.02	0.38	1.00	1.00
$SizeCorr_{ijt}$	0.58677	0.40741	-0.62	0.02	0.44	0.69	0.88	1.00	1.00
$Bk/MktCorr_{ijt}$	0.07081	0.40127	-0.52	-0.36	-0.22	-0.01	0.26	0.69	1.00
$MomCorr_{ijt}$	0.08049	0.45197	-0.65	-0.45	-0.26	0.01	0.33	1.00	1.00
$IndCorr_{ijt}$	0.04493	0.29741	-0.41	-0.25	-0.15	0.00	0.14	0.38	1.00

Table 2 provides summary statistics for all of the regression variables appearing in the article. The variables in Panel A are calculated using simple returns in excess of the risk free rate, and Panel B reports only the included variables that change as a result of calculating the return correlation using market model residuals. The average daily return correlation ρ_{ijt} across all firm-pairs and all 6 month periods is about 17% when

calculated with simple excess returns and 1.67% when calculated from market model residuals. The variable is predictably more centered near zero after removing the effect of the market factor, but neither distribution has any noticeable skewness. About 62% of the firm-pairs consist of two firms with some news coverage, while only 28% consist of two-firms with a volume of text above the median for the period. Very few of the included firms appear frequently in the same take, but the standard deviation 0.091% is considerably larger than the mean of 0.001%. The news similarity interaction terms are predictably truncated from below by the dummy variables, but neatly bounded from above around 60%. Summary statistics and definitions for all other variables are discussed as needed.

The first group of regression results is reported in Table 3, with Panel A including specifications where return correlation is calculated using simple excess returns, while Panel B uses residuals from the market model. The independent variables in each specification are also included as predetermined instruments. So any specification including $\pi_{ijt}\omega_{ijt}^{pos}$ as a right hand side variable will use all current and lagged levels and differences $\{ \pi_{ijt}\omega_{ijt}^{pos} , \pi_{ijt-1}\omega_{ijt-1}^{pos} ,\ldots, \pi_{ij1}\omega_{ij1}^{pos} , \Delta\pi_{ijt}\omega_{ijt}^{pos} , \Delta\pi_{ijt-1}\omega_{ijt-1}^{pos} ,\ldots, \Delta\pi_{ij1}\omega_{ij1}^{pos}\}$ as instruments, in addition to the lagged levels and differences of the dependent variable $\{\rho_{ijt-1}, \rho_{ijt-2},\ldots, \rho_{ij1}, \Delta\rho_{ijt-1}, \Delta\rho_{ijt-2},\ldots, \Delta\rho_{ij2}\}$ (Arellano and Bond, 1991). The number of instruments increases as independent variables are added to the model. The fit statistic, recommended by Bloom, Bond, and van Reenen (2007), is simply the squared correlation of the predicted values $\widehat{\rho_{ijt}}$ with the actual values ρ_{ijt}. It is not possible to calculate a typical R^2 measure because differencing the model removes the constant term.

The second order test for serial correlation (p-values reported) was suggested by Arellano and Bond (1991) to ensure that there is no pattern (null hypothesis) in the differenced times series residuals of the individual cross sections. All regression coefficients are reported in terms of the variable's economic impact on the dependent variable. The economic impact measures how many standard deviation changes in the dependent variable are caused by a single standard deviation change in the independent variable. The t-statistics of the original regression coefficients are reported in parenthesis.

Table 3: Regression results for news similarity and the information diffusion view

The dependent variable in all specification is the future Pearson daily return correlation ρ_{ijt+1}. Panel A describes the case when the Pearson correlations ρ_{ijt} and ρ_{ijt+1} are calculated from the daily returns of firms i and j in excess of the risk free rate for each six month period t from 2003 to 2011. Panel B describes the case when the Pearson correlations ρ_{ijt} and ρ_{ijt+1} are calculated from the market model residuals of firms i and j in each six month period t. The correlation varaibles in Panels C and D are computed in a similar fashion using the Carhart (1997) four-factor model and the Lewellen and Nagel (2006) specification with lagged systematic factors respectively. All of the included dependent variables are described in Table 2. Each specification is estimated with the Arellano and Bond (1991) dynamic panel estimation methodology, and all of the independent variables in a particular specification are used as predetermined instruments.

Panel A: Correlation calculated from excess returns

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ρ_{ijt}	0.050 (22.20)	0.097 (43.00)	0.078 (36.31)	0.054 (24.06)	0.168 (85.53)	0.101 (49.03)	0.175 (90.51)
ω_{ijt}^{pos}		-0.160 (-16.39)		0.064 (4.44)	0.005 (0.67)		0.013 (1.77)
ω_{ijt}^{P50}			-0.026 (-12.77)			-0.154 (-50.17)	-0.012 (-4.85)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$				-0.001 (-10.88)	0.004 (50.31)	0.005 (52.23)	0.004 (53.91)
$\pi_{ijt}\,\omega_{ijt}^{pos}$					0.145 (91.82)		0.139 (76.21)
$\pi_{ijt}\,\omega_{ijt}^{P50}$						0.147 (61.56)	0.020 (8.17)
Fit Statistic		0.279	0.151	0.283	0.273	0.127	0.211
AR(2) Test	1.000	0.000	0.861	1.000	0.000	0.024	0.000

Panel B: Correlation calculated from market model residuals

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ρ_{ijt}	0.019 (11.31)	0.028 (18.72)	0.027 (18.41)	0.021 (13.48)	0.029 (20.43)	0.028 (19.70)	0.029 (20.48)
ω_{ijt}^{pos}		-0.068 (-3.48)		0.233 (6.97)	-0.038 (-3.54)		-0.083 (-8.47)
ω_{ijt}^{P50}			-0.016 (-5.15)			-0.075 (-17.67)	-0.053 (-16.57)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$				0.001 (7.06)	0.010 (70.11)	0.011 (68.20)	0.010 (70.47)
$\pi_{ijt}\,\omega_{ijt}^{pos}$					0.048 (23.52)		0.044 (20.45)
$\pi_{ijt}\,\omega_{ijt}^{P50}$						0.066 (19.98)	0.026 (8.10)
Fit Statistic	0.020	0.010	0.003	0.000	0.009	0.001	0.004
AR(2) Test	0.732	0.004	0.013	1.000	0.000	0.006	0.000

Panel C: Correlation calculated from residuals of market model with lagged factors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ρ_{ijt}	0.005	0.011	0.011	0.007	0.011	0.011	0.011
	(3.06)	(7.98)	(8.10)	(5.05)	(7.94)	(8.09)	(7.90)
ω_{ijt}^{pos}		-0.026		-0.383	-0.024		-0.042
		(-1.26)		(-8.65)	(-1.70)		(-3.25)
ω_{ijt}^{P50}			-0.001			-0.047	-0.036
			(-0.39)			(-10.81)	(-11.05)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$				-0.001	0.010	0.010	0.010
				(-3.60)	(98.18)	(58.00)	(66.23)
$\pi_{ijt}\,\omega_{ijt}^{pos}$					0.031		0.019
					(15.70)		(9.16)
$\pi_{ijt}\,\omega_{ijt}^{P50}$						0.049	0.033
						(15.02)	(9.29)
Fit Statistic	0.003	0.000	0.000	0.001	0.000	0.000	0.000
AR(2) Test	0.999	0.381	0.394	1.000	0.308	0.392	0.347

Panel D: Correlation calculated from residuals of four-factor model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ρ_{ijt}	0.015	0.018	0.017	0.015	0.020	0.018	0.020
	(8.69)	(12.27)	(11.94)	(10.06)	(14.21)	(13.30)	(14.42)
ω_{ijt}^{pos}		-0.061		0.062	-0.042		-0.081
		(-3.11)		(2.40)	(-4.08)		(-10.07)
ω_{ijt}^{P50}			-0.014			-0.070	-0.047
			(-4.34)			(-16.51)	(-14.18)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$				0.002	0.010	0.010	0.010
				(12.66)	(50.49)	(54.90)	(52.15)
$\pi_{ijt}\,\omega_{ijt}^{pos}$					0.047		0.044
					(23.55)		(20.29)
$\pi_{ijt}\,\omega_{ijt}^{P50}$						0.062	0.020
						(18.91)	(5.75)
Fit Statistic	0.014	0.009	0.002	0.002	0.008	0.001	0.003
AR(2) Test	0.989	0.124	0.211	1.000	0.041	0.135	0.011

The economic impact of a single standard deviation change in news similarity $\pi_{ijt}\omega_{ijt}^{pos}$ results in a 0.145 standard deviation change in return correlation ρ_{ijt+1} when calculated from excess returns, and a 0.048 standard deviation change when calculated from market model residuals. Thus, the similarity of news text between two firms is capable of predicting future price comovement even after controlling for the current periods return correlation. This result supports my main hypothesis by providing evidence of a meaningful relationship between the similarity of contemporaneous information environments and the future price comovement of two firms.

A positive relationship between the text volume dummy variables and future return correlation would support the prediction of the information diffusion view that firms experiencing a high level of news coverage share a common factor in their returns. However, if a higher text volume reduces a firm-pair's return correlation, then increased news coverage must actually produce fundamental information investors can use to form more firm-specific expectations. The results in Table 3 do not provide support for the information diffusion view. The coefficient on ω_{ijt}^{pos} is only positive and significant in specification (4) of both panels, and in both cases the second order serial correlation test cannot reject the null of structure in the disturbances. The coefficient on ω_{ijt}^{P50} is negative and significant in every specification indicating that firms sharing high levels of news coverage actually experience less price comovement. This casts doubt on the prediction that the returns of firms with similar levels of news coverage share a common factor. Instead, it seems that investors are able to use the information provided by higher volumes

of news to remove "excess" correlation. This finding is much more consistent with the theoretical model of information-driven comovement provided in Veldkamp (2006b). In this model, comovement between assets would fall as additional signals about the fundamental values of those assets are purchased by investors. If the volume of text written about a firm is positively related to the amount of resources expended by investors to discover firm-specific information, then the negative coefficient on ω_{ijt}^{P50} provides empirical support for this alternative model.

The main inferences do not change when additional factors are included in the model of excess returns. The excess return correlation is also calculated using the within-period residuals of the Carhart (1997) four-factor model and the Lewellen and Nagel (2006) specification with lagged systematic factors. The economic impact of the news similarity variable is still positive and significant in all specifications using the residuals of these alternative factor models, however, including the additional factors does decrease the magnitudes. For example, Panels C and D of Table 3 show that the news similarity measure has an economic impact of 0.031 and 0.047 when return correlation is calculated from the residuals of the four-factor model and the specification with lagged systematic factors respectively.

The persistence of pairwise return correlation, the impact reported for ρ_{ijt}, has not been well documented in the finance literature though a variety of portfolio construction techniques that rely on the implicit assumption that future return correlation is strongly

related to its historical time series[7]. Table 3 does not support this assumption. When using only the current period's return correlation to predict future correlation, a one standard deviation change in the right hand side variable only results in a 0.050 standard deviation change in the dependent variable when calculated with excess returns, and a 0.019 standard deviation change when using market model residuals. However, the p-value on the second order test of serial correlation suggests that this simple model is badly misspecified in both panels. As additional variables are added to the model, both as independent variables and instruments, the presence of second order serial correlation becomes less likely. The fit statistic is usually difficult to interpret because, as is the case in most instrumental variables estimations, it does not increase with the addition of right hand side variables. At best, the current period's return correlation has an economic impact of 0.175 standard deviation in Panel A and 0.029 standard deviations in Panel B[8]. This finding does not offer strong support for the use of historical return correlations by themselves when predicting future comovement. The limitations of forming portfolios based on covariance matrices computed with past returns has been well documented, however this result provides insight into a potential cause of this shortcoming. If return covariance is time-varying, then sample estimated covariance matrices will not provide useful inputs for portfolio formation. An alterntive, but not necessarily mutually exclusive explanation, is that return-based estimates of covariance are prohibitively noisy.

[7] Chan, Karceski, and Lakonishok (1999) provide evidence that historical covariance is not very useful for constructing optimal portfolios.

[8] An empirical measure of return correlation persistence might also be useful for choosing a data-generating process in structural models like the one presented in Engle (2012).

5.2 THE FUNDAMENTALS VIEW

The fundamentals view states that comovement in prices reflects comovement in fundamental values. Tests of the fundamentals view will include variables related to firm beta, size, book-to-market, momentum, and industry. While the relationship between firm beta (Ledoit and Wolf, 2003), size (Pindyck and Rotemberg, 1993), book-to-market (Bekaert et al. 2009), industry (Campbell et al. (2001), Irvine and Pontiff (2009), and Brandt et al. (2010)) and price comovement might have some fundamental basis; the relationship between momentum and comovement might have less theoretical support. Thus, the subsequent analysis will test an admittedly broad interpretation of the fundamentals view.

Firm market values are calculated on the final trading day of the previous year, and each firm is placed in a market weighted portfolio based on NYSE size deciles for the next year[9]. The correlation between the size portfolios of two firms in one period is then used to predict their return correlation in the next period. For instance, if firm *i* is NYSE size decile 3 and firm *j* is in NYSE size decile 7, the correlation between the size portfolios 3 and 7 will be used to predict the correlation between firms *i* and *j* in the next period.

All additional explanatory variables included in Table 4 are calculated using these portfolio return correlations. The market model beta in the current period is used to place each firm in a NYSE beta decile. The values used to construct the NYSE book-to-market

[9] All independent variables constructed from stock returns are calculated using the entire sample of CRSP firms trading on the three major exchanges with share codes 10 or 11, not the truncated sample that only includes the 4,000 largest firms.

deciles are calculated in accordance with Fama and French (1993), where all financial statement information is taken from the most recent quarter ending before the start of the current formation period. Firms lacking the information necessary to calculate book-to-market in Compustat are included in their own market weighted portfolio, and their observations are retained in the sample. The average monthly return for all CRSP firms is calculated using the $t-2$ to $t-12$ monthly returns where t refers to the ending month of the current period. These average monthly returns are used to construct NYSE market weighted momentum portfolios for every period in the sample. Finally, the SIC codes for all firms are taken Compustat when they are available, and from CRSP when they are not. Each firm is assigned to one of the 49 industry portfolios, as defined on Kenneth French's website, and the market weighted correlation of these portfolios is used to measure the effect of industry on future return correlation.

Table 4: Regression results for the fundamentals view

The dependent variable in all specification is the future Pearson daily return correlation ρ_{ijt+1}. Panel A describes the case when the Pearson correlations ρ_{ijt} and ρ_{ijt+1} are calculated from the daily returns of firms i and j in excess of the risk free rate for each six month period t from 2003 to 2011. Panel B describes the case when the Pearson correlations ρ_{ijt} and ρ_{ijt+1} are calculated from the market model residuals of firms i and j in each six month period t. The correlation varaibles in Panels C and D are computed in a similar fashion using the Carhart (1997) four-factor model and the Lewellen and Nagel (2006) specification with lagged systematic factors respectively. All of the included dependent variables are described in Table 2. Each specification is estimated with the Arellano and Bond (1991) dynamic panel estimation methodology, and all of the independent variables in a particular specification that are not calculated from portfolio returns are used as predetermined instruments. The other explanatory variables $BetaCorr_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktCorr_{ijt}$, $MomCorr_{ijt}$, and $IndCorr_{ijt}$ are included as correlated predetermined instruments.

Panel A: Correlation calculated from excess returns

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
ρ_{ijt}	0.049 (27.17)	0.116 (71.15)	0.054 (30.32)	0.122 (72.89)	0.055 (30.78)	0.131 (80.14)	0.055 (30.82)	0.134 (80.36)	0.051 (28.63)	0.090 (57.37)	0.078 (47.82)
ω_{ijt}^{pos}		-0.127 (-25.73)		-0.116 (-22.92)		-0.164 (-33.71)		-0.159 (-32.52)		-0.063 (-12.01)	-0.126 (-25.40)
ω_{ijt}^{P50}		-0.011 (-4.48)		0.038 (13.33)		-0.013 (-4.76)		-0.017 (-6.47)		0.048 (14.59)	0.049 (15.99)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$		0.004 (61.15)		0.004 (61.90)		0.004 (43.35)		0.004 (73.62)		-0.002 (-28.71)	0.000 (-2.88)
$\pi_{ijt}\,\omega_{ijt}^{pos}$		0.139 (75.17)		0.140 (70.99)		0.142 (73.23)		0.137 (74.68)		0.062 (27.39)	0.079 (36.68)
$\pi_{ijt}\,\omega_{ijt}^{P50}$		0.021 (8.24)		-0.019 (-6.90)		0.026 (9.49)		0.028 (10.51)		-0.033 (-10.54)	-0.033 (-10.94)
$BetaCorr_{ijt}$	0.010 (5.32)	0.062 (32.72)									0.032 (27.42)
$SizeCorr_{ijt}$			-0.008 (-2.18)	0.310 (83.96)							0.061 (25.84)
$Bk/MktCorr_{ijt}$					-0.013 (-6.18)	-0.129 (-75.17)					-0.007 (-6.28)
$MomCorr_{ijt}$							0.001 (0.81)	-0.053 (-36.74)			-0.002 (-1.75)
$IndCorr_{ijt}$									0.423 (28.05)	0.896 (205.82)	0.741 (187.67)
Fit Statistic	0.083	0.226	0.009	0.209	0.007	0.210	0.006	0.212	0.024	0.190	0.204
AR(2) Test	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.277	0.000

Panel B: Correlation calculated from market model residuals

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
ρ_{ijt}	0.022 (13.85)	0.029 (20.27)	0.027 (18.25)	0.029 (20.67)	0.024 (15.80)	0.029 (20.62)	0.024 (15.19)	0.029 (20.24)	0.022 (13.67)	0.028 (19.99)	0.026 (19.14)
ω_{ijt}^{pos}		-0.078 (-8.06)		-0.019 (-2.35)		-0.069 (-7.03)		-0.083 (-8.42)		-0.066 (-5.71)	-0.012 (-1.33)
ω_{ijt}^{P50}		-0.053 (-16.68)		-0.040 (-11.78)		-0.052 (-15.81)		-0.052 (-16.26)		-0.027 (-8.15)	-0.014 (-4.16)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$		0.010 (69.79)		0.010 (74.17)		0.010 (66.66)		0.010 (65.15)		0.006 (38.12)	0.006 (35.80)
$\pi_{ijt}\,\omega_{ijt}^{pos}$		0.043 (19.91)		0.044 (19.95)		0.044 (20.96)		0.044 (20.67)		0.013 (5.47)	0.017 (6.70)
$\pi_{ijt}\,\omega_{ijt}^{P50}$		0.027 (8.31)		0.023 (6.84)		0.024 (7.26)		0.025 (7.71)		0.002 (0.66)	-0.003 (-0.88)
$BetaCorr_{ijt}$	0.005 (3.07)	0.011 (7.19)									0.010 (7.30)
$SizeCorr_{ijt}$			0.017 (1.70)	0.084 (13.67)							0.072 (11.61)
$Bk/MktCorr_{ijt}$					-0.003 (-1.45)	0.004 (1.92)					0.004 (1.97)
$MomCorr_{ijt}$							0.008 (5.57)	0.013 (10.53)			0.015 (12.48)
$IndCorr_{ijt}$									0.136 (4.87)	0.331 (25.97)	0.318 (25.36)
Fit Statistic	0.004	0.006	0.009	0.008	0.000	0.004	0.020	0.005	0.011	0.007	0.010
AR(2) Test	0.079	0.000	0.003	0.000	0.004	0.000	0.037	0.000			

Panel C: Correlation calculated from residuals of market model with lagged factors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
ρ_{ijt}	0.010	0.010	0.011	0.010	0.011	0.010	0.011	0.011	0.011	0.010	0.009
	(7.60)	(7.48)	(7.92)	(7.70)	(7.96)	(7.77)	(8.11)	(7.91)	(7.85)	(7.53)	(7.14)
ω_{ijt}^{pos}		-0.043		-0.043		-0.036		-0.038		-0.032	-0.034
		(-3.43)		(-3.93)		(-2.82)		(-3.01)		(-2.41)	(-3.05)
ω_{ijt}^{P50}		-0.037		-0.036		-0.037		-0.036		-0.022	-0.022
		(-11.19)		(-10.99)		(-11.09)		(-11.03)		(-6.46)	(-6.33)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$		0.010		0.010		0.010		0.010		0.008	0.008
		(100.53)		(98.90)		(65.68)		(64.12)		(88.21)	(84.15)
$\pi_{ijt}\,\omega_{ijt}^{pos}$		0.019		0.019		0.020		0.020		0.007	0.007
		(9.03)		(9.06)		(9.16)		(9.18)		(2.83)	(3.09)
$\pi_{ijt}\,\omega_{ijt}^{P50}$		0.034		0.033		0.033		0.033		0.021	0.021
		(9.54)		(9.39)		(9.35)		(9.25)		(5.85)	(5.72)
$BetaCorr_{ijt}$	0.004	0.008									0.007
	(2.25)	(6.14)									(4.98)
$SizeCorr_{ijt}$			0.003	0.008							0.007
			(0.67)	(2.26)							(1.92)
$Bk/MktCorr_{ijt}$					-0.001	-0.002					-0.002
					(-0.53)	(-0.94)					(-1.07)
$MomCorr_{ijt}$							0.000	0.000			0.000
							(0.04)	(-0.07)			(0.15)
$IndCorr_{ijt}$									0.080	0.169	0.167
									(2.88)	(11.45)	(11.38)
Fit Statistic	0.003	0.000	0.003	0.000	0.003	0.000	0.003	0.000	0.005	0.000	0.000
AR(2) Test	0.600	0.388	0.400	0.380	0.301	0.394	0.404	0.343			

Panel D: Correlation calculated from residuals of four-factor model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
ρ_{ijt}	0.014 (9.40)	0.019 (14.27)	0.017 (11.71)	0.019 (13.85)	0.016 (10.87)	0.019 (14.09)	0.014 (9.52)	0.018 (13.55)	0.016 (10.20)	0.018 (13.54)	0.016 (12.31)
ω_{ijt}^{pos}		-0.061 (-7.52)		-0.007 (-0.87)		-0.065 (-8.14)		-0.074 (-9.04)		-0.080 (-7.61)	-0.008 (-0.93)
ω_{ijt}^{P50}		-0.047 (-14.10)		-0.031 (-9.11)		-0.047 (-14.07)		-0.046 (-14.03)		-0.024 (-6.76)	-0.010 (-2.73)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$		0.010 (53.05)		0.010 (55.03)		0.010 (52.12)		0.010 (53.12)		0.006 (36.10)	0.006 (40.45)
$\pi_{ijt}\,\omega_{ijt}^{pos}$		0.044 (20.23)		0.045 (21.00)		0.045 (20.54)		0.045 (20.85)		0.016 (6.16)	0.018 (7.43)
$\pi_{ijt}\,\omega_{ijt}^{P50}$		0.020 (5.73)		0.016 (4.52)		0.020 (5.60)		0.019 (5.48)		-0.002 (-0.48)	-0.007 (-1.99)
$BetaCorr_{ijt}$	0.004 (2.57)	0.009 (5.72)									0.007 (4.84)
$SizeCorr_{ijt}$			0.028 (2.54)	0.097 (13.43)							0.079 (11.43)
$Bk/MktCorr_{ijt}$					-0.003 (-1.37)	0.002 (1.01)					0.003 (1.41)
$MomCorr_{ijt}$							0.007 (4.64)	0.013 (9.87)			0.015 (11.77)
$IndCorr_{ijt}$									-0.045 (-1.55)	0.306 (21.82)	0.303 (26.17)
Fit Statistic	0.003	0.005	0.008	0.007	0.002	0.004	0.014	0.004	0.009	0.006	0.009
AR(2) Test	0.148	0.013	0.085	0.005	0.095	0.008	0.346	0.014			

An alternative to this portfolio correlation approach commonly used in the literature would be to use dummy variables indicating that firms are in the same category (Muslu, Rebello, and Xu (2012) and Chen, Chen, and Li (2012)). However, this approach could mask the complexity of the actual correlation structure. Accounting only for the comovement implied by shared industry membership would likely bias the results towards finding support for news similarity as a determinant of individual firm-pair return correlation. Moreover, investors are likely to infer information about one portfolio from its return correlation with all the other portfolios. For instance, investors interested in a sector rotation strategy might pay more attention to the return correlation between industries than the correlation within.

Panel A of Table 2 reports the summary statistics for these additional variables whenever correlation is calculated using returns in excess of the risk free rate. Continuing with the example above, the correlation between the daily returns of size portfolios 3 and 7 would be calculated after subtracting the risk free rate. Panel B computes correlations using the residuals from the market model. In this case, residuals from a regression of portfolio returns on a constant and the market return are used to calculate the portfolios' comovement. Firms assigned to the same portfolio, across any of the five characteristics, will have a portfolio correlation value of 1 in each case. In both panels, constructing portfolios from market model betas seems to induce the greatest amount of correlation variability, both in range and standard deviation. Organizing portfolios by firm size or book-to-market does not seem to induce as much variability in return correlation regardless of the specification.

Unlike the explanatory variables from the tests reported Table 3, the levels of the additional return-based measures will not be included as instruments due to their probable correlation with the error term. For instance, two large firms from different industries are likely to account for a

great deal of their respective industry portfolio returns. Thus the correlation of their portfolio returns is probably related to their actual return correlation, which, according to section chapter 4, might produce biased parameter estimates. The current and lagged differences for all of these return-based measures are still included as instruments when they appear as independent variables in the specification.

The economic impact of news similarity $\pi_{ijt}\omega_{ijt}^{pos}$ is positive and significant in every specification, and only about half of the variable's predictability is subsumed by industry correlation. The economic impact of the interaction term $\pi_{ijt}\omega_{ijt}^{P50}$, between news similarity and above-median text volume, becomes negative whenever the size or industry correlation variable is included in the model. This is likely the result of positive relationships between size and news coverage and industry and news coverage. In Panel A of Table 4, the test for second order serial correlation fails to reject the presence of structure in the disturbances for all cases where only a single return-based measure is included in the model, so these equations are likely to be misspecified. In the models including all of the news-based measures and one return-based measure, the economic impact of a change in the correlation is significantly positive for the beta, size, and industry portfolios and zero or negative for the book-to-market and momentum portfolios. Surprisingly, the economic impact of the correlation implied by firm betas, 0.062 return correlation standard deviations, is somewhat muted compared with the other explanatory variables. A considerable volume of asset pricing theory would suggest that more price comovement might be explained by a firm's relationship to the broader market. Not surprisingly, the economic impact of a standard deviation change in industry portfolio return correlation is very high at 0.896. When all of the variables are combined into one specification, the significance of the momentum portfolio correlation disappears and the effects of all other variables are attenuated. In Table 4, pairwise

excess return correlation is found to be even less persistent than in the specifications reported in Table 3, and almost all of the variables with significantly positive economic impacts seem to be better predictors of future comovement.

When return correlation is calculated from the residuals of the market model, most of the economic impacts are attenuated but the inferences remain the same. The economic impact of news similarity $\pi_{ijt}\omega_{ijt}^{pos}$ remains positive and significant for all specifications. Thus after controlling for a variety of possible fundamental explanations, there is still information in the news similarity measure capable of predicting return correlation. The correlations of the beta portfolios retain some predictability for future market model residual correlation even though the calculation of return correlation is intended to account for the effect of firm beta. The economic impact of momentum switches sign in Panel B, and is significantly positive in all specifications. The excess return correlation of the industry portfolio is a very strong predictor of future market model residual correlation, ranging between 0.136 and 0.331. As in Panel A, the addition of the industry correlation variable only reduces the economic impact of news similarity on return correlation by about half. Overall, the results in Table 4 provide support for the fundamentals view of comovement, at least in terms of return correlation implied by firm beta, size, and industry. These fundamental characteristics and news similarity variables are more important for predicting the future return correlation between firms than their own current return correlation.

5.3 THE CATEGORY VIEW

The category view, proposed by Barberis and Shleifer (2003), predicts that in order to simplify portfolio decisions, many investors group assets into categories and then allocate funds at the level of these categories rather than at the individual asset level. As in Barberis, Shleifer, and Wurgler (2005), membership in the S&P 500 Index will be used to separate firms into categories that should

have no measurable relationship to return comovement if only the fundamentals view is correct. Finding that the members of the S&P 500 have higher future return correlation with each other than with nonmembers, after accounting for the fundamental variables introduced in the previous section, would provide support for the category view. Similary, membership in one of the S&P Value or Growth indices, as suggested by Boyer (2011), can also be used to separate firms in to categories. Finally, Green and Hwang (2009) proposed that investors might also categorize stocks based on price, so a variable to account for differences in stock price will also be included.

Index membership is taken from the Compustat Index Constituents file, and all firms that are listed as members of a particular index on the last day of the formation period are considered index members for that period. Boyer (2011) uses the S&P/Barra Value and Growth Indices, defined by dividing all S&P 500 stocks into two mutually exclusive categories according to simple mechanical rules. Unfortunately, those indices were discontinued in 2005, so the nonmutually exclusive S&P 1500 Value and Growth Indices must be used instead. According to Table 2, Only 1.6% of the firm-pairs contain two members of the S&P 500, whereas 7.6% and 5.4% contain two members of the S&P 1500 Value and Growth Indices respectively. The nonmutual exclusivity of the latter indices is demonstrated by the positive percentage, 1.5%, of firm-pairs consisting of two firms that are members of both indexes. NYSE deciles based on the closing stock price on the last trading day in the previous year will be used to form price portfolios, and firms will remain in those portfolios for the following year. Similar to the return-based variables created to test the fundamentals view, the return correlation between these price portfolios will be used to test the category view. There does seem to be measurable variation in the return correlation of the price portfolios. The standard deviation of the price correlations are larger than that of the size and book-to-market portfolios, suggesting that there may be price related structure in stock returns.

For all subsequent tests, variables related to the information diffusion and fundamental views will be retained in some of the specifications as additional control variables. Table 5 reports the tests of the category hypothesis with the dummy variables for index membership and the price portfolio correlations included as regressors. After controlling for the information environment and fundamental variables, index membership still explains a meaningful portion of future return correlation, with economic impacts of 0.069 for the S&P 500, and 0.113 and 0.073 for the S&P 1500 Value and Growth indices respectively. The postive impacts of membership in the value and growth indices are more remarkable because the correlation of book-to-market portfolios is also included as an independent variable. Inclusion in both of these categories does not seem to be a predictor of positive future return correlation, with an economic impact of -0.048 when both firms appear in the union of the indices constituents. This implies that investors recognize that firms classified as both value and growth are less correlated with firms that are only members of one index. Because the price correlation variable is a return based measure, it's levels are not used as intruments in the model estimations. The economic impact of a standard deviation change in the correlation of the firm's price portfolios is a positive and significant. 0.012. Even after controlling for all of the included news related and fundamental variables, price is still able to explain a small amount of future return correlation.

Table 5: Regression results for the category view.

The dependent variable in all specification is the future Pearson daily return correlation ρ_{ijt+1}, calculated from the daily returns of firms i and j in excess of the risk free rate for each six month period t from 2003 to 2011. All of the included dependent variables are described in Table 2. Each specification is estimated with the Arellano and Bond (1991) dynamic panel estimation methodology, and all of the independent variables in a particular specification that are not calculated from portfolio returns are used as predetermined instruments. The other explanatory variables $BetaCorr_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktCorr_{ijt}$, $MomCorr_{ijt}$, $IndCorr_{ijt}$, and $PrcCorr_{ijt}$ are included as correlated predetermined instruments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ρ_{ijt}	0.049	0.175	0.080	0.058	0.200	0.097	0.057	0.135	0.080
	(21.84)	(91.22)	(50.11)	(26.60)	(109.81)	(59.27)	(31.63)	(85.26)	(49.99)
ω_{ijt}^{pos}		0.025	-0.117		0.130	-0.046		-0.125	-0.119
		(3.42)	(-23.72)		(20.91)	(-9.92)		(-26.54)	(-24.12)
ω_{ijt}^{P50}		-0.008	0.052		-0.010	0.055		-0.015	0.050
		(-3.46)	(16.85)		(-3.95)	(17.82)		(-5.91)	(16.33)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$		0.003	-0.001		0.003	-0.001		0.004	0.000
		(41.85)	(-18.76)		(43.22)	(-12.55)		(58.67)	(-5.90)
$\pi_{ijt}\,\omega_{ijt}^{pos}$		0.135	0.077		0.123	0.070		0.138	0.079
		(75.99)	(35.90)		(69.33)	(31.72)		(75.85)	(36.86)
$\pi_{ijt}\,\omega_{ijt}^{P50}$		0.012	-0.040		0.008	-0.045		0.027	-0.033
		(4.82)	(-13.12)		(3.13)	(-14.93)		(10.55)	(-11.13)
$BetaCorr_{ijt}$			0.032			0.029			0.032
			(27.15)			(23.52)			(26.95)
$SizeCorr_{ijt}$			0.063			0.060			0.056
			(26.18)			(23.80)			(23.07)
$Bk/MktCorr_{ijt}$			-0.008			-0.007			-0.011
			(-6.82)			(-5.22)			(-9.26)
$MomCorr_{ijt}$			-0.001			-0.001			-0.001
			(-1.13)			(-0.97)			(-1.50)
$IndCorr_{ijt}$			0.736			0.737			0.738
			(188.02)			(192.12)			(187.66)
$SP500_{ijt}$	-0.151	0.052	0.069						
	(-8.30)	(43.74)	(37.62)						
$SPVal_{ijt}$				-0.004	0.116	0.113			
				(-1.16)	(97.31)	(68.20)			
$SPGrw_{ijt}$				-0.027	0.082	0.073			
				(-9.64)	(77.69)	(52.19)			
$SPValGrw_{ijt}$				0.014	-0.043	-0.048			
				(8.61)	(-48.43)	(-48.34)			
$PrcCorr_{ijt}$							0.018	0.009	0.012
							(7.04)	(4.27)	(7.62)
Fit Statistic	0.237	0.212	0.206	0.140	0.226	0.220	0.007	0.214	0.208
AR(2) Test	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000

For all of the specifications reported in Table 5, the news similarity and lagged return correlation variables are able to predict a meaningful portion of future stock price comovement. The results related to the index and price variables also provide support for the category hypothesis, however the similarities between individual firm's information environment are still incorporated into future price comovement.

5.4 THE HABITAT VIEW

The final explanation for comovement presented by Barberis, Shleifer, and Wurgler (2005) is the habitat view, which starts from the observation that many investors choose to trade only in a subset of all available securities. Such preferred habitats could arise because of transaction costs, international trading restrictions, or lack of information. To test the habitat view, institutional ownership (Pindyck and Rotemberg, 1993) and firm headquarters location (Pirinsky and Wang, 2006) will be used to identify investor groups. Lee, Shleifer, and Thaler (1991) were the first to suggest that the percentage of the firm's shares owned, or not owned, by institutional investors might explain excess comovement in the stock returns. Because some institutions might be restricted from owning firms with particular characteristics, the investment universe for these organizations could be paritioned into identifiable habitats. Pirinsky and Wang (2006) do not motivate their hypothesis with the habitat view, however it is plausable that a variety of factors related to investor sentiment and information flow around the geographical headquarters location, a firm's literal habitat, could cause the stock prices to comove.

The Thomson Reuters Institutional Holdings (13F) Database is used to calculate the level of institutional ownership, on the final day of each formation period, for all of the sample firms. Similar to the approach introduced in the section discussing the fundamentals view, NYSE institutional holding decile portfolios are formed for each period, and the correlations between

those portfolios are used as an explanatory variable. Table 2 reports that the magnitude and range of the institutional correlation variable is similar to the other return-based measures. The county and state of a firm's headquarters locations are taken from the CRSP/Compustat Merged Company Header History file, and merged with the list of Metropolitan Statistical Areas (MSA) defined by the Office of Management and Budget and reported on the Census Bureau's website. All observations with firm-pairs headquartered in the same MSA will have a value of 1 for the *MSA* variable. According to Table 2, about 2.9% of the firm-pairs consist of firms located in the same region.

Table 6: Regression tests of the habitat view

The dependent variable in all specification is the future Pearson daily return correlation ρ_{ijt+1}, calculated from the daily returns of firms i and j in excess of the risk free rate for each six month period t from 2003 to 2011. All of the included dependent variables are described in Table 2. Each specification is estimated with the Arellano and Bond (1991) dynamic panel estimation methodology, and all of the independent variables in a particular specification that are not calculated from portfolio returns are used as predetermined instruments. The other explanatory variables $BetaCorr_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktCorr_{ijt}$, $MomCorr_{ijt}$, $IndCorr_{ijt}$, and $InstCorr_{ijt}$ are included as correlated predetermined instruments.

	(1)	(2)	(3)	(4)	(5)	(6)
ρ_{ijt}	0.056 (31.45)	0.134 (80.63)	0.079 (48.57)	0.049 (21.90)	0.180 (93.95)	0.079 (48.95)
ω_{ijt}^{pos}		-0.140 (-29.47)	-0.124 (-25.07)		0.033 (4.65)	-0.117 (-23.91)
ω_{ijt}^{P50}		-0.016 (-6.16)	0.048 (15.59)		-0.014 (-5.52)	0.050 (15.94)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$		0.004 (59.98)	0.000 (-5.55)		0.004 (52.35)	0.000 (-2.43)
$\pi_{ijt}\,\omega_{ijt}^{pos}$		0.141 (74.57)	0.080 (37.25)		0.136 (75.64)	0.079 (36.48)
$\pi_{ijt}\,\omega_{ijt}^{P50}$		0.031 (11.67)	-0.031 (-10.42)		0.022 (8.66)	-0.032 (-10.55)
$BetaCorr_{ijt}$			0.032 (26.86)			0.033 (28.02)
$SizeCorr_{ijt}$			0.054 (22.39)			0.061 (24.73)
$Bk/MktCorr_{ijt}$			-0.011 (-9.11)			-0.009 (-7.41)
$MomCorr_{ijt}$			-0.003 (-3.00)			-0.001 (-1.08)
$IndCorr_{ijt}$			0.739 (186.89)			0.741 (188.72)
$InstCorr_{ijt}$	0.007 (3.46)	-0.050 (-29.73)	0.000 (-0.28)			
MSA_{ijt}				-0.163 (-3.04)	-0.023 (-14.76)	-0.069 (-28.57)
Fit Statistic	0.013	0.213	0.207	0.142	0.207	0.196
AR(2) Test	1.000	0.000	0.000	1.000	0.000	0.000

Table 6 does not provide much support for the habitat view. Because the institutional correlation variable is calculated using market returns, its levels are not included as instruments in the estimation. When all control variables are included in the specifications, the institutional ownership variable has an impact of 0.000 and the headquarters location variable has a

significantly negative impact of -0.069. Thus it seems unlikely that frictions related to investor habitats explain an economically meaningful amount of return comovement. As before, the economic impact of news similarity remains positive and significant in all specifications.

CHAPTER 6 ROBUSTNESS

In all reported specifications, the similarity of firm information environments has predicted future return comovement. It may be possible that the news similarity variable only proxies for aspects of a firm's information environment that investors discover from other sources. Because the frequency of firm-specific news seems to be highly correlated with earnings season, as reported in Table 1, analyst forecasts of these earnings may also explain future return correlation. It may also be the case that the current and lagged level of return correlation is actually causing news similarity. If the news similarity variable is sufficiently persistent, it may appear to predict future comovement even when causality runs the other direction. The following additional tests will shed light on these questions.

6.1 COMMON ANALYST COVERAGE

Israelsen (2012) and Muslu, Rebello, and Xu (2012) both examine the effect of common analyst coverage on stock return comovement. Because analysts often use models or standardized methodologies when making predictions about earnings, many of the inputs used in their models, such as projected GDP or industry growth may be used across multiple stocks. This may cause systematic errors in their earnings forecasts among the stocks that they follow. If investors use these forecasts to make trading decisions but do not account for these systematic errors, abnormal return correlation might be observed between stocks with similar analyst followings. For this study, the measure of analyst coverage provided by Israelsen (2012) will be used to determine the proportion of firm-pairs information related comovement that is attributable to commonality in their analyst following. This variable is defined as:

$$EPSCorr_{ijt} = {N_{ijt}^{an}} \Big/ \sqrt{N_{it}^{an} N_{jt}^{an}} \qquad 7$$

where N_{ij}^{an} is the number of analysts following both firms i and j in a period t, and N_{it}^{an} and N_{jt}^{an} are the number of analysts following firms i and j respectively. Variables for instiutional $S34Corr_{ijt}$ and mutual fund ownership $S12Corr_{ijt}$ are constructed in an analgolous way to account for organizational commonalities in forecasts attributed to the same type of imperfectly estimated model inputs.

Each period the common analyst coverage is calculated from the I/B/E/S database by counting the number of unique analyst making earnings forecasts during that span. Likewise, the number of institions, from the Thomson Reuters Institutional Holdings (13F) database, and mutual funds, from the Thomson Reuters Mutual Fund Holdings database, that reported ownership of particular firms over the period are used to calculate the $S34Corr_{ijt}$ and $S12Corr_{ijt}$ variables respectively. Table 2 indicates that the commonality of ownership is significantly higher in terms of institutions, 27.3%, than mutual funds, 17.4%, though a sizable quantity of firm-pairs are owned by the same organizations in either case. The commonality of analyst following, however, is much more sporadic, 0.2%, because most firms do not have analyst coverage.

Specifications including these additional variables are reported in Table 7. When all appropriate controls are included in the model, only the institutional ownership variable is able to explain future return correlation. Common mutual fund ownership and analayst coverage have a negative impact on stock price comovement after accounting for other sources of correlation. This coud imply that the proprietary research conducted by mutual funds and analysts actually provide

the markets with information that can be used to make better firm-specific valuations. As before, the economic impact of news similarity remains positive and significant in all specifications.

Table 7: Resgression tests of analyst coverage and institutional ownership

The dependent variable in all specification is the future Pearson daily return correlation ρ_{ijt+1}, calculated from the daily returns of firms i and j in excess of the risk free rate for each six month period t from 2003 to 2011. All of the included dependent variables are described in Table 2. Each specification is estimated with the Arellano and Bond (1991) dynamic panel estimation methodology, and all of the independent variables in a particular specification that are not calculated from portfolio returns are used as predetermined instruments. The other explanatory variables $BetaCorr_{ijt}$, $SizeCorr_{ijt}$, $Bk/MktCorr_{ijt}$, $MomCorr_{ijt}$, and $IndCorr_{ijt}$ are included as correlated predetermined instruments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ρ_{ijt}	0.051 (24.07)	0.175 (100.84)	0.112 (73.02)	0.078 (34.77)	0.203 (114.30)	0.103 (65.36)	0.050 (22.39)	0.178 (93.40)	0.080 (48.64)
ω_{ijt}^{pos}		0.405 (94.94)	0.123 (35.81)		0.303 (69.25)	0.006 (1.60)		0.040 (5.54)	-0.115 (-23.42)
ω_{ijt}^{P50}		-0.008 (-2.88)	0.044 (14.78)		-0.003 (-1.34)	0.050 (16.42)		-0.009 (-3.61)	0.046 (14.94)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$		0.003 (39.70)	-0.001 (-6.85)		0.003 (30.94)	0.000 (3.04)		0.002 (30.09)	0.002 (29.27)
$\pi_{ijt}\,\omega_{ijt}^{pos}$		0.045 (19.89)	0.076 (35.60)		0.107 (58.71)	0.104 (47.09)		0.138 (77.84)	0.090 (42.27)
$\pi_{ijt}\,\omega_{ijt}^{P50}$		0.011 (3.91)	-0.028 (-9.42)		0.021 (8.53)	-0.036 (-12.01)		0.015 (5.93)	-0.026 (-8.81)
$BetaCorr_{ijt}$			0.027 (22.44)			0.032 (25.90)			0.030 (25.24)
$SizeCorr_{ijt}$			0.026 (10.59)			0.032 (13.42)			0.068 (28.63)
$Bk/MktCorr_{ijt}$			-0.005 (-4.21)			-0.005 (-3.61)			-0.010 (-7.57)
$MomCorr_{ijt}$			-0.002 (-2.51)			-0.001 (-0.97)			-0.003 (-2.72)
$IndCorr_{ijt}$			0.701 (197.23)			0.748 (197.66)			0.712 (193.97)
$S34Corr_{ijt}$	-0.478 (-36.82)	0.456 (97.96)	0.088 (23.30)						
$S12Corr_{ijt}$				0.495 (34.53)	0.216 (68.36)	-0.089 (-29.62)			
$EPSCorr_{ijt}$							0.017 (4.48)	0.027 (22.35)	-0.027 (-21.32)
Fit Statistic	0.283	0.227	0.226	0.220	0.221	0.215	0.313	0.211	0.205
AR(2) Test	0.000	0.000	0.000	0.889	0.000	0.000	1.000	0.000	0.000

6.2 NEWS SIMILARITY PERSISTENCE

The final tests in this study focus on the causal relationship between return correlation and news similarity. All of the results to this point have implied that news similarity in the current period explains return comovement in the following period. It may also be of interest to determine

how much return correlation in the current period can predict news similarity in the future. To test the null hypothesis that return correlation does not Granger (1969) cause news similarity, the contemporaneous values of return correlation must be non-significant predictors of the news similarity variable when contemporaneous values of the dependent variable are also included. By this criteria, it is possible to say that news similarity Granger causes return correlation. However, it is not uncommon for two variables to Granger cause each other in tests of this nature.

To estimate this relationship, only observations with some positive volume of news text are retained in the sample. Because the dependent variable is no longer a return-based measure, the levels and differences of the firm-pair and portfolio correlations can be included as predetermined instruments. Table 8 provides conflicting evidence on the causal relationship between price comovement and news similarity. The economic impact of return correlation and news similarity is not significant in a right-tailed test for specifications (2) and (3), but is very significant when contemporaneous portfolio correlations are also included. Because the latter specifications are likely to be better specified, there does seem to be Granger causality flowing in both directions. However, the persistence of the news similarity measure is not high enough to warrant concern about potential feedback effects.

Table 8: Regression tests of Granger (1969) causality.

The dependent variable in all specification is the future news similarity variable π_{ijt+1} defined as the cosine similarity between the firm vectors i and j in the term-document matrix for period $t+1$ for each six month period t from 2003 to 2011. Each specification is estimated with the Arellano and Bond (1991) dynamic panel estimation methodology, and all of the independent variables in a particular specification are used as predetermined instruments.

	(1)	(2)	(3)	(4)	(5)
ρ_{ijt}		-0.044 (-25.52)	-0.038 (-30.47)	0.045 (47.07)	0.044 (49.35)
ω_{ijt}^{P50}			0.020 (20.26)		0.060 (72.73)
$\rho_{ijt}^{take}\,\omega_{ijt}^{pos}$			0.006 (81.90)		0.005 (83.26)
$BetaCorr_{ijt}$				-0.011 (-14.06)	-0.014 (-17.29)
$SizeCorr_{ijt}$				0.010 (7.12)	0.016 (11.93)
$Bk/MktCorr_{ijt}$				0.004 (4.01)	0.004 (4.60)
$MomCorr_{ijt}$				0.005 (7.29)	0.006 (9.79)
$IndCorr_{ijt}$				0.041 (26.76)	0.049 (33.58)
π_{ijt}	-0.020 (-3.36)	0.044 (8.35)	0.086 (27.29)	0.160 (76.22)	0.165 (91.37)
Fit Statistic	0.449	0.022	0.003	0.173	0.147
AR(2) Test	0.000	0.000	0.000	0.000	0.000

CHAPTER 7 CONCLUDING REMARKS

In this article, I introduce a novel approach for quantifying a firm's information environment and use the cross-firm similarity of this quantity to predict future price comovement. Commonality across information environments is measured by the textual similarity of firm-specific news articles appearing on the Reuters Integrated Data Network from 2003 to 2011. This measure of news similarity is able to predict an economically meaningful portion of future return correlation after controlling for numerous alternative explanations of comovement that have been introduced in prior literature. Prior literature has shown that the financial media is informative about future stock returns, but this article is the first to show that media content also anticipates return comovement.

The intuition for this result draws from the dimensionality of the inputs in the price generating function relative to the historical time series of the prices themselves. Existing methodologies for predicting future return comovement rely on a single-dimensional price history and a lengthy time series of returns. Innovations in the price function resulting from mergers, government policy, or macroeconomic conditions will not be reflected in the historical returns of an asset, so past observations are inevitably poor predictors of future market dynamics. I propose an alternative methodology for predicting future comovement that incorporates the depth of qualitative information provided by the financial media into expectations about innovations in future covariance. This methodology provides a quantifiable measure that is able to mimic the process by which human beings integrate qualitative knowledge of these innovations into their own expectations.

The inability of past prices to predict future comovement has been detrimental to the field of portfolio choice. My results indicate that the time series of inter-firm return correlations has very

little persistence from period to period. Thus even if individual firm variances could be predicted with certainty, anticipating future covariance would be infeasible with sample estimates of return correlation. This lack of persistence makes the well-documented deficiencies of using sample covariance matrices to construct portfolios seem inevitable. Related work by Box (2012) investigates whether incorporating predicted correlations into estimates of future return covariance can resolve these shortcomings. In out-of-sample tests, the author shows that minimum variance portfolios formed with predicted correlations experience a lower volatilities than those formed with covariance matrices estimated from historical data.

The general approach described in this article has a variety of additional applications that would benefit academics and practitioners in all areas of finance. The methodology for predicting correlations described above could also be used to identify matching firms for a pairs-trading strategy, or as an alternative approach for clustering firms into industries. With only minor adjustments, this basic framework could also be used to predict the correlation of a particular company's stock return with that of the broader market. These predicted correlations could then be used to estimate stock Betas that are less susceptible to measurement error resulting from short-sample estimates, but still responsive to time series variation in the fundamental relationships.

WORKS CITED

Anderson, T. W. and Cheng Hsiao. "Estimation of Dynamic Models with Error Components." *Journal of the American Statistical Association 76*, 1981: 598-606.

Arellano, Manuel and Stephen Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies 58 (2)*, 1991: 277-297.

Barber, Brad M. and Douglas Loeffler. "The "Dartboard" Column: Second-Hand Information and Price Pressure." *Journal of Financial and Quantitative Analysis Vol. 28., No. 2*, 1993: 273-284.

Barberis, Nicholas and Andrei Schleifer. "Style Investing." *Journal of Financial Economics 75*, 2003: 161-200.

Barberis, Nicholas, Andrei Shleifer, and Jeffrey Wurgler. "Comovement." *Journal of Financial Economics 75*, 2005: 283-317.

Bekaert, Geert, Robert J. Hodrick, and Xiaoyan Zhang. "International Stock Return Comovements." *The Journal of Finance Vol. LXIV, No. 6*, 2009: 2591-2627.

Bilisoly, Roger. *Practical Text Mining with Perl.* Hoboken, New Jersey: John Wiley & Sons. Inc., 2008.

Binongo, José Nilo G. and M. W. A. Smith. "The application of principal components analysis to stylometry." *Literary and Linguistic Computing, 14*, 1999: 445-466.

Bloom, Nick, Stephen Bond, and John van Reenen. "Uncertainty and Investment Dynamics." *Review of Economic Studies 74*, 2007: 397-415.

Box, Travis. "Minimum Variance Portfolios with Predicted Correlations." *Working paper*, 2012.

Boyer, Brian H. "Style-Related Comovement: Fundamentals or Labels." *The Journal of Finance Vol. LXVI, No. 1*, 2011: 307-332.

Brandt, Michael W., Alon Brav, John R. Graham, Alok Kumar. "The Idiosyncratic Volatility Puzzle: Time Trend of Specualtive Episode." *The Review of Financial Studies / v 23 n 2*, 2010: 865-899.

Campbell, John Y., Martin Lettau, Burton G. Malkiel, and Yexiao Xu. "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk." *The Journal of Finance Vol. LVI, No. 1*, 2001: 1-43.

Carhart, Mark M. "On Persistence in Mutual Fund Performance." *The Journal of Finance Vol. LII, No. 1*, 1997: 57-82 .

Chan, Luois K. C., Jason Karceski, Josef Lakonishok. "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model." *The Review of Financial Studies Vol. 12*, 1999: 937-974.

Chen, Huafeng, Shaojun Chen, and Feng Li. "Empirical Investigation of an Equity Pairs Trading Strategy." *Working paper*, 2012.

DeMiguel, Victor, Lorenzo Garlappi, Raman Uppal. "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?" *The Review of Financial Studies / v 22 n 5*, 2009: 1915-1953.

Engelberg, Joseph E. and Christopher A. Parsons. "The Causal Impact of Media in Financial Markets." *The Journal of Finance Vol. LXVI, No. 1*, 2011: 67-97.

Engle, Robert. "Dynamic Conditional Correlation: A Simple Class of Multivariate Gerneralized Autoregressive Conditional Heteroskedasticity Models." *Journal of Business & Economic Statistics*, 2012: 339-350.

Fama, Eugene and Keneth French. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics 33*, 1993: 3-56.

Fang, Lily and Joel Peress. "Media Coverage and the Cross-section of Stock Returns." *The Journal of Finance Vol. LXIV, No. 5*, 2009: 2023-2052.

Granger, C. W. J. "Investigating causal relations by econometric models and cross-spectral methods." *Econometrica 37 (3)*, 1969: 424-438.

Green, T. Clifton, and Byoung-Hyoun Hwang. "Price-based return comovement." *Journal of Financial Economics 93*, 2009: 37-50.

Greene, William H. "Minimum Distance and the Generalized Method of Moments." In *Econometric Analysis Sixth Edition*, 428-481. Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

Grossman, Sanford J. and Joesph E. Stiglitz. "On the Impossibility of Informationally Efficient Markets." *The American Economic Review, Vol. 70, No. 3*, 1980: 393-408.

Hasbrouck, Joel. "One Security, Many Markets: Determining the Contributions to Price Discovery." *The Journal of Finance Vol. L, No. 4*, 1995: 1175-1199.

Hoberg, Gerard and Gordon Phillips. "Dynamic Text-Based Industries and Endogenous Product Differentiation." *NBER Working Papers 15991, National Bureau of Economic Research, Inc.*, 2010b.

Hoberg, Gerard and Gordon Phillips. "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis." *The Review of Financial Studies / v 23 n 10*, 2010a: 3773-3811.

IBM Corporation. September 2012. www.ibm.com/energy/.

Irvine, Paul J., Jeffrey Pontiff. "Idiosyncratic Return Volatility, Cash Flows, and Product Markets." *The Review of Financial Studies / v 22 n 3*, 2009: 1149-1177.

Israelsen, Ryan D. "Does Common Analyst Coverage Explain Excess Comovement?" *working paper*, 2012.

Klarreich, Erica. "Bookish Math." *Science News, 164*, 2003: 392-394.

Langville, Amy N. and Carl D. Meyer. *Google's Page Rank and Beyond: The Science of Search Engine Rankings.* Princeton, New Jersey: Princeton University Press, 2006.

Ledoit, Olivier and Michael Wolf. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection." *Journal of Empirical Finance 10*, 2003: 603-621.

Lee, Charles M.C., Andrei Shleifer, and Richard H. Thaler. "Investor Sentiment and the Closed-End Fund Puzzle." *Journal of Finance XLVI*, 1991: 75-109.

Lewellen, Jonathan and Stefan Nagel. "The conditional CAPM does not explain asset-pricing anomalies." *Journal of Financial Economics 82*, 2006: 289-314.

Manning, Christopher D. and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: MIT Press, 1999.

Muslu, Volkan, Michael Rebello, and Yexiao Xu. "Sell-Side Analyst Research and Stock Comovement." *Working Paper*, 2012.

Pindyck, Robert S. and Julio J. Rotemberg. "The Excess Co-movement of Commodity Prices." *The Economic Journal, 100*, 1990: 1173-1189.

Pindyck, Robert S. and Julio Rotemberg. "The Comovement of Stock Prices." *The Quarterly Journal of Economics, Vol. 108, No. 4*, 1993: 1073-1104.

Pirinsky, Christo and Qinghai Wang. "Does Corporate Headquarters Location Matter for Stock Returns." *The Journal of Finance Vol. LXI, No. 4*, 2006: 1991-2015.

Tetlock, Paul C. "All the News That's Fit to Reprint: Do Investors React to Stale Information." *Review of Financial Studies 24*, 2011: 1481-1512.

Tetlock, Paul C. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance Vol. LXII, No. 3*, 2007: 1139-1168.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *The Journal of Finance Vol. LXIII, No. 3*, 2008: 1437-1467.

Veldkamp, Laura L. "Information Markets and the Comovement of Asset Prices." *Review of Economic Studies 73*, 2006b: 823-845.

Veldkamp, Laura L. "Media Frenzies in Markets for Financial Information." *The American Economic Review, Vol. 96, No. 3*, 2006a: 577-601.